

Microbial Systems Biology

9

microbiologynow

DNA Sequencing in the Palm of Your Hand

DNA sequencing technologies are revolutionizing microbiology at a remarkable pace. Innovations in next-generation sequencing have even tackled the issues of cost and portability. The world's first mobile nucleic acid sequencer—the MinION—is a palm-sized device that possesses 2000 tiny pore-containing proteins called nanopores. As single strands of nucleic acid travel through the nanopores, individual nucleotides are identified based on changes in electrical current. These current changes are relayed to a computer through a USB connection, which also powers the MinION. This miniature but mighty machine can display nucleic acid sequences from critical field samples in real time on a computer screen.


The utility of the MinION was clearly on display during the 2014–2015 Ebola virus hemorrhagic fever outbreak in West Africa. Scientists traveled to Guinea with three MinIONS in their luggage, a feat in itself as most DNA sequencers are too large and delicate to travel in baggage. Once in Guinea, scientists were able to survey the spread of different Ebola virus strains by analyzing unique nucleotide sequences present in each strain's genome. In as little as 48 hours after sample collection, Ebola virus genomes from 14 patients were determined using MinION sequencing. The photo here shows a researcher loading a patient's sample onto a MinION set up in a mobile field laboratory (inset).

Because the Ebola genome mutates on average every two weeks, the astonishing turnaround time provided by the MinION allowed epidemiologists to track geographical movements of different strains of the virus. This real-time analysis indicated that two major viral strains were the cause of Ebola persistence and that cross-border transmission between Sierra Leone and Guinea severely prolonged the outbreak. Traditional sequencing methods would not have supported such surveillance, as it requires weeks to obtain results after shipment of samples to remote laboratories.

While field biologists have envisioned a myriad of uses for the MinION, developers are currently attempting to modify it to operate from a smartphone instead of a computer. Also in its immediate future is outer space—NASA plans on testing the MinION on the International Space Station. And, because of its size, relatively low cost, and ease of use, the next frontier for the MinION will undoubtedly be the classroom.



- I Genomics 278
- II The Evolution of Genomes 288
- III Functional Omics 293
- IV The Utility of Systems Biology 302

 **Source:** Quick, J., et al. 2016. Real-time, portable genome sequencing for Ebola surveillance. *Nature* 530: 228–232. Photo credits: ©EMLab/Tommy Trenchard 2016.

Traditional approaches to studying microbes have focused on the analysis of individual biochemical pathways or molecular responses under specific conditions. While informative, this reductionist approach can only target a specific gene or subset of genes (or gene products) and fails to address the dynamic nature of organisms and how a network of biological molecules controls their behavior. By contrast, **systems biology** integrates different methodologies to yield an overview of an organism's response to its environment. Systems biology has been bolstered by the “*omics*” revolution—the ability to characterize and quantify large pools of biological molecules. Because the ability to store and analyze massive amounts of biological information by computer is essential to systems biology, the understanding of entire biological systems is evolving in parallel with computing power and storage and retrieval capabilities.

I • Genomics

The foundation of omics and systems biology lies in nucleic acid and protein *sequences*, characteristics ultimately controlled by the cell's genome. The **genome** is an organism's entire complement of genetic information, including genes that encode proteins, RNAs, and regulatory sequences, as well as any noncoding DNA that may be present. The genome sequence of an organism not only reveals its genes, but also yields important clues to how the organism functions. While new omics are coined with regularity, this chapter focuses on the major omics of systems biology—*genomics*, *transcriptomics*, *proteomics*, and *metabolomics*—and describes how these various pieces of the puzzle are integrated into microbial systems biology today (Figure 9.1).

The word **genomics** refers to the discipline of mapping, sequencing, analyzing, and comparing genomes. Here we review how genomes are sequenced and some techniques used to analyze these genomes and their gene content.

9.1 Introduction to Genomics

Advances in genomics rely heavily on improvements in molecular technology and computing power. The automation of DNA sequencing and the development of powerful computational tools for DNA and protein sequence analysis have reduced the cost and increased the speed at which genomes are analyzed. Thus the number of sequenced genomes has grown rapidly, with the major genomics bottleneck being the digestion of vast amounts of nucleic acid sequence data.

Genomics: Then and Now

The first genomes sequenced were those of small viruses over 40 years ago, and the first bacterial genome sequence was published in 1995. Today, DNA sequences from over 50,000 *Bacteria*, *Archaea*, and viruses, as well as datasets from metagenomic projects (Section 9.8), are available in public databases such as the Genomes Online Database (GOLD; see <https://gold.jgi.doe.gov> for an up-to-date list). With the goal of using genome sequences to advance systems and ecosystems biology, the United States Department of Energy's *Joint Genome Institute* (JGI) sponsors GOLD. Table 9.1 lists some representative genomes from *Bacteria*

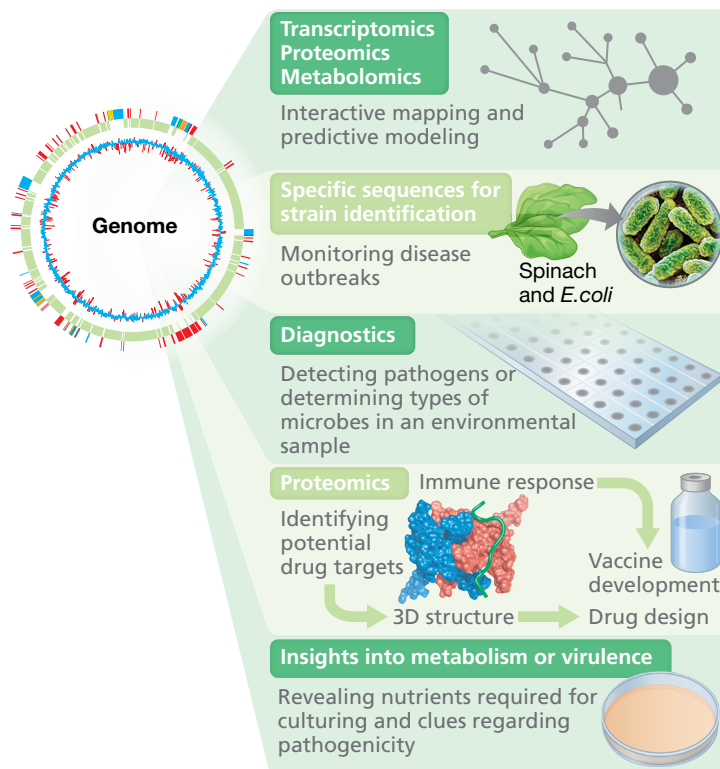


Figure 9.1 Utility of microbial genome sequences. A genome sequence allows for the development of omics approaches and tools for understanding, investigating, and monitoring microorganisms. It can also provide targets for drug and vaccine design.

and *Archaea*. The genomes of many eukaryotic organisms have also been sequenced, including the haploid human genome, which contains about 3.2 billion bp (~20,000 protein-encoding genes, Figure 9.2).

What Can Genomes Tell Us?

As we will discuss throughout this chapter, modern microbiology thrives on genome sequences; indeed, little in microbiology has been left untouched by genomic sequences. Microbial genome sequencing has discovered everything from genes encoding heat-stable enzymes in microbes that thrive in boiling water to genes that encode virulence factors in the most vicious pathogens. Genome sequencing has also been instrumental in developing microarrays for studying gene expression (Section 9.9), detecting horizontal transfer events (between microbes of different species, genera, and even kingdoms), monitoring and diagnosing disease outbreaks (based on the presence of “signature genes” of different pathogens), discovering CRISPRs (see Sections 10.13 and 11.12), understanding metabolic pathways, and discerning the growth requirements of microbes that have thus far defied laboratory culture.

The ability to sequence genomes has also been used to solve obscure medical mysteries. An excellent example is the genomics that revealed the causative agent of the “Black Death,” which swept through Europe in the middle of the fourteenth century (Figure 9.3a). While it was believed that the Black Death was caused by a massive outbreak of bubonic plague, a typically fatal disease caused by the bacterium *Yersinia pestis* (see Section 31.7),

TABLE 9.1 Genomes of select species of *Bacteria* and *Archaea*^a

Organism	Lifestyle ^b	Size (bp)	ORFs ^c	Comments
Bacteria				
<i>Nasuia deltocephalica</i>	E	112,091	137	Degenerate sap-feeding insect endosymbiont
<i>Tremblaya princeps</i>	E	138,931	121	Degenerate mealybug endosymbiont
<i>Hodgkinia cicadicola</i>	E	143,795	169	Degenerate cicada endosymbiont
<i>Buchnera aphidicola</i> BCc	E	422,434	362	Aphid endosymbiont
<i>Mycoplasma genitalium</i>	P	580,070	470	Smallest nonsymbiotic bacterial genome
<i>Borrelia burgdorferi</i>	P	910,725	853	Spirochete, linear chromosome, causes Lyme disease
<i>Rickettsia prowazekii</i>	P	1,111,523	834	Obligate intracellular parasite, causes epidemic typhus
<i>Treponema pallidum</i>	P	1,138,006	1,041	Spirochete, causes syphilis
<i>Methylophilaceae</i> family, strain HTCC2181	FL	1,304,428	1,354	Marine methylophilic, smallest free-living genome
<i>Thermotoga maritima</i>	FL	1,860,725	1,877	Hyperthermophile
<i>Deinococcus radiodurans</i>	FL	3,284,156	2,185	Radiation resistant, multiple chromosomes
<i>Bdellovibrio bacteriovorus</i>	FL	3,782,950	3,584	Predator of other bacteria
<i>Bacillus subtilis</i>	FL	4,214,810	4,100	Gram-positive genetic model
<i>Mycobacterium tuberculosis</i>	P	4,411,529	3,924	Causes tuberculosis
<i>Escherichia coli</i> K-12	FL	4,639,221	4,288	Gram-negative genetic model
<i>Escherichia coli</i> O157:H7	FL	5,594,477	5,361	Enteropathogenic strain of <i>E. coli</i>
<i>Pseudomonas aeruginosa</i>	FL	6,264,403	5,570	Metabolically versatile opportunistic pathogen
<i>Streptomyces coelicolor</i>	FL	8,667,507	7,825	Linear chromosome, produces antibiotics
<i>Bradyrhizobium japonicum</i>	FL	9,105,828	8,317	Nitrogen fixation, nodulates soybeans
<i>Sorangium cellulosum</i>	FL	14,782,125	11,559	Forms multicellular fruiting bodies
Archaea				
<i>Nanoarchaeum equitans</i>	P	490,885	552	Smallest nonsymbiotic cellular genome
<i>Methanocaldococcus jannaschii</i>	FL	1,664,976	1,738	Methanogen, hyperthermophile
<i>Pyrococcus horikoshii</i>	FL	1,738,505	2,061	Hyperthermophile
<i>Sulfolobus solfataricus</i>	FL	2,992,245	2,977	Hyperthermophile, sulfur chemolithotroph
<i>Haloarcula marismortui</i>	FL	4,274,642	4,242	Extreme halophile, bacteriorhodopsin
<i>Methanosarcina acetivorans</i>	FL	5,751,000	4,252	Acetate using methanogen

^aInformation on prokaryotic genomes can be found at <https://gold.jgi.doe.gov>.

^bE, endosymbiont; P, parasite; FL, free-living.

^cOpen reading frames. Genes encoding known proteins are included, as well as ORFs that could encode a protein greater than 100 amino acid residues. Smaller ORFs are not included unless they show similarity to a gene from another organism or unless the codon bias is typical of the organism being studied.

scientists could not be positive until they recovered and sequenced DNA samples from the teeth and bones of corpses of people known to have died from the Black Death. By comparisons of this ancient DNA with the genome of *Y. pestis*, the mystery behind this devastating medieval disease was unraveled: The Black Death was indeed bubonic plague.

Microbial genomics has also been used to identify new microbial phyla. For example, until recently, only three phyla of *Archaea* were known—*Euryarchaeota*, *Crenarchaeota*, and *Nanoarchaeota*. Because every cultured species was isolated from an extreme environment, many microbiologists concluded that *Archaea* were mainly extremophiles and that they did not inhabit oceans, lakes, and soil in significant numbers. However, based on environmental 16S rRNA gene sequencing, *Archaea* only marginally affiliated with *Crenarchaeota* were detected in marine and freshwater samples. Who were these organisms, and how were

they making a living? Subsequently, *Nitrosopumilus*, the first ammonia-oxidizing (nitrifying) archaeon known, was isolated (Figure 9.3b) (↗ Section 14.11). Using the powerful analytical tools of genomics, the genomes of two distinct ammonia-oxidizing *Archaea* were compared with those of all other *Archaea*. This genomic analysis clearly showed that these ammonia-oxidizing *Archaea* belonged in a new phylum, now called the *Thaumarchaeota* (↗ Section 17.5).

The above is just a taste of how genomics has impacted microbiology. Other relevant examples will appear regularly as you make your way through this book. The major message here is twofold: (1) We are clearly living in the era of microbial genomics, and (2) the genomics revolution has spawned a wealth of powerful tools to attack old problems in new ways. Indeed, in the past 40 years or so, microbiology as a science has leapt forward farther and faster than at any time in its history.

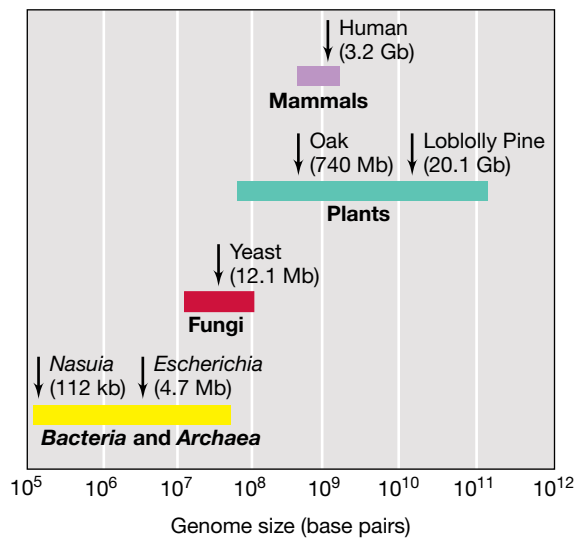


Figure 9.2 Genome sizes of microbial cells and higher organisms. Compare with viral genome sizes in Figure 10.1.

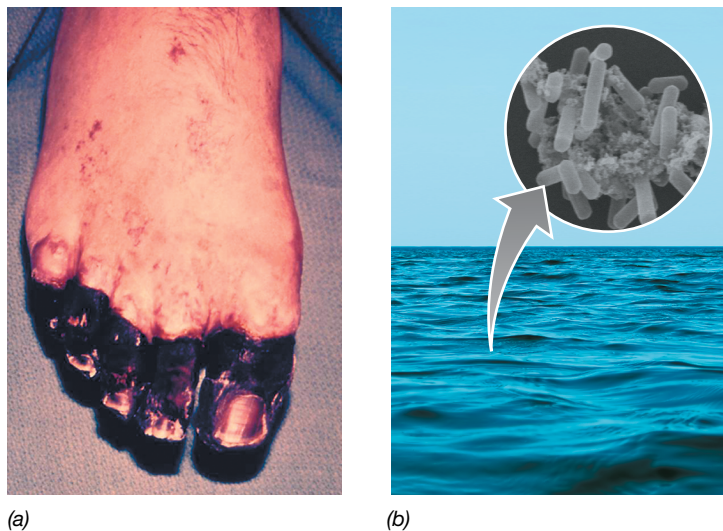


Figure 9.3 What genomes can tell us. (a) Genomics helped solve an ancient medical mystery surrounding plague. (The blackened skin on the toes of this modern plague victim originates from hemorrhaging due to systemic infection with *Yersinia pestis*.) (b) Genome sequencing was used to assign the marine ammonia-oxidizing archaeon *Nitrosopumilus* to a new phylum of Archaea, the *Thaumarchaeota*.

MINIQUIZ

- How many protein-encoding genes are in the human genome?
- List three examples of how genomics has led to major new discoveries in microbiology.

9.2 Sequencing and Annotating Genomes

In biology, the term **sequencing** refers to determining the precise order of subunits in a macromolecule. In the case of DNA (or RNA), the sequence is the *order* in which the nucleotides are aligned. DNA sequencing today forms the heart of the omics revolution and its technology is advancing so quickly that new

methods appear every year. Interestingly, however, despite the technological breakthroughs that have catapulted us into the omics age, some of the earliest sequencing methodologies—born of simple yet brilliant basic science—form the foundation of the latest methods today (see next subsection).

After sequencing and assembly of the gene fragments, the next step is *genome annotation*, the conversion of raw sequence data into a list of the genes and other functional sequences present in the genome. The term **bioinformatics** refers to the use of computers to store and analyze the sequences and structures of nucleic acids and proteins. Improved sequencing methods are now generating data faster than it can be properly analyzed. Thus, at present, annotation is the major “bottleneck” in genomics. Here we focus on the process of genome sequencing, assembly, and annotation.

DNA Sequencing

The first widely used method for sequencing DNA was the dideoxy method developed by the British scientist Fred Sanger, who won a Nobel Prize (his second) for this accomplishment. In the Sanger procedure the sequence is determined by making a copy of the original single-stranded DNA in a process similar to the polymerase chain reaction (PCR, [see Section 12.1](#)). The secret behind the Sanger method was the addition of a mixture of normal deoxyribonucleotides (dNTPs) and small amounts of the corresponding *dideoxynucleotides* (ddNTPs), one for each of the four bases—adenine, guanine, cytosine, and thymine—to the mixture used to make the DNA copy (**Figure 9.4a**). The dideoxy analog is a specific *chain-terminator*; because it lacks a 3'-hydroxyl, the analog prevents further elongation of the chain after its insertion. Because ddNTPs insert randomly, DNA chains of varying length are produced in the synthesis reaction (**Figure 9.4a**). Sanger sequencing originally used radioactive labels, but automated systems were quickly developed that used a separate fluorescent label for each different ddNTP and that detected the DNA products (separated by passing through a sizing column) with a laser (**Figure 9.4a**).

Because the original Sanger method was dependent on primers binding to a known sequence and was limited to around 800 nucleotides per reaction, chromosomes or large DNA molecules could not be sequenced in a single reaction. Instead large DNA molecules had to be cut into smaller fragments and cloned into vectors for sequencing. This led to the development of new sequencing technologies, which appear now with such regularity that the term “next-generation sequencing” is commonly used to describe the latest and greatest in nucleic acid sequencing. For example, *pyrosequencing*, a *second-generation sequencing method* still widely used today, was developed to improve the process by employing the light-emitting enzyme luciferase to detect incorporation of dNTPs by emitting a pulse of light (**Figure 9.4b**). **Table 9.2** summarizes modern sequencing methods and illustrates how the cost of sequencing 1 megabase (Mbp, million base pairs) of DNA has dropped over 100,000-fold in the last 15 years.

Genome Assembly and Annotation

Regardless of which sequencing system is used, the sequences obtained must be *assembled* before they can be analyzed. Genome assembly consists of putting the fragments in the correct order and eliminating overlaps. Then, for assembled genomic sequences

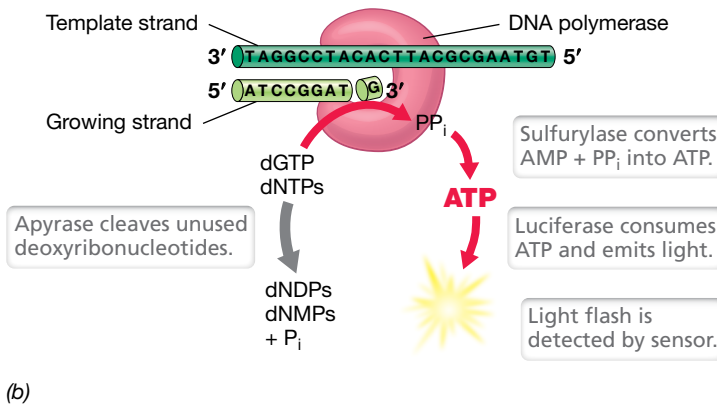
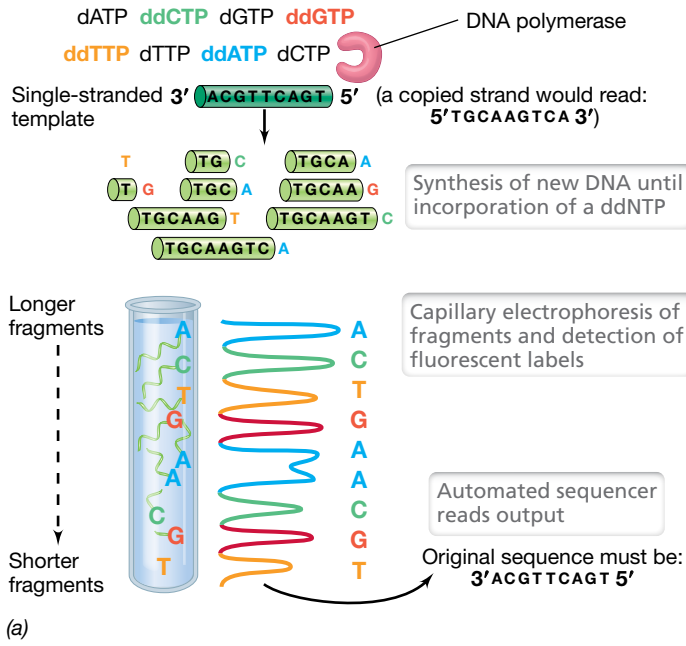


Figure 9.4 DNA sequencing. (a) Sanger sequencing. When a polymerase incorporates a ddNTP during synthesis, the chain of DNA is terminated. The identity of the terminal ddNTP can be determined by capillary electrophoresis and fluorescence detection. (b) Pyrosequencing. Whenever a new dNTP is inserted into the growing strand of DNA (red arrows), pyrophosphate (PP_i) is released and is used to make ATP from AMP by the enzyme sulfurylase. The ATP is consumed by the enzyme luciferase, which releases light. Unused dNTPs are degraded by the enzyme apyrase (gray arrow).

to be useful, they must be *annotated* in order to identify genes and other functional regions. Many of the tasks surrounding genome assembly and annotation are highly computational. For genome assembly, a computer examines many short DNA fragments that have been sequenced and deduces their order by detecting all of the instances where two fragments of DNA possess overlapping sequence (Figure 9.5). These overlaps are used to merge sequencing reads into *contigs*, or contiguous consensus sequences. Individual contigs with overlapping ends are then aligned to form scaffolds (contigs as well as gaps) that are ultimately used to generate a map representing the complete genome.

From the genome map, the annotation process can begin. Because the genomes of *Bacteria* and *Archaea* possess very few intervening sequences (introns, Section 4.6), their genomes essentially consist of a series of **open reading frames (ORFs)**

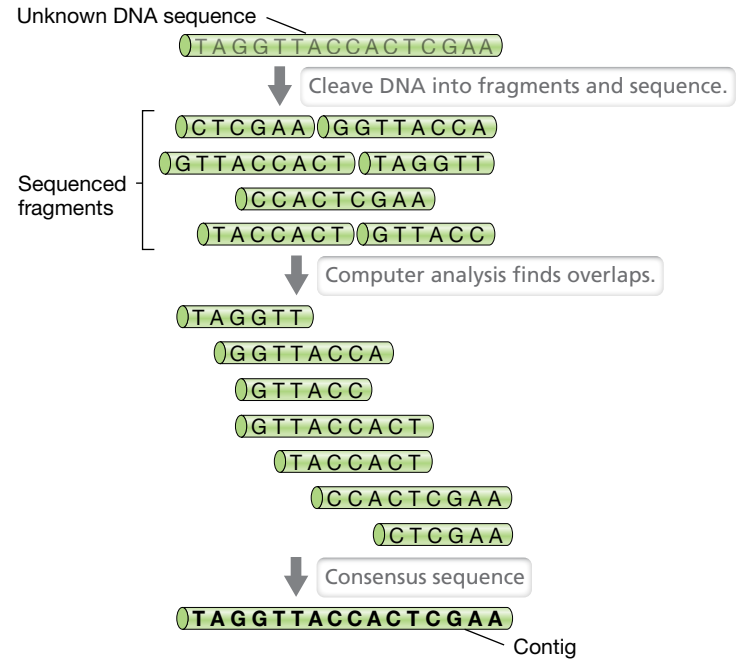


Figure 9.5 Computer assembly of DNA sequence. Most DNA sequencing methods generate vast numbers of short sequences (30 to several hundred bases) that must be assembled. The computer searches for overlaps in the short sequences and then arranges them to form contigs, or a consensus sequence.

separated by short regulatory regions and transcriptional terminators. A *functional ORF* is one that actually encodes a protein (Section 4.9) and can be identified from a computer search of the sequence (Figure 9.6). Although any given cellular gene is always transcribed from one DNA strand, a gene can actually be

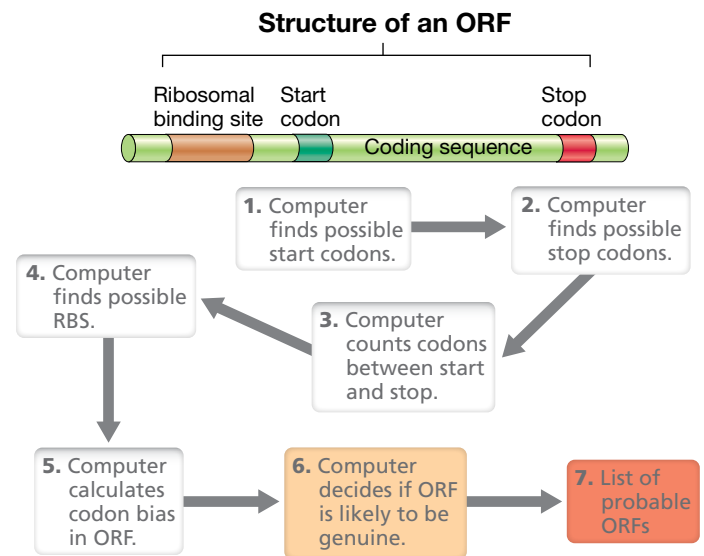


Figure 9.6 Computer identification of possible ORFs. The computer scans the DNA sequence looking first for start and stop codons. It then counts the number of codons in each uninterrupted reading frame and rejects those that are too short. The probability of a genuine ORF is made stronger if a likely ribosomal binding site (RBS) is found the correct distance in front of the reading frame. Codon bias calculations are used to test whether an ORF complies with the codon usage of the organism being examined.

TABLE 9.2 DNA sequencing methods

Generation	Method	Features
First generation	Sanger dideoxy method (radioactivity or fluorescence; DNA amplification)	Read length: 700–900 bases Used for the Human Genome Project
Second generation	454 Pyrosequencing (fluorescence; DNA amplification; massively parallel) Illumina/Solexa method (fluorescence; DNA amplification; massively parallel) SOLiD method (fluorescence; DNA amplification; massively parallel)	Read length: 400–500 bases Used to sequence genome of James Watson (completed 2007) Read length: 50–100 bases Giant panda genome (2009; Beijing Genome Institute); Denisovan genome (2010) Read length 50–100 bases
Third generation	HeliScope Single Molecule Sequencer (fluorescence; single molecule) Pacific Biosciences SMRT (fluorescence; single molecule; zero mode waveguide)	Read length: up to 55 bases Fossil DNA accuracy greatly improved Read length: 2500–3000 bases
Fourth generation	Ion torrent (electronic—pH; DNA amplification) Oxford nanopore (electronic—current; single molecule; real time)	Read length: 100–200 bases Sequenced genome of Intel cofounder Gordon Moore (originator of Moore's law), 2011 Read length: thousands of bases Portable MinION unit is approximately the size of a flash drive

located on either strand and thus computer inspection of both strands of DNA is required.

Finding and Identifying ORFs

The first step in finding an ORF is to locate *start* and *stop* codons in the sequence (🔗 Section 4.9 and Table 4.4). However, in-frame start and stop codons appear randomly with reasonable frequency; thus, further clues are needed. In *Bacteria*, translation begins at start codons located immediately downstream of a ribosome-binding sequence (Shine–Dalgarno site) on the mRNA (🔗 Section 4.9). Thus, locating potential ribosome-binding sequences in addition to start and stop codons helps decide both whether an ORF is functional and which start codon is actually used. In addition, an ORF is more likely to be functional if its sequence is similar to those of ORFs in the genomes of other organisms (regardless of whether they encode known proteins) or if the ORF includes a sequence known to encode a protein functional domain. This is because proteins with similar functions in different cells tend to share a common evolutionary origin and thus share sequence and structural features (Section 9.5). A computer can search for sequence similarities in major databases such as GenBank (<http://www.ncbi.nlm.nih.gov/Genbank/>) using *BLAST* (Basic Local Alignment Search Tool), an algorithm that can compare a nucleic acid or protein sequence with all other such sequences in the database.

Other issues must be considered in a genome annotation as well. For example, more than one codon exists for many of the 20 common amino acids (🔗 Table 4.4), and some codons are used more frequently than others. The latter is known as **codon bias** (codon usage) and differs greatly between organisms. For example, Table 9.3 shows the different usage of the six arginine codons in *Escherichia coli* compared to their usage in humans and fruit flies. If the codon bias in a given ORF differs greatly from the consensus for the organism containing it, that ORF may be nonfunctional or may be functional but obtained by horizontal gene transfer (Section 9.6).

Genomic Analyses: The Final Tally

No genome sequence project ends with 100% of the genome identified. In fact, this is one of the exciting findings of genomic analyses: Many genes in microbes almost certainly encode proteins whose function(s) remain unknown. Although there are differences among organisms, in most genomes the percentage of genes whose role can be clearly identified is approximately 70% of the total number of ORFs detected. Uncharacterized (or unknown) ORFs are said to encode *hypothetical proteins*, proteins that probably exist although their function is unknown. These ORFs have uninterrupted reading frames of reasonable length and the necessary start and stop codons and ribosome-binding site (Figure 9.6); however, the proteins they encode lack sufficient amino acid sequence homology with any known protein to be unambiguously identified. Some gene annotations can only assign a gene to a protein family or to a general function (such as “transport protein”) without being more specific. Many of the unidentified genes in *E. coli* are thought to encode proteins that play a role in some unidentified regulatory process or are proteins required only for special nutritional or environmental conditions. A few may also function as “backups” of key enzymes.

TABLE 9.3 Examples of codon bias

Arginine codon ^a	Usage of each arginine codon (%)		
	<i>Escherichia coli</i>	Fruit fly	Human
AGA	1	10	22
AGG	1	6	23
CGA	4	8	10
CGC	39	49	22
CGG	4	9	14
CGU	49	18	9

^aArginine has six codons; see Table 4.6.

In addition to protein-encoding genes, some genes encode RNA molecules that are not translated. Such genes therefore lack start codons and may well have multiple stop codons within the gene. Some noncoding RNAs, such as tRNAs and rRNAs, are easy to detect because they are well characterized and are highly conserved. However, many noncoding regulatory RNA molecules (Section 6.11) are conserved only in their three-dimensional structure, with little sequence homology. Thus transcriptomics, specifically RNA-Seq (Section 9.9), has become instrumental in identifying these noncoding genes.

With this general background in nucleic acid sequencing and the coding features of genomes, we move on to compare the nature of genomes in various microbial groups. We begin with the *Bacteria* and *Archaea* where thousands of genome sequences are available for comparative analyses.

MINIQUIZ

- What key molecules are essential for Sanger sequencing?
- What is an open reading frame (ORF)? What is a hypothetical protein?
- What is the major problem in identifying genes encoding nontranslated RNA?

9.3 Genome Size and Gene Content in Bacteria and Archaea

Once a genome has been assembled, *comparative genomics* using databases such as MicrobesOnline (<http://www.microbesonline.org>)—which contains nearly 4000 microbial genome sequences—can be used to probe its biological secrets. By using comparative genomics, it has been determined that genomes of *Bacteria* and *Archaea* show a strong correlation between genome size and open reading frame (ORF) content (Figure 9.7). Regardless of the organism, each megabase pair of DNA in a prokaryotic cell encodes about 1000 ORFs, and as the size of these genomes increases, the gene number also increases proportionally.

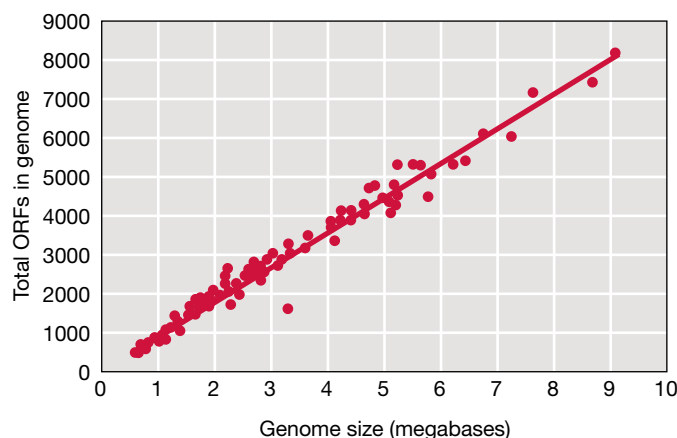


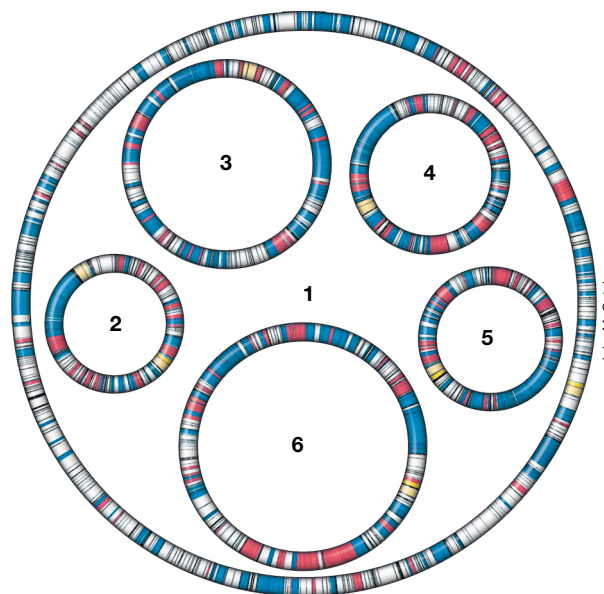
Figure 9.7 Correlation between genome size and ORF content in prokaryotic cells. Analyses of 115 completed genomes from species of both *Bacteria* and *Archaea*. Data from *Proc. Natl. Acad. Sci. (USA)* 101: 3160–3165 (2004).

This contrasts markedly with the genomes of eukaryotes, in which noncoding DNA (introns, Section 4.6) may constitute a large fraction of the genome, especially in organisms with large genomes (Figure 9.2).

Bacterial genomes range in size from the insect symbiont *Tremblaya princeps* with 121 protein-encoding genes to the soil-dwelling *Sorangium cellulosum* with nearly 100 times as many genes (Figure 9.2 and Table 9.1). While about 1300 genes has been the benchmark for the number of genes necessary for a cell to have a free-living existence, the recent discovery of free-living marine *Actinobacteria* containing approximately 800 genes has called this earlier estimate into question.

Small Genomes

The smallest cellular genomes belong to bacteria that are parasitic or endosymbiotic (cells that live inside other cells), with the insect symbionts *Tremblaya* (in mealybugs) and *Hodgkinia* (in cicadas) possessing some of the smallest genomes (around 140 kbp, Table 9.1 and Figure 9.8). The absolute smallest genome discovered thus far is that of *Nasuia deltocephalicol*, a sap-feeding insect symbiont whose genome is only 112 kbp. Because of their reduced



1. <i>Mycoplasma genitalium</i> (Mollicutes) 580.1 kbp GC: 31.7%	4. <i>Carsonella</i> (Gammaproteobacteria) 159.6 kbp GC: 16.6%
2. <i>Tremblaya</i> (Betaproteobacteria) 138.9 kbp GC: 58.8%	5. <i>Hodgkinia</i> (Alphaproteobacteria) 143.7 kbp GC: 58.4%
3. <i>Zinderia</i> (Betaproteobacteria) 208.5 kbp GC: 13.5%	6. <i>Sulcia</i> (Bacteroidetes) 245.5 kbp GC: 22.4%

Figure 9.8 Symbiont genomes. Five symbiont genomes are shown drawn to scale inside the circle representing the genome of a *Mycoplasma*. Blue: genes encoding genetic information processing; Red: genes encoding amino acid and vitamin biosyntheses; Yellow: rRNA genes; White: other genes; Gaps indicate noncoding DNA. Kbp, kilobase pairs. GC indicates percentage of nucleotides that are guanine or cytosine.

genome size, such symbionts are totally dependent on their insect host cells for survival and nutrients. In turn, the symbionts provide the insect with essential amino acids and other nutrients that the insect cannot synthesize.

With genomes of around 500 kbp, *Mycoplasma* (Bacteria) and *Nanoarchaeum equitans* (Archaea) have the smallest genomes among parasitic prokaryotic cells (Table 9.1). *N. equitans* is a hyperthermophile and a parasite of another hyperthermophile, the archaeon *Ignicoccus* (↪ Section 17.6). *N. equitans* lacks virtually all genes that encode metabolic proteins and presumably depends on its host for most catabolic as well as anabolic functions. While some pathogens such as *Mycobacterium tuberculosis* have quite large genomes (4.4 Mbp), the genomes of most human pathogens, such as *Mycoplasma*, *Chlamydia*, and *Rickettsia*, are smaller than the largest known viral genome, that of *Pandoravirus* (2.5 Mbp, ↪ Section 10.1).

Using *Mycoplasma*, which has around 500 genes, as a starting point, it has been estimated that around 250–300 genes are the minimum number possible for a viable cell. These estimates rely partly on comparisons with other small genomes. In addition, systematic mutagenesis has been performed to identify essential genes. For example, experiments with *Escherichia coli* and *Bacillus subtilis*, both of which have about 4000 genes, indicate that approximately 300–400 genes are essential depending on the growth conditions. However, in these experiments the bacteria were provided with many nutrients, allowing them to survive without many genes that encode biosynthetic functions. Most of the “essential genes” identified are present in other Bacteria as well and approximately 70% have also been found in Archaea and eukaryotes.

Large Genomes

Some Bacteria have genomes that are as large as those of some eukaryotic microbes. In fact, because eukaryotes tend to have significant amounts of noncoding DNA and bacteria do not, some bacterial genomes actually have more genes than microbial eukaryotes, despite having less DNA. For example, the genome of *Bradyrhizobium japonicum*, a bacterium that forms nitrogen-fixing root nodules on leguminous plants such as soybeans (↪ Section 23.3), has 9.1 Mbp of DNA and 8300 ORFs, whereas the genome of the baker's yeast *Saccharomyces cerevisiae*, a eukaryote, has 12.1 Mbp of DNA and only 5800 ORFs (see Tables 9.1 and 9.5).

The largest bacterial genome known is that of *S. cellulosum*, a species of the gliding myxobacteria (↪ Section 15.17). With just under 14.8 Mbp on a single circular chromosome, it has more DNA than several eukaryotes including yeast and the pathogenic protozoans *Cryptosporidium* and *Giardia* (see Table 9.5). The *S. cellulosum* genome is composed of roughly 10.5% noncoding DNA and 11,559 protein-encoding genes, making it over three times larger than the genome of *E. coli*. Interestingly, the *S. cellulosum* genome encodes 508 kinases (enzymes that phosphorylate other proteins to regulate their activity), which is over three times that of any other genome including those of eukaryotes. This suggests that the lifestyle of *S. cellulosum* is highly diverse and that its ecological success requires extensive regulation. In contrast to Bacteria, the

largest genomes found in species of Archaea thus far are only about 5 Mbp (Table 9.1).

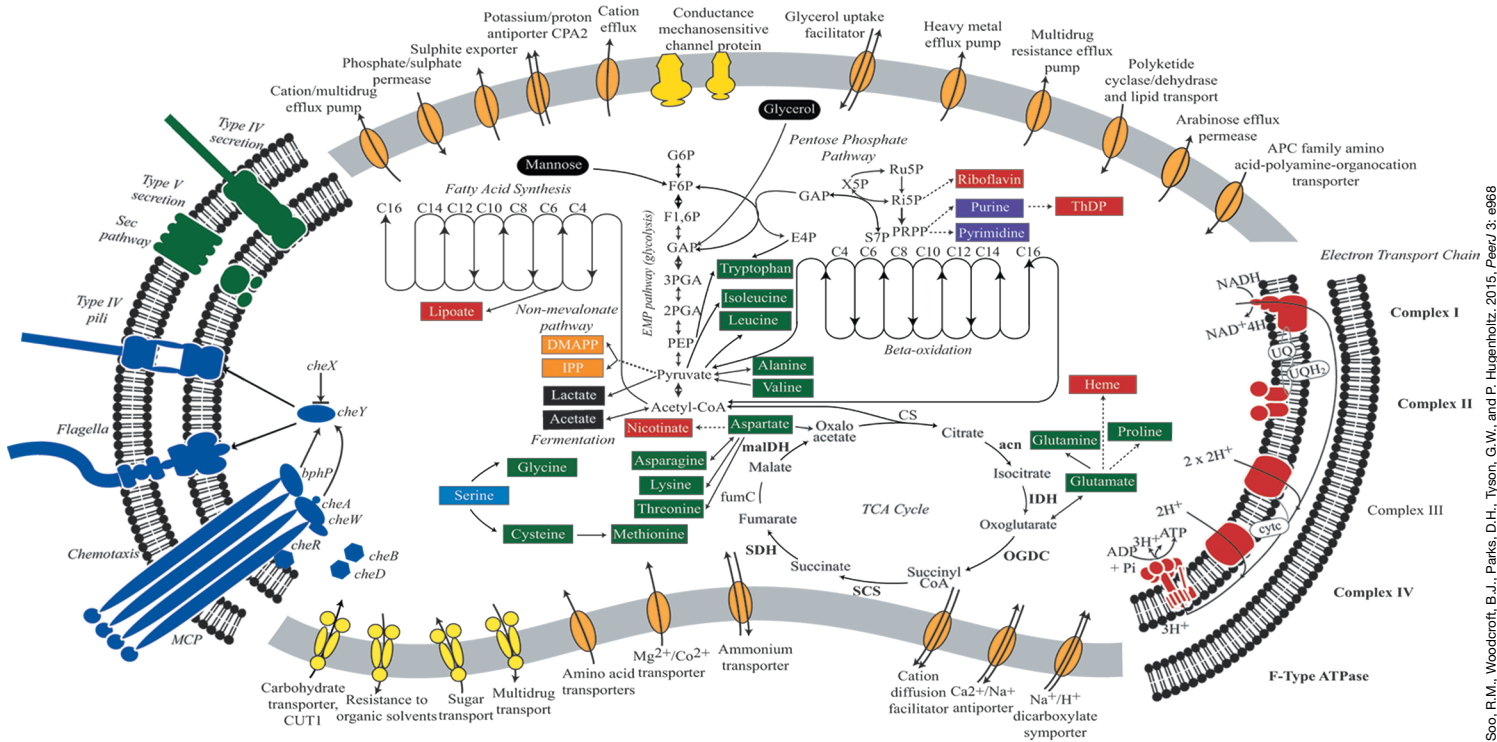
Gene Content of Bacterial Genomes

The complement of genes in a particular organism reveals its capabilities. Conversely, genomes are molded by adaptation to particular lifestyles. Comparative analyses are useful when searching for genes that encode enzymes that probably exist because of the lifestyle of an organism, and in some cases these searches yield big surprises. For example, *Vampirovibrio chlorellavorus* has been reported to be a predatory bacterium that attacks its host, the green alga *Chlorella*, by surface attachment and ultimate ingestion of its cellular contents (thus the terms “vámpir” from Hungarian, meaning “blood sucker,” and “vorus” from Latin, meaning “to devour,” in the organism's name). Isolates of *V. chlorellavorus* existed only as 36-year-old freeze-dried samples that had not been successfully revived. However, by using advanced sequencing techniques, the genome of *V. chlorellavorus* was recovered, and surprisingly, genomic analyses indicated that *V. chlorellavorus* falls within the phylum Cyanobacteria, even though it lacks genes for photosynthesis.

Figure 9.9, reprinted from a scientific journal, is included here to give you an idea of why a microbe's genome should be sequenced and the amazing amount of information that can be gleaned from annotation, though the details are beyond the scope of this chapter. The figure summarizes some of the metabolic pathways and transport systems of *V. chlorellavorus* deduced from analysis of its genome. These include an electron transport chain for microaerobic growth, the ability to ferment, chemotaxis abilities, and the synthesis of 15 of the 20 essential amino acids. Comparative genomics also indicated that *V. chlorellavorus* used a conjugative type IV secretion system (↪ Section 4.13) to attack its prey, the first discovery of this strategy in a predatory bacterium.

A functional analysis of genes and their activities in several bacteria is given in Table 9.4. Thus far, a distinct pattern of gene distribution in Bacteria has emerged. Metabolic genes are typically the most abundant class in bacterial genomes, although genes for protein synthesis overtake metabolic genes on a percentage basis as genome size decreases (Table 9.4 and Figure 9.10). Although many genes can be dispensed with, genes that encode the protein-synthesizing apparatus cannot. Thus, the smaller the genome, the greater the percentage of genes that encode translational processes. Conversely, the larger the genome, the more genes there are for transcriptional regulation and signal transduction.

Analyses of gene categories have also been done for several Archaea. On average, Archaea devote a higher percentage of their genomes to energy and coenzyme production than do Bacteria (this result is undoubtedly skewed a bit due to the large number of novel coenzymes produced by methanogenic Archaea, ↪ Section 14.17). On the other hand, Archaea appear to contain fewer genes for carbohydrate metabolism and membrane functions (such as transport and membrane biosynthesis) than do Bacteria. However, this conclusion may also be skewed a bit because the corresponding pathways have been less studied in Archaea than in Bacteria and many of the relevant archaeal genes remain unidentified.



Soo, R.M., Woodcroft, B.J., Tyson, G.W., and P. Hugenholtz. 2015. *PeerJ* 3: e968

Figure 9.9 Functional and metabolic predictions for *Vampirovibrio chlorellavorus* based on genomic annotation. Although the details are beyond our discussion, the figure illustrates the power of genomic sequencing and annotation. Within the membrane, the following systems are highlighted:

secretion (green), chemotaxis and movement (blue), electron transport (red), ATP-binding cassette transporters (yellow), and permeases/pumps/transporters (orange). Black ovals indicate substrates that enter the glycolysis pathway, while fermentation end-products are indicated as black rectangles. Colors of internal

compounds correspond to the following: green (amino acids), red (cofactors and vitamins), purple (nucleotides), and orange (non-mevalonate pathway products). Note that genes for synthesis of serine (highlighted in blue) are not present, so presumably it is transported into the cell. Adapted from Soo, R.M., et al. 2015. *PeerJ* 3: e968.

MINIQUIZ

- What lifestyle is typical of *Bacteria* and *Archaea* that contain fewer than 500 protein-encoding genes?
- Which is likely to have more genes, a species of *Bacteria* with 8 Mbp of DNA or a eukaryote with 10 Mbp? Explain.
- In prokaryotic cells with the largest genomes, which gene category contains the largest percentage of genes?

9.4 Organelle and Eukaryotic Microbial Genomes

Mitochondria and chloroplasts are eukaryotic cell organelles derived from endosymbiotic bacteria (↔ Sections 2.15 and 18.1), and thus share many fundamental traits with *Bacteria* to which they are phylogenetically related. The genomes of both organelles encode the machinery necessary for protein synthesis including

TABLE 9.4 Gene function in some genomes of *Bacteria*

Functional categories	Percentage of genes		
	<i>Escherichia coli</i> (4.64 Mbp) ^a	<i>Haemophilus influenzae</i> (1.83 Mbp) ^a	<i>Mycoplasma genitalium</i> (0.58 Mbp) ^a
Metabolism	21.0	19.0	14.6
Structure	5.5	4.7	3.6
Transport	10.0	7.0	7.3
Regulation	8.5	6.6	6.0
Translation	4.5	8.0	21.6
Transcription	1.3	1.5	2.6
Replication	2.7	4.9	6.8
Other, known	8.5	5.2	5.8
Unknown	38.1	43.0	32.0

^aChromosome size, in megabase pairs. Each organism listed contains only a single circular chromosome.

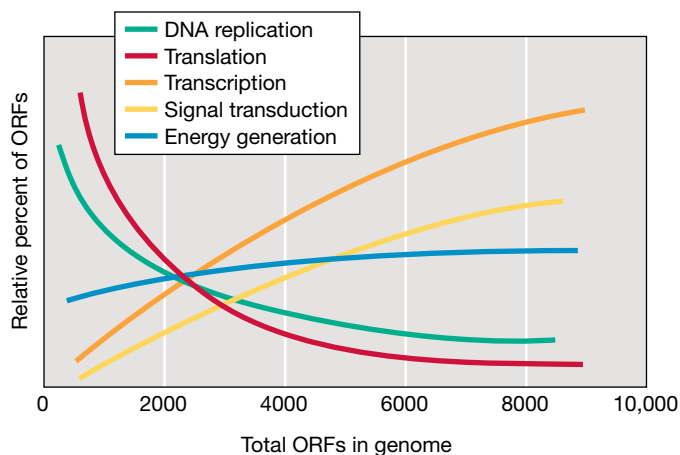


Figure 9.10 Functional category of genes as a percentage of the genome. The percentage of genes encoding products for translation or DNA replication is greater in organisms with small genomes, whereas the percentage of transcriptional regulatory genes is greater in organisms with large genomes.

ribosomes, transfer RNAs, and the other components necessary to drive translation. The genomes of several microbial eukaryotes have also been sequenced (Table 9.5), and their size varies widely (Figure 9.2). Certain single-celled protozoans, including the free-living ciliate *Paramecium* (40,000 genes) and the pathogen *Trichomonas* (60,000 genes), have significantly more genes than do humans (Table 9.5). In this section we focus on organellar genomes and the genomes of a few select microbial eukaryotes.

The Chloroplast Genome

Green plant cells contain chloroplasts, the organelles that perform photosynthesis (↻ Section 14.1). All known chloroplast genomes are circular DNA molecules, and each chloroplast contains several identical copies of the genome. The typical chloroplast genome is about 120–160 kbp and contains two inverted repeats of 6–76 kbp that each encode copies of the three rRNA genes (Figure 9.11). As might be expected, many chloroplast genes encode proteins for photosynthetic reactions and autotrophy. For example, the enzyme RubisCO catalyzes the first step in CO₂ fixation in the Calvin cycle (↻ Section 14.5). The *rbcl* gene encoding the large subunit of RubisCO is present on the chloroplast genome (see Figure 9.11), whereas the gene for the small subunit, *rbcs*, resides in the plant cell nucleus and its protein product must be imported from the cytoplasm into the chloroplast after synthesis.

The chloroplast genome also encodes tRNAs used in translation, several proteins used in transcription and translation, and some other proteins. Not all chloroplast proteins are encoded by the chloroplast genome; some are nuclear encoded. These are thought to be genes that migrated to the nucleus as the chloroplast evolved from an endosymbiotic cell into a photosynthetic organelle. Introns are common in chloroplast genes and are primarily of the self-splicing type (↻ Section 4.6).

Mitochondrial Genomes and Proteomes

Mitochondria are the eukaryotic cell's respiratory organelles and are present in all but a few eukaryotes (↻ Sections 2.15 and 18.1). Mitochondrial genomes primarily encode proteins for oxidative

TABLE 9.5 Some eukaryotic nuclear genomes^a

Organism	Comments	Lifestyle ^b	Genome size (Mbp)	Haploid chromosomes	ORFs
Nucleomorph of <i>Bigeloviella natans</i>	Degenerate endosymbiotic nucleus	E	0.37	3	331
<i>Encephalitozoon intestinalis</i>	Smallest known eukaryotic genome, human pathogen	P	2.3	11	1,800
<i>Cryptosporidium parvum</i>	Parasitic protozoan	P	9.1	8	3,800
<i>Plasmodium falciparum</i>	Malignant malaria	P	23	14	5,300
<i>Saccharomyces cerevisiae</i>	Yeast, a model eukaryote	FL	13.4	16	5,800
<i>Ostreococcus tauri</i>	Marine green alga, smallest free-living eukaryote	FL	12.6	20	8,200
<i>Aspergillus nidulans</i>	Filamentous fungus	FL	30	8	9,500
<i>Giardia intestinalis</i> (also called <i>Giardia lamblia</i>)	Flagellated protozoan, causes acute gastroenteritis	P	12	5	9,700
<i>Drosophila melanogaster</i>	Fruit fly, model organism for genetic studies	FL	180	4	13,600
<i>Caenorhabditis elegans</i>	Roundworm, model for animal development	FL	97	6	19,100
<i>Mus musculus</i>	Mouse, a model mammal	FL	2,500	23	25,000
<i>Homo sapiens</i>	Human	FL	2,850	23	25,000
<i>Arabidopsis thaliana</i>	Model plant for genetics	FL	125	5	26,000
<i>Paramecium tetraurelia</i>	Ciliated protozoan	FL	72	>50	40,000
<i>Pinus taeda</i>	Loblolly pine tree	FL	20,000	19	50,000
<i>Trichomonas vaginalis</i>	Flagellated protozoan, human pathogen	P	160	6	60,000

^aAll data are for the haploid nuclear genomes of these organisms in megabase pairs. For most large genomes, both size and ORFs listed are best estimates due to large numbers of repetitive sequences and/or introns in the genomes.

^bE, endosymbiont; P, parasite; FL, free-living.

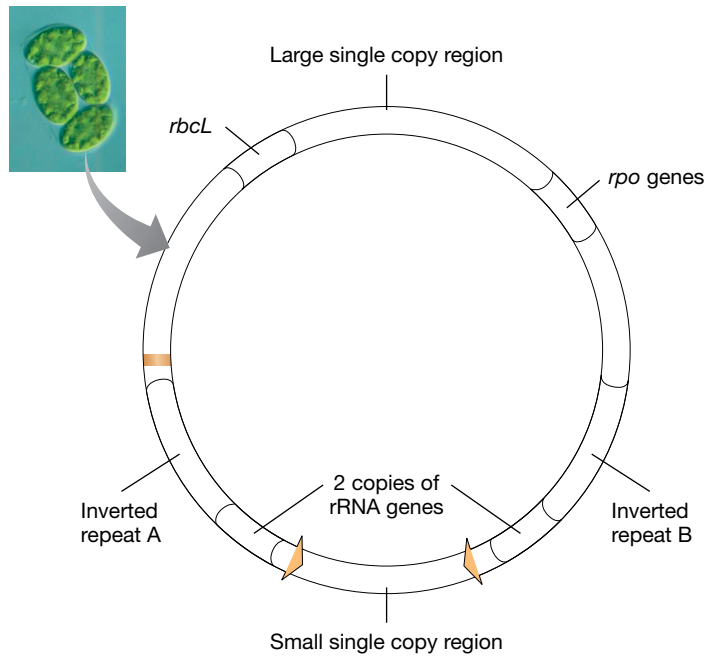
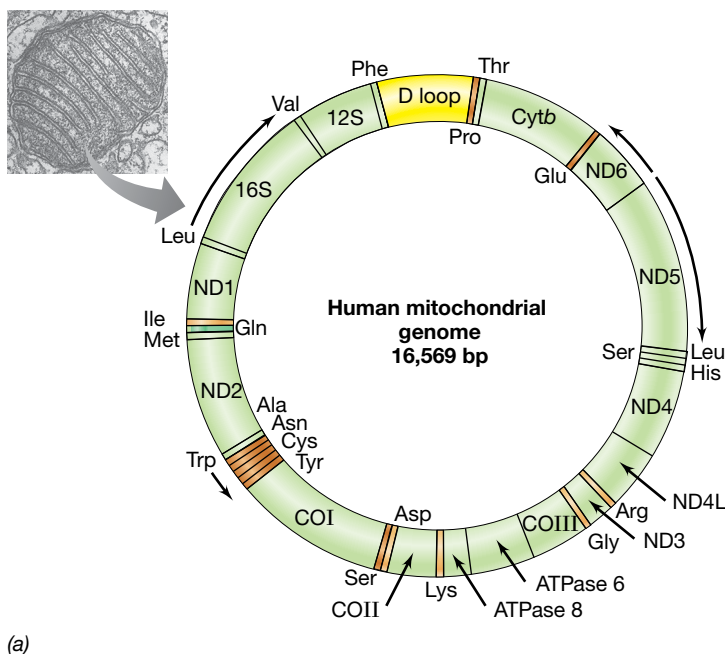


Figure 9.11 Map of a typical chloroplast genome. The inverted repeats each contain a copy of the three genes for rRNA (5S, 16S, and 23S). The large subunit of RubisCO is encoded by *rbcL* and the chloroplast RNA polymerase by *rpo* genes. Inset: Photo of four cells of the green alga *Makinoella* with chloroplasts clearly visible.

phosphorylation and, like chloroplast genomes, also encode proteins, rRNAs, and tRNAs for protein synthesis. However, most mitochondrial genomes encode far fewer proteins than those of chloroplasts. The largest mitochondrial genome known has only 62 protein-encoding genes, but others contain as few as three. The mitochondria of almost all mammals, including humans, encode only 13 proteins in addition to 22 tRNAs and 2 rRNAs. **Figure 9.12a** shows a map of the 16,569-bp human mitochondrial genome. While human mitochondrial genomes are circular, diverse arrangements exist in other organisms. For example, some mitochondrial genomes are linear, including those of certain algae, protozoans, and fungi. Finally, the mitochondria of many fungi and flowering plants contain, in addition to the mitochondrial genome, small circular or linear plasmids (↔ Section 4.2).

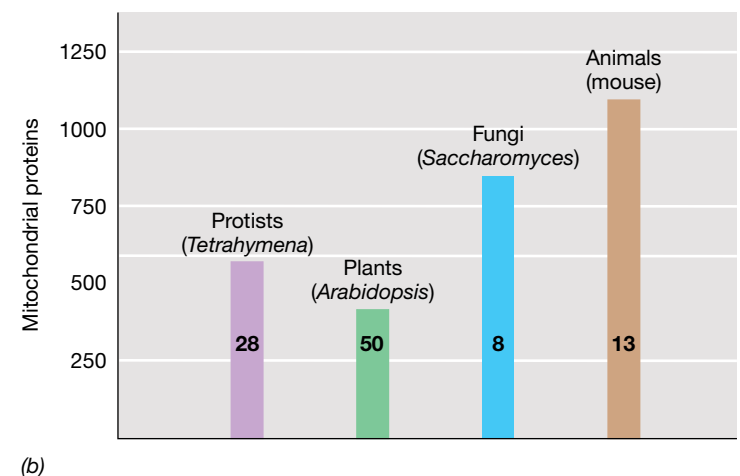
Mitochondria require many more proteins than their genome encodes (in particular, proteins needed for translation), and thus many mitochondrial proteins are encoded by genes in the nucleus. The yeast mitochondrion contains as many as 800 different proteins in its proteome (all the proteins encoded by a genome; Section 9.10). However, only eight (~1%) of them are encoded by the yeast mitochondrial genome, the remaining proteins being encoded by nuclear genes (Figure 9.12b). However, the nuclear-encoded proteins required for translation and energy generation in mitochondria are more closely related to their counterparts in *Bacteria* than to those in the eukaryotic cytoplasm, consistent with both the evolutionary history of the mitochondrion and with a scenario—like that seen in the chloroplast—of genes having migrated from the original endosymbiont to the host cell nucleus.



(a)

Figure 9.12 Map of the human mitochondrial genome and the mitochondrial proteome. (a) The genome encodes rRNAs, 22 tRNAs, and several proteins. Arrows show direction of transcription for genes of a given color, and the three-letter amino acid designations for tRNA genes are also shown. The 13 protein-encoding

genes are in green. *Cytb*, cytochrome *b*; ND1–6, components of the NADH dehydrogenase complex; COI–III, subunits of the cytochrome oxidase complex; ATPase 6 and 8, polypeptides of the mitochondrial ATPase complex. The two promoters are in the region called the D loop, which is also involved in DNA



(b)

replication. Inset: Transmission electron micrograph of a mitochondrion (credit, D. W. Fawcett). (b) Mitochondrial proteomes. The numbers in each colored bar are the number of proteins encoded on the mitochondria of some model eukaryotes.

Genomes and Introns in Some Microbial Eukaryotes

Apart from the human pathogen *Trichomonas*, which contains almost three times more genes than human cells, parasitic eukaryotic microorganisms typically have relatively small genomes of 10–40 Mbp containing between 4000 and 11,000 genes. For example, *Trypanosoma brucei*, the agent of African sleeping sickness, has 11 chromosomes, 35 Mbp of DNA, and almost 11,000 genes. The four species of *Plasmodium* that infect humans (causing malaria, [↻](#) Section 33.5) have genomes ranging from 23 to 27 Mbp arranged in 14 chromosomes and about 5500 genes.

As in *Bacteria*, the smallest eukaryotic genome belongs to an endosymbiont. Known as a *nucleomorph*, it is the degenerate remains of a eukaryotic endosymbiont of a certain green alga that has acquired the ability to photosynthesize by secondary endosymbiosis ([↻](#) Section 18.1). Nucleomorph genomes range from about 0.37 to 0.85 Mbp. The smallest genome in a parasitic eukaryote belongs to *Encephalitozoon intestinalis*, an intracellular pathogen of humans and other animals. *E. intestinalis* even lacks mitochondria, and although its haploid genome contains 11 chromosomes, the genome size is only 2.3 Mbp with approximately 1800 genes (Table 9.5); this is smaller than many bacterial genomes (Table 9.1).

The baker's yeast *Saccharomyces cerevisiae* is widely used as a model eukaryote and its genome contains 16 chromosomes (13.4 Mbp of DNA). Yeast has approximately 6000 ORFs, which is fewer than that of some genomes of *Bacteria* (Tables 9.1 and 9.5). How many of these yeast genes are actually essential? This question has been addressed by systematically inactivating each gene in turn with *knockout mutations* (mutations that completely inactivate genes, [↻](#) Section 12.4). Knockout mutations cannot normally be obtained in essential genes in a haploid organism. However, yeast can be grown in both diploid and haploid states ([↻](#) Section 18.9). By generating knockout mutations in diploid cells and then investigating whether they can also exist in haploid cells, it is possible to determine whether a particular gene is essential for cell viability. Using knockout mutations, it has been shown that around 900 yeast ORFs (17% of its genome) are absolutely essential. Note that this number of essential genes is much greater than the approximately 300 genes (Section 9.3) estimated to be the minimal number required in a bacterial cell.

Being a eukaryote, the yeast genome contains introns ([↻](#) Section 4.6). However, the total number of introns in the protein-encoding genes of yeast is a mere 225. Most yeast genes that contain introns have only a single small intron near the 5' end of the gene. This situation differs greatly from that seen in more complex eukaryotes ([Figure 9.13](#)). For example, in the worm *Caenorhabditis elegans*, the average gene has five introns, and in the fruit fly *Drosophila*, the average gene has four. Introns are also common in the genes of plants, averaging around four per gene. The model flowering plant *Arabidopsis* averages five introns per gene, and over 75% of *Arabidopsis* genes have introns. In humans almost all protein-encoding genes have introns, and it is common for a single gene to have 10 or more. Moreover, introns in human genes are typically much longer than exons, the DNA that actually encodes proteins. Indeed, exons make up only about 1% of the human genome, whereas introns account for 24%. The remaining DNA is made up of repetitive sequences, noncoding RNA, and regulatory regions. Nevertheless, as we will see later, much of this DNA is indeed functional (Section 9.14).

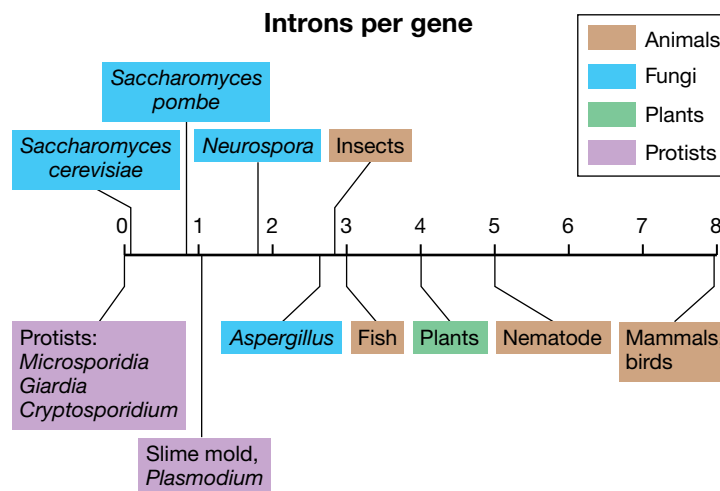


Figure 9.13 Intron frequency in the genes of different eukaryotes. The average number of introns per gene is shown for a range of eukaryotic organisms.

MINIQUIZ

- What is unusual about the genes that encode mitochondrial proteins?
- What do chloroplast genomes typically encode?
- What is unusual about the genome of the eukaryote *Encephalitozoon*?

II • The Evolution of Genomes

In addition to revealing how genes function and how organisms interact with their environments, comparative genomics can illuminate evolutionary relationships between organisms. Reconstructing evolutionary trees from genome sequences helps to distinguish between primitive and derived characteristics and can resolve ambiguities in phylogenetic trees based on analyses of a single gene, such as an rRNA gene ([↻](#) Section 13.3). Genomics is also a link to understanding early life forms and may eventually help answer the most fundamental of all questions in biology: How did life originate?

9.5 Gene Families, Duplications, and Deletions

Genomes from both prokaryotic and eukaryotic cells often contain multiple copies of genes that are related in sequence due to shared evolutionary ancestry; such genes are called *homologous genes*, or **homologs**. Groups of gene homologs are called **gene families**. Not surprisingly, larger genomes tend to contain more individual members from a particular gene family than do smaller genomes. *Gene duplication* is thought to be a major driving force behind the evolution of gene families and the organisms that contain them.

Paralogs, Orthologs, and Gene Duplications

Comparative genomics shows that many genes have arisen by duplication of other genes. Such homologs may be subdivided,

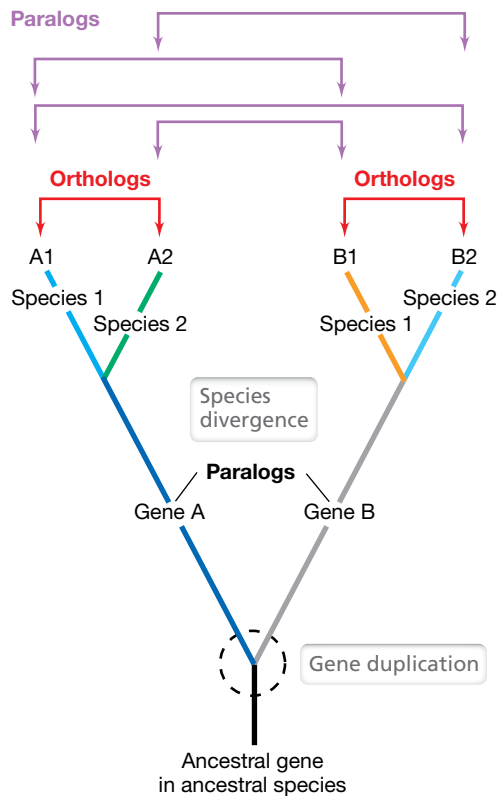
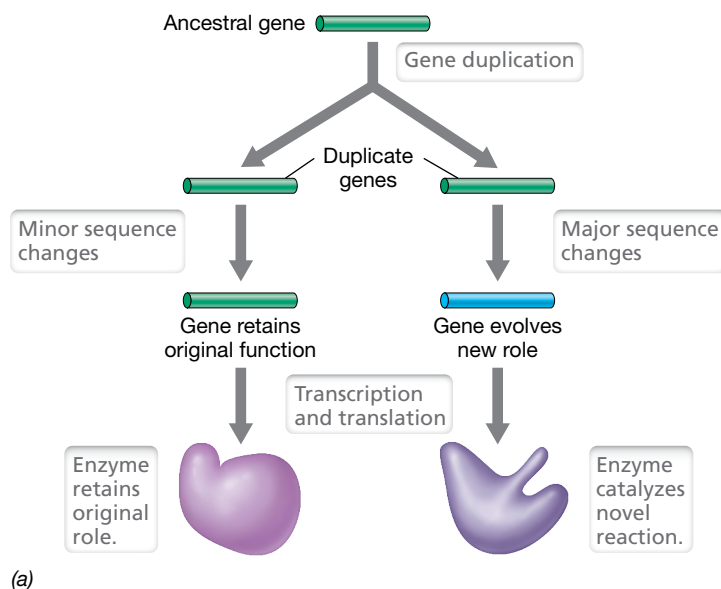


Figure 9.14 Orthologs and paralogs. This family tree depicts an ancestral gene that duplicated and diverged into two paralogous genes, A and B. Later, the ancestral species diverged into species 1 and species 2, both of which have genes for A and B (designated A1 and B1 and A2 and B2, respectively). Each such pair are paralogs. However, because species 1 and 2 are now separate species, A1 is an ortholog of A2 and B1 is an ortholog of B2.

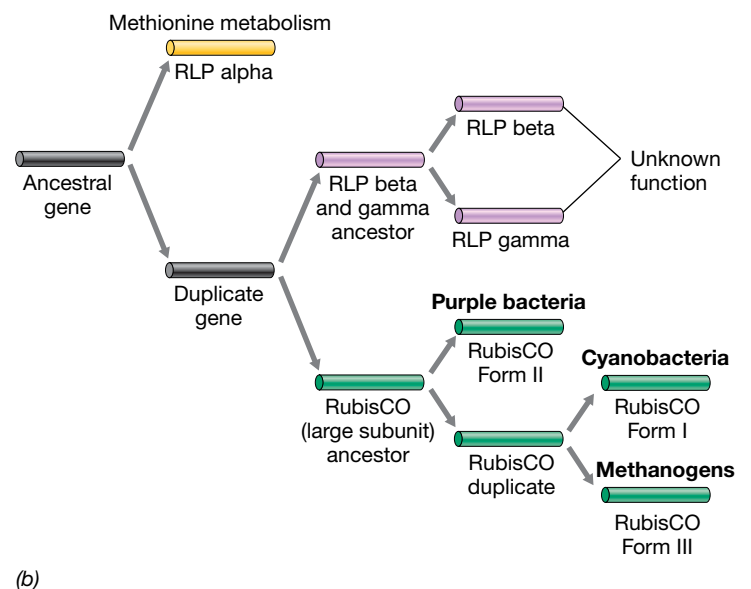


(a)

Figure 9.15 Evolution by gene duplication. (a) The principle of gene duplication. After duplication, the “spare” copy of a gene is free to evolve to encode a new function. (b) The RubisCO (*rbcL*) family of genes. The large subunit of the enzyme RubisCO that fixes CO₂

depending on their origins. Genes whose similarity is the result of gene duplication at some time in the evolution of an organism are called **paralogs**. Genes found in one organism that are similar to genes in another organism because of descent from a common ancestor are called **orthologs** (Figure 9.14). An example of paralogous genes is evident in the genes that encode variant human lactate dehydrogenases (LDH), which recycle NAD⁺ through the conversion of pyruvate to lactate in tissues where oxygen is absent. These variants, called *isoenzymes*, are structurally distinct yet all highly related and carry out the same enzymatic reaction. By contrast, the corresponding LDH from the lactic acid bacterium *Lactobacillus* would be said to be orthologous to the human LDH isoenzymes. Thus, gene families contain both paralogs and orthologs.

If a segment of duplicated DNA is long enough to include an entire gene or group of genes, the organism with the duplication has multiple copies of these particular genes. After duplication, one of the duplicates is free to undergo spontaneous mutations while the other copy continues to supply the cell with the original function (Figure 9.15a). In this way, evolution can “experiment” with one copy of the gene. Such gene duplication events, followed by diversification of one copy, are thought to be the major events that fuel microbial evolution. Genomic analyses have revealed many examples of protein-encoding genes that were clearly derived from gene duplication. Figure 9.15b shows this for the enzyme RubisCO, a key enzyme of autotrophic metabolism (Section 14.5): An ancestral RubisCO gene gave rise to enzymes with different but related catalytic activities.



(b)

during photosynthesis is an ancestor of three closely related forms (I, II, and III) that all retain the original function (green bars). However, RubisCO is in turn derived from an ancestral gene (black bars) of unknown function that divided to produce a gene encoding an

enzyme in methionine metabolism (yellow bar) and several genes whose function is still unknown (purple bars). RLP, RubisCO-like protein.

Entire Genome Duplications

Duplications of genetic material may include just a handful of bases, one or more genes, or even whole genomes. For example, comparison of the genomes of the yeast *Saccharomyces cerevisiae* and other fungi indicates that the ancestor of *S. cerevisiae* duplicated its entire genome. This was followed by extensive deletions that eliminated much of the duplicated genetic material. Analysis of the genome of the model plant *Arabidopsis* suggests that there were one or more whole genome duplications in the ancestor of the flowering plants, as well.

Some bacterial genomes show evidence of having once been duplicated. The distribution of duplicated genes and gene families in the genomes of *Bacteria* and *Archaea* suggest that several duplications have occurred. For example, the soil bacterium *Myxococcus* has a genome of 9.1 Mbp. This is approximately twice that of the genomes of its close relatives. However, while genomic analyses point to frequent small-scale gene duplications being rather common in prokaryotic cells, entire genome duplications appear to be rather rare. Conversely, in parasitic bacteria, frequent successive gene *deletions* have eliminated genes no longer needed for a parasitic lifestyle. This has been the driving force behind the unusually small genomes of many endosymbiotic bacteria. Insect endosymbionts have carried this theme to an extreme and show the smallest of all cellular genomes (Section 9.3, Table 9.1, and Figure 9.8).

MINIQUIZ

- What is a homologous gene?
- What is a gene family?
- Contrast gene paralogs with gene orthologs.

9.6 Horizontal Gene Transfer and the Mobilome

Genetic traits are transferred from one generation to the next by what's called a vertical process (from mother to daughter). However, vertical transfer in *Bacteria* and *Archaea* can be embellished by **horizontal gene transfer** (sometimes called *lateral gene transfer*), and this can complicate the analysis of genomes. Horizontal transfer refers to gene transfer from one cell to another by means other than the vertical process (Figure 9.16). In prokaryotic cells, at least three mechanisms of horizontal gene transfer are known: *transformation*, *transduction*, and *conjugation*, and these are discussed in detail in Chapter 11.

Detecting Horizontal Gene Flow

Horizontal gene transfers can be detected in genomes once the genes have been annotated (Section 9.2). The presence of genes that encode proteins typically found only in distantly related species is one signal that the genes originated from horizontal transfer. However, another clue to horizontally transferred genes is the presence of a stretch of DNA whose guanine/cytosine (GC) content or codon bias (↻ Section 4.9) differs significantly from that of the rest of the genome. Using these clues, many likely examples of horizontal transfer have been documented in the genomes of both *Bacteria* and *Archaea*. A classic example exists

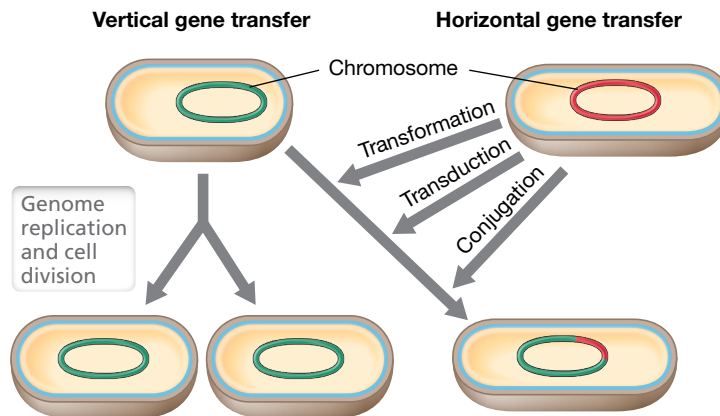


Figure 9.16 Vertical versus horizontal gene transfer. Vertical gene transfer occurs when cells divide. Horizontal gene transfer occurs when a donor cell contributes genes to a recipient cell. In *Bacteria* and *Archaea*, horizontal transfer occurs through one of three mechanisms: transformation, transduction, and conjugation.

with the thermophilic bacterium *Thermotoga maritima*, whose genome was shown to contain over 400 genes (greater than 20% of the entire genome) of archaeal origin. Of these genes, 81 were found in discrete clusters. The latter is a strong indication that the archaeal genes were transferred to *T. maritima* by horizontal gene flow from thermophilic *Archaea* that share its hot habitat.

For horizontal gene transfer to be readily detectable by comparative genomics, the phylogenetic difference between the organisms must be rather large. For example, several eukaryotic genes have been found in *Chlamydia*, a bacterial pathogen that causes both a sexually transmitted disease and an eye infection called trachoma in humans. In particular, two genes encoding histone H1-like proteins have been found in the *Chlamydia trachomatis* genome, suggesting horizontal transfer from a eukaryotic source, possibly even its human host. Note that this is conceptually the reverse of mitochondrial and chloroplast gene flow in which genes from the ancestor of the mitochondrion and the chloroplast were transferred to the eukaryotic nucleus (Section 9.4). In contrast to the *Chlamydia* findings, more subtle horizontal transfers, such as those between fairly closely related organisms, can be easily hidden and therefore missed during genome annotation.

Horizontally transferred genes typically encode metabolic functions distinct from the core molecular processes of DNA replication, transcription, and translation, and may account for observed similarities of metabolic genes in *Archaea* and *Bacteria*. In addition, there are several examples of virulence genes of pathogenic bacteria that have been transferred by horizontal means. It is clear that prokaryotic cells are actively exchanging genes in nature, and the process likely functions to “fine-tune” an organism’s genome to a particular ecological situation or habitat. Nevertheless, it is necessary to be cautious when invoking horizontal gene transfer to explain the distribution of genes in a given organism. Homologs of the genes in question may be present in close relatives whose genome sequences are not yet available.

The Mobilome

Horizontal gene flow is facilitated by the **mobilome**, the sum total of all mobile genetic elements in a genome. The mobilome

includes plasmids, prophages (integrated virus genomes), integrons, insertion sequences, and transposons (↻ Section 11.11). Integrons are genetic elements that can capture gene cassettes (mobile DNA containing a recombination site) through the activity of an enzyme called integrase. All constituents of the mobilome also play important roles in genome evolution by shuffling genes between species.

Transposons are mobile genetic elements that move between different host DNA molecules, including chromosomes, plasmids, and viruses (Figure 9.17), by the activity of an enzyme called *transposase* (↻ Section 11.11). In doing so, transposons may pick up and horizontally transfer genes encoding various characteristics, including resistance to antibiotics and production of toxins. Because of this tendency, transposable elements are a strong driver of genome evolution. However, transposons may also mediate a variety of large-scale chromosomal changes (Figure 9.17). Bacteria undergoing rapid evolutionary change often contain relatively large numbers of mobile elements, especially *insertion sequences*, simple transposable elements whose genes encode only transposition. Recombination among identical elements generates chromosomal rearrangements such as deletions, inversions, or translocations, and these provide a source of genomic diversity upon which natural selection can act. Thus, chromosomal rearrangements that accumulate in bacteria during stressful growth conditions are often flanked by repeats or insertion sequences.

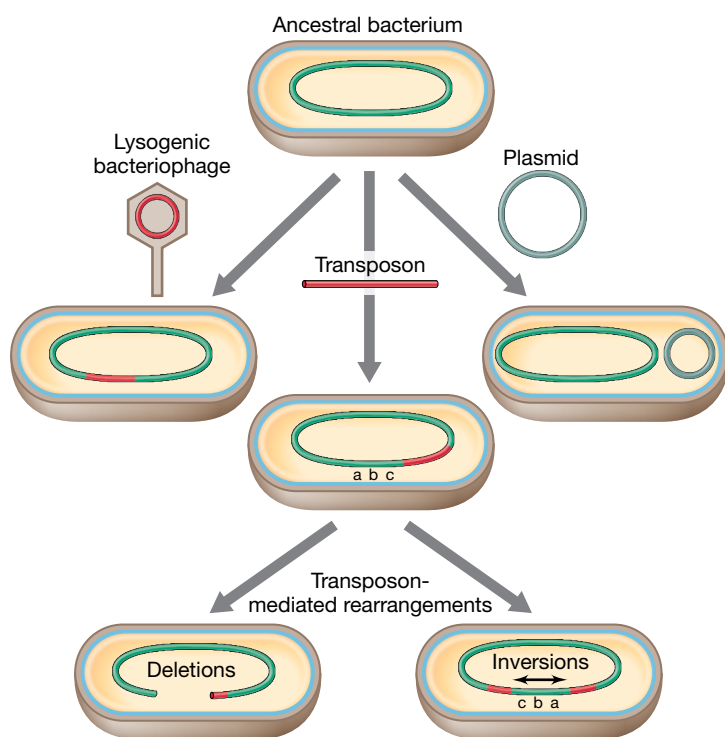


Figure 9.17 Mobile elements promote genome evolution. A variety of mobile genetic elements can move from one organism to another, thus adding genes to the genome of the recipient. The most common of these are plasmids, bacteriophages, and transposons. When a transposon moves, chromosomal rearrangements, such as deletions and inversions of DNA neighboring the transposon, may be mediated by the activity of the transposase.

Chromosomal rearrangements due to insertion sequences have apparently contributed to the evolution of several bacterial pathogens, increasing their pathogenic potential. In the genera *Bordetella*, *Yersinia*, and *Shigella*, the more highly pathogenic species show a much greater content of insertion sequences. For example, *Bordetella bronchiseptica*, which causes a bronchitis-like cough in dogs and other domestic animals, has a genome of 5.3 Mbp but carries no known insertion sequences. Its more pathogenic relative, *Bordetella pertussis*, the causative agent of whooping cough in humans (↻ Section 30.3), has a smaller genome (4.1 Mbp) but has more than 260 insertion sequences. Comparison of these genomes suggests that the insertion sequences are responsible for major genome rearrangements, including the deletions that reduced the genome size in *B. pertussis*, possibly as a means for streamlining its genome to allow for more virulence factors to be encoded.

The mobilome, especially the horizontal transfer of transposable elements and prophages, is also responsible for the presence of *chromosomal islands* in certain strains. We turn our attention to these islands now and examine their relationship to the rest of the genome.

MINIQUIZ

- Which class of genes is rarely transferred horizontally? Why?
- List the major mechanisms by which horizontal gene transfer occurs in *Bacteria* and *Archaea*.
- How might transposons be especially important in the evolution of pathogenic bacteria?

9.7 Core Genome versus Pan Genome

One of the most important concepts to emerge from comparing the genome sequences of multiple strains of the same bacterial species is the distinction between the **pan genome** and the **core genome**. The *core genome* is that shared by all strains of a given species, whereas the *pan genome* of an organism includes the core plus genes not shared by all strains of that species. As we have seen, horizontal gene transfer of entire genetic elements such as plasmids, viruses, or transposable elements is widespread. Consequently, there may be major differences in the total amount of DNA and the suite of accessory capabilities (virulence, symbiosis, biodegradation, and the like) between strains of a single bacterial species. In other words, one could say that the core genome is typical of the species as a whole, whereas the other components that make up the pan genome, frequently including mobile elements, are restricted to particular strains within a species.

It is difficult to define the size of the pan genome precisely because it increases as the genomes of more strains of a species are sequenced. In some cases, such as the enteric bacteria *Escherichia coli* and *Salmonella enterica*, many different isolates have been found that carry a wide range of different mobilome elements. Consequently the pan genome is extremely large. Figure 9.18 illustrates the pan genome for serovars (strains distinguished by their immunological properties) of the important human pathogen

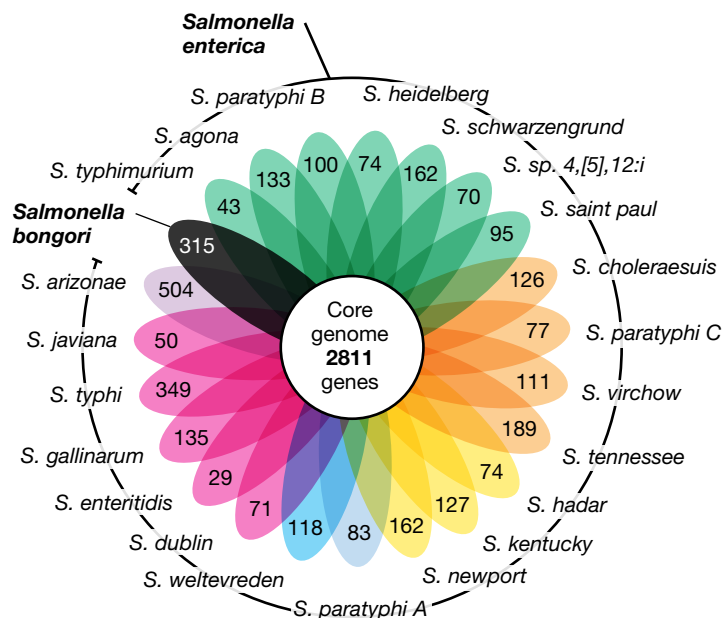


Figure 9.18 Flowerplot of the *Salmonella enterica* pan genome.

A “flowerplot” of gene families in serovars (strains) of the gram-negative pathogenic bacterium *Salmonella enterica* (the names surrounding the flowerplot are immunologically unique serovars [S.] of *S. enterica*). The figure presents the average number of gene families found in each genome as being unique to each serovar. *Salmonella bongori* is a species distinct from *S. enterica*. Serovar 4,[5],12:i has not yet been named. Data from Jacobsen, A., R.S. Hendriksen, F.M. Aarestrup, D.W. Ussey, and C. Friis. 2011. The *Salmonella enterica* pan-genome. *Microb Ecol* 62: 487–504.

S. enterica (salmonellosis) depicted in a “flowerplot” schematic. As is evident, although all strains contain at least 2811 genes, some contain several hundred more (Figure 9.18).

Chromosomal Islands

Comparison of the core and pan genomes of a particular species with their close relatives sometimes reveals extra blocks of genetic material that are part of the chromosome, rather than existing as plasmids or integrated viruses. These so-called **chromosomal islands**, or *genomic islands*, contain clusters of genes for specialized functions that are not essential for survival (Figure 9.19). Consequently, two strains of the same bacterial species may show significant differences in genome size. Chromosomal islands are presumed to be of “foreign” origin based on several lines of evidence. First, these extra genes are often flanked by inverted repeats, implying that the whole region was inserted into the chromosome by transposition (Section 9.6) at some point. Second, the base composition and codon bias (Table 9.3) in chromosomal islands often differ significantly from that of the genome proper. Third, chromosomal islands are found in some strains of a particular species but not in others.

Chromosomal islands in pathogenic bacteria have drawn the most attention because in many cases genes encoding disease-associated functions have been linked to an island. However, chromosomal islands are also known that encode the biodegradation of pollutants such as aromatic hydrocarbons and herbicides. In addition, many of the genes essential for the symbiotic relationship of species of the nitrogen-fixing bacterium *Rhizobium* with

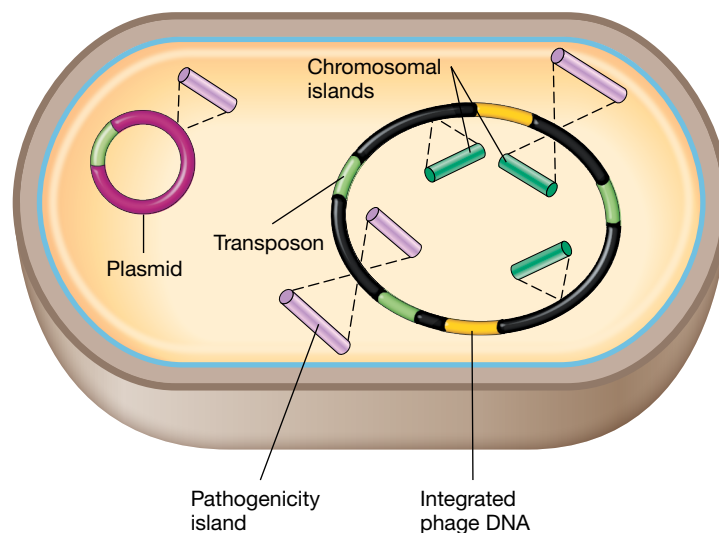


Figure 9.19 Possible genome insertions. The core genome is represented by the black regions of the chromosome and is present in all strains of a species. Each colored wedge indicates a single insertion. Where two wedges emerge from the same location, they represent alternative chromosomal islands that can insert at that site. However, only one insertion can be present at a given location. Plasmids, like the chromosome, may also have insertions.

the root nodules of plants (↔ Section 23.3) are carried in chromosomal islands. Perhaps the most unique chromosomal island is the magnetosome island of the bacterium *Magnetospirillum*; this DNA fragment carries genes that encode the formation of magnetosomes, intracellular magnetic particles that orient the cell in a magnetic field and influence the direction of its movement (↔ Section 2.8).

Some chromosomal islands carry a gene encoding an integrase enzyme, suggesting that the islands move within the genome in a manner similar to conjugative transposons (Section 9.6). Chromosomal islands are typically inserted into a gene for a tRNA; however, because the target site is duplicated upon insertion, an intact tRNA gene is regenerated during the insertion process. In a few cases, transfer of a whole chromosomal island between related bacteria has been demonstrated in the laboratory, and transfer can presumably occur by any of the mechanisms of horizontal transfer (Figure 9.17). It is thought that after insertion into the genome of a new host cell, chromosomal islands gradually accumulate mutations, and hence, over many generations, chromosomal islands tend to lose their ability to move.

Pathogenicity Islands and the Evolution of Virulence

Comparison of the genomes of pathogenic bacteria with those of their harmless or less virulent relatives often reveals chromosomal islands that encode *virulence factors*: special proteins, toxins, enzymes, or other molecules or structures that facilitate disease symptoms (Chapter 25). Some virulence genes are carried on lysogenic bacteriophages or plasmids (↔ Sections 8.7 and 11.8); however, many others are clustered in chromosomal regions called **pathogenicity islands**.

The pathogenicity islands of uropathogenic strains of *E. coli* have been particularly well studied (Figure 9.20). Only a few strains

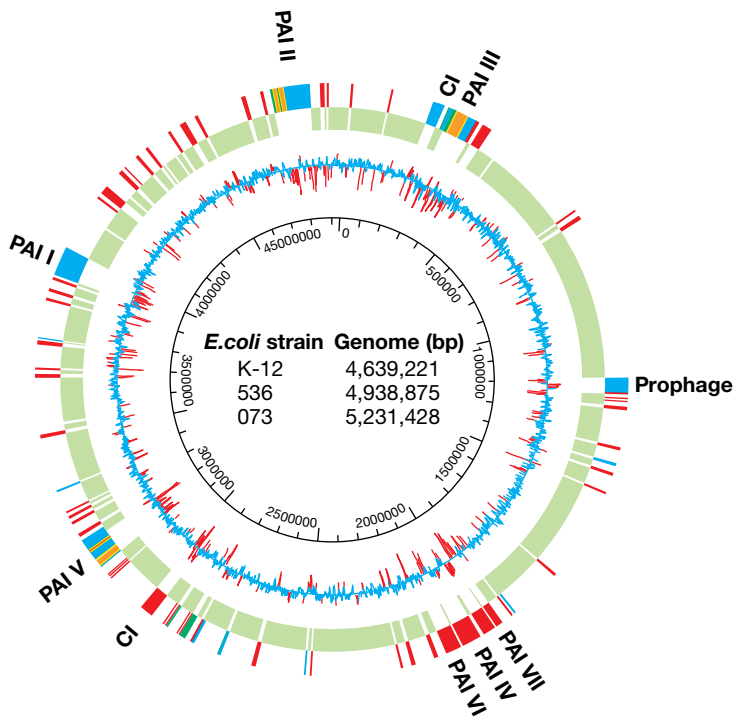


Figure 9.20 Pathogenicity islands in *Escherichia coli*. Genome comparisons of uropathogenic *E. coli* strain 536, uropathogenic strain 073, and the nonpathogenic strain K-12. The uropathogenic strains (urinary tract pathogens) contain pathogenicity islands and thus their chromosomes are larger than that of K-12. The inner circle represents nucleotide base pairs. The jagged circle shows the DNA GC distribution; regions where GC content varies dramatically from the genome average are in red. The outermost circle is a three-way genomic comparison: green, genes common to all strains; red, genes present in the pathogenic strains only; blue, genes found only in strain 536; orange, genes of strain 536 present in a different location in strain 073. Some very small inserts are deleted for clarity. PAI, pathogenicity island; CI, chromosomal island. Prophage, DNA from a temperate bacteriophage. Note the correlation between chromosomal islands and skewed GC content. Data adapted from *Proc. Natl. Acad. Sci. (USA)* 103: 12879–12884 (2006).

of *E. coli* cause bladder and kidney infections, but those that do contain pathogenicity islands that encode a variety of virulence factors including adhesins that facilitate binding to host tissues and a capsule that helps cells evade immune surveillance. For example, although *E. coli* strain K-12 (a harmless laboratory strain) and strain 073 (a urinary tract pathogen) share a core genome of some 4.6 Mbp, strain 073 contains 11% more DNA than strain K-12, much of which is devoted to its pathogenic lifestyle.

Small pathogenicity islands that encode a series of virulence factors are also present in certain strains of the gram-positive pathogenic bacterium *Staphylococcus aureus* (skin infections, boils, and the like) and can be moved between cells in temperate bacteriophages by transduction, a major mechanism of horizontal gene transfer (see Section 11.7). The islands are smaller than the phage genome, and when the islands excise from the chromosome and replicate, they induce the formation of defective phage particles that carry the genes for the islands but are too small to carry the phage genome. In this way, when strains of *S. aureus* that lack the islands are infected, they are not killed by the phage but instead acquire the islands and become more potent pathogens.

MINIQUIZ

- What is the difference between core genome and pan genome?
- What is a chromosomal island and how can one be identified as being of foreign origin?
- What is a pathogenicity island and how does one move between bacterial species?

III • Functional Omics

Despite the major effort required to generate an annotated genome sequence, the net result is simply a “list of parts.” To understand how a cell *functions*, we need to know more than which genes are present. We must also understand (1) gene expression, (2) the function of gene products, (3) the activity of the proteins made, and (4) the metabolites produced during growth.

In analogy to the term “genome,” the entire complement of RNA, proteins, or metabolites produced under a given set of conditions is known as the **transcriptome**, **translatome**, and **metabolome**, respectively. Adding the suffix “omic” denotes their corresponding areas of study. **Table 9.6** summarizes some of the “omics” terminology used in microbiology today.

9.8 Metagenomics

Microbial communities contain many microbial species, many of which have never been cultured or formally identified. **Metagenomics** is the science that analyzes pooled DNA or RNA

TABLE 9.6 Some omics terminology

DNA	Genome the total complement of genetic information of a cell or a virus
	Metagenome the total genetic complement of all the cells present in a particular environment
	Epigenome the total number of possible epigenetic changes
	Methylome the total number of methylated sites on the DNA (whether epigenetic or not)
	Mobilome the total number of mobile genetic elements in a cell
RNA	Transcriptome the total RNA produced in an organism under a specific set of conditions
Protein	Proteome the total set of proteins encoded by a genome; sometimes also used in place of <i>translatome</i>
	Translatome the total set of proteins present under specified conditions
	Interactome the total set of interactions between proteins (or other macromolecules)
	Secretome the total set of proteins secreted by a cell
Metabolites	Metabolome the total complement of small molecules and metabolic intermediates
	Glycome the total complement of sugars and other carbohydrates
Organisms	Microbiome the total complement of microorganisms in an environment (including those associated with a higher organism)
	Virome the total complement of viruses in an environment
	Mycobiome the total complement of fungi in a natural environment

from an environmental sample containing organisms that have not been isolated and identified (Figure 9.21). Just as the total gene content of an organism is its *genome*, so the total gene content of a microbial community is its **metagenome** (Table 9.6). In addition to metagenomic analyses based on DNA sequencing, analyses based on RNA or proteins (metatranscriptomics or metaproteomics, respectively) may be used to explore the patterns of gene expression in natural microbial communities. With today's molecular technology, these studies can even be done on individual cells (Section 9.12).

Examples of Metagenomic Studies

Several environments have been surveyed by large-scale metagenome sequencing projects. Extreme environments, such as highly acidic runoff waters from mining operations, tend to have low microbial species diversity. Consequently it has been possible to isolate community DNA (and metabolites, Section 9.11) and assemble much of it into nearly complete individual genomes. Conversely, complex environments such as fertile soils or aquatic environments are much more challenging, and complete genome assemblies here are much more difficult. Nonetheless, a surprising finding that has emerged from metagenomic studies thus far is that most genes recovered from natural habitats do not originate from cells but from viruses. This is discussed further in Chapter 10 where we consider the genomics and phylogeny of viruses.

Even if complete genomes cannot be assembled from environmental DNA, much useful information can be derived from metagenomic surveys. For example, environments can be analyzed for the presence and distribution of specific microbial groups. These vary greatly in relative abundance in different environments, and Figure 9.21*b* illustrates this for subgroups of *Proteobacteria* (a major

phylum of gram-negative *Bacteria*, Chapter 16) at a sampling site in the Pacific Ocean near the Hawaiian Islands. Light, oxygen, nutrients, and temperature all change with depth in a water column, and these factors can be correlated with proteobacterial subgroups to show which are most competitive at each depth (Figure 9.21*b*). One curious observation that has emerged from such metagenomic studies is that much cellular DNA in natural habitats does not reside in *living* cells. Around 50–60% of the DNA in the oceans is extracellular DNA present in deep-sea sediments. Presumably this was DNA deposited when dead organisms from the upper layers of the ocean sank to the bottom and lysed. Because nucleic acids are major reservoirs of phosphate, marine sediment DNA is thought to be a major component of the global phosphorus cycle.

Metagenomics and “Biome” Studies

The human body is estimated to contain about 10 trillion (10^{13}) cells, but each of us also carries around ten times more prokaryotic cells than human ones. This collection of prokaryotic cells is called the human *microbiome* (Chapter 24). Most of these organisms inhabit the large intestine, with the majority belonging to one of two phylogenetic groups of *Bacteria*, the *Bacteroidetes* and the *Firmicutes* (Chapter 16). A fascinating finding is that the composition of the gut microbiome correlates with obesity in both humans and experimental mouse models. The data show that the higher the proportion of *Firmicutes* (mostly species of *Clostridium* and relatives) in the gut, the more obese is the human or mouse. A suggested mechanism behind this finding is that fermentative species of *Firmicutes* convert more dietary fiber into fatty acids than can be absorbed by the host (⇨ Section 24.8). In this way, the obese host gets more organic carbon than the thinner host from the same amount of food.

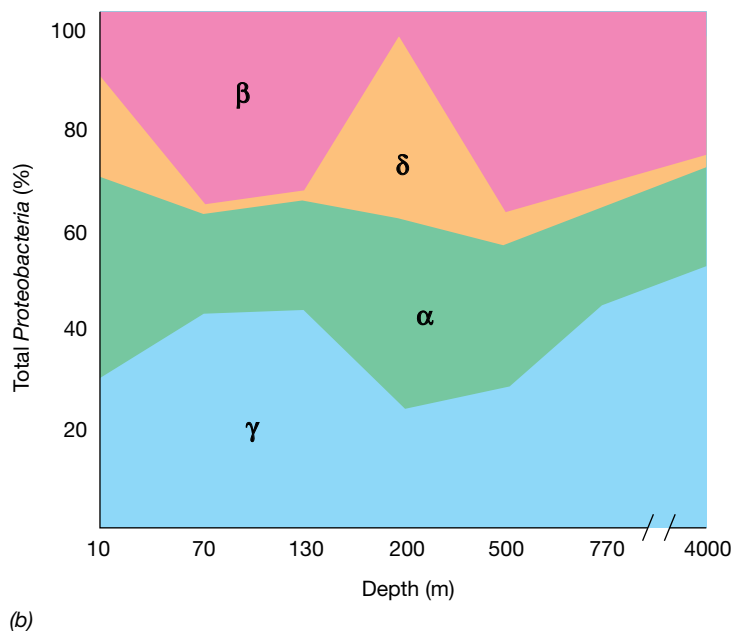
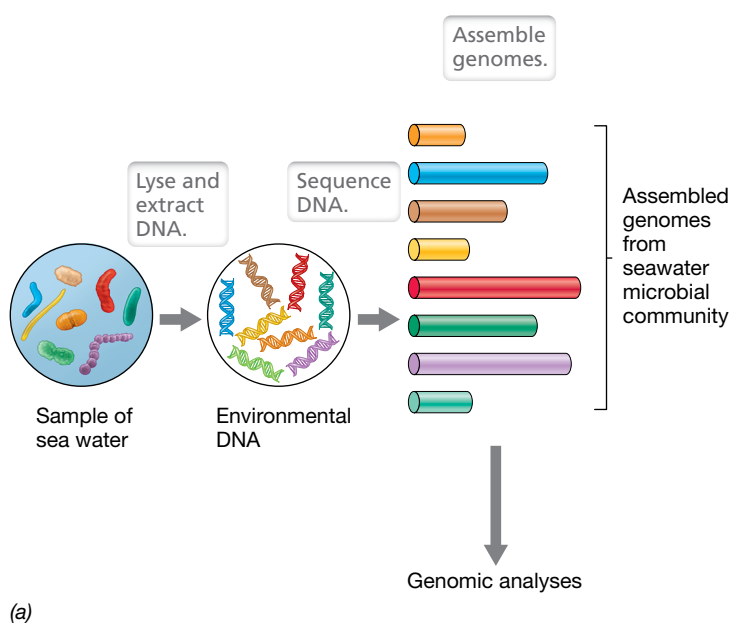


Figure 9.21 Metagenomics and the microbiome. (a) Isolation, sequencing, and identification of DNA from a sample of seawater. (b) *Proteobacteria* in the ocean. The distribution with depth of the major subgroups (alpha α , beta β , gamma γ , and delta δ) of *Proteobacteria* in the Pacific Ocean is shown. Many other types of bacteria are also present (not shown). Data adapted from Kembel, S.W., J.A. Eisen, K.S. Pollard, and J.L. Green. 2011. *PLoS One* 6: e23214.

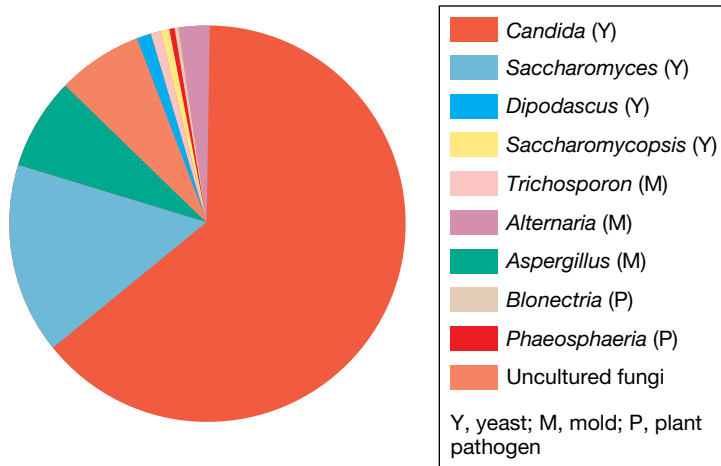


Figure 9.22 The mouse mycobiome. The data shown represent the relative amount of different fungal genera of the mouse intestine. The pie chart shows the most common fungi present are yeasts. Data adapted from Iliev, I.D., et al. *Science* 336: 1314–1317 (2012).

Recent surveys of the human and mouse gut microbiome have also revealed the rather surprising finding that over 60 species of fungi are present (Figure 9.22). These constitute the gut *mycobiome* (the prefix “myco” means fungal). Many fungi, typically non-pathogenic yeasts, inhabit the skin, the oral cavity, and virtually all moist surfaces on the human body. Many of these are common and generally harmless yeasts, such as *Saccharomyces*, *Cladosporium*, and most species of *Candida*. Most of these also are found in the gut, although some gut fungi—such as *Aspergillus* and *Trichosporon*—are potential serious pathogens. Moreover, although gut fungi constitute less than 1% of the microbiome, it is known that certain conditions such as inflammatory bowel disease and some cases of obesity correlate strongly with specific fungal populations. Thus metagenomics holds great promise for exploring possible connections between specific microbial populations and specific diseases in humans and other animals. Moreover, in cases where a clear cause-and-effect relationship is strongly suspected, metagenomics also holds great promise as a clinical tool for making medical diagnoses.

MINIQUIZ

- What is a metagenome? The mycobiome?
- How is a metagenome analyzed?

9.9 Gene Chips and Transcriptomics

Once a genome sequence is available, the sequence can be used to synthesize gene chip devices that can be used to detect specific microbes, determine genome differences between closely related strains of the same species (for example, the presence of chromosomal islands), identify sequences specifically bound by DNA-binding proteins, and measure gene expression (transcription). *Transcriptomics* refers to the global study of transcription and is done by monitoring the total RNA generated under a chosen

growth condition. In the case of genes whose role is still unknown, discovering the conditions under which they are transcribed may yield clues to their function. Two main approaches are used in transcriptomics: *microarrays* and *RNA-Seq*.

Microarrays and the DNA Gene Chip

Microarrays are small, solid supports to which genes or, more often, oligonucleotides corresponding to segments of genes are fixed and arrayed spatially in a known pattern; they are often called **gene chips** (Figure 9.23a). Microarrays measure the DNA or RNA that hybridizes to the DNA sequences on the chip. When DNA is denatured (that is, the two strands are separated), the single strands can form hybrid double-stranded molecules with other nucleic acid molecules by complementary or almost complementary base pairing (Figure 9.23b; see Section 12.1). This process is called *nucleic acid hybridization*, or **hybridization** for short, and is widely used in detecting, characterizing, and identifying segments of DNA or RNA. The single-stranded segments of nucleic acid, whose identity is already known, are called **nucleic acid probes** or, simply, *probes*. To detect hybridization to the probes, the nucleic acid added to the chip must be labeled with a fluorescent dye and then the hybridized chip is scanned with a laser fluorescence detector that measures which of the probes contain hybridized DNA (Figure 9.23b, c).

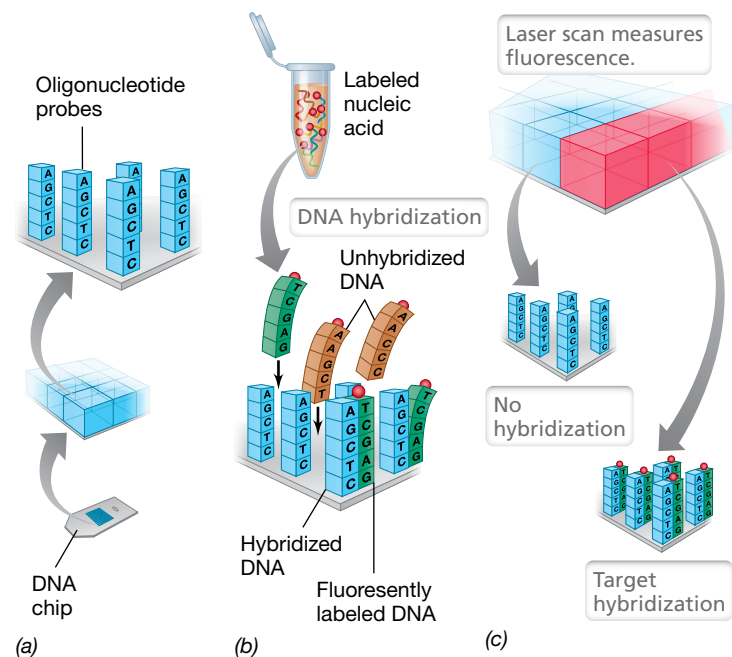


Figure 9.23 DNA chip design and application. (a) DNA chip design. Short single-stranded oligonucleotides (probes) corresponding to each gene in an organism or to diagnostic sequences corresponding to numerous organisms are synthesized and affixed at known locations to make a microarray. (b) Microarray hybridization. The presence of specific DNA or RNA (in the form of cDNA) is assayed by hybridizing fluorescently labeled samples (DNA or cDNA) to the DNA probes on the chip. Labeled DNA or cDNA will bind to the probes on the chip if they possess sequence complementarity. (c) Analysis of microarray hybridization. A scanning laser is used to identify regions of the chip where labeled nucleic acid has bound to the probes.

Gene chips are typically about 1 to 2 cm and are inserted into a plastic holder that can easily be manipulated (Figure 9.24a); each chip holds thousands of different DNA fragments. In practice, each gene is usually represented more than once in the array to increase reliability. Whole genome arrays contain DNA segments that cover the entire genome of an organism. For example, a chip that covers the entire human genome (Figure 9.24a) can analyze over 47,000 human transcripts and has room for 6500 additional oligonucleotides for use in clinical diagnostics.

Measuring Gene Expression and Other Uses of Gene Chips

In a gene expression microarray, the probes are designed and synthesized for each gene based on the genomic sequence. Once attached to the solid support, the DNA segments can be hybridized with labeled RNA from cells grown under specific conditions and scanned and analyzed by computer. Because mRNA levels are typically too low for use directly, the mRNA sequences are first amplified and converted into DNA using a modified version of the polymerase chain reaction (PCR) that converts RNA to *complementary DNA* (cDNA, Section 12.1).

To monitor global gene expression, total RNA (or cDNA) from a test sample is hybridized to an array of oligonucleotides corresponding to the entire genome. Figure 9.24b shows part of a chip containing probes for over 6000 protein-encoding genes of the yeast *Saccharomyces cerevisiae*. After hybridizing yeast cDNA to the chip, a distinct hybridization pattern is observed, and the fluorescence and its intensity reveal both which genes were expressed and at what level (Figure 9.24b); these data yield the *transcriptome* of the yeast culture grown under specified conditions (Table 9.6). Using such analyses, gene expression under different growth conditions can be measured. For example, in yeast—which can grow by either fermentation or respiration—transcriptome analyses have shown that genes that control production of ethanol (a key fermentation product) are strongly repressed while genes encoding citric acid cycle functions (needed for aerobic growth) are strongly activated when the organism is shifted from anaerobic to aerobic conditions. Overall, over 700 genes are turned on and over

1000 turned off during this metabolic transition. In “shift” experiments of this type, the expression pattern of genes of unknown function is also revealed, and analysis of these expression patterns sometimes yields valuable clues to the cellular function of these unknown proteins.

Microarrays can also be used to identify specific microbes. For example, identification (ID) chips have been used in the food industry to detect DNA sequences unique to specific pathogens, such as the gastrointestinal pathogen *Escherichia coli* O157:H7, an occasional foodborne pathogen. In environmental work, microarrays called *PhyloChips* have been used to assess microbial diversity. These contain oligonucleotides complementary to the 16S rRNA of different bacterial species, a molecule widely used in microbial systematics (Chapter 13). After extraction of bulk DNA or RNA from an environment, the presence or absence of a given species can be assessed by the hybridization response on the chip (Section 19.7). Although ID chips and *PhyloChips* can be made highly specific, the inexpensive nature of DNA isolation, sequencing, and analysis have made metagenomic approaches to the identification of specific pathogens or phylogenetic groups in natural samples the preferred method of assessment.

RNA-Seq Analysis

RNA-Seq analysis is a transcriptomic method in which all the RNA molecules from a cell are converted to DNA (cDNA, Section 12.1) and then sequenced. Provided that the genome sequence is available for comparison, *RNA-Seq* reveals both which genes were transcribed and how many RNA copies of each gene were made. Because *RNA-Seq* targets *all* transcripts, it is ideal for measuring the expression of mRNA, to identify long untranslated regions, and to discover noncoding RNAs. *RNA-Seq* requires high-throughput sequencing (second- or third-generation sequencing, Section 9.2) and is complicated a bit by the fact that the most abundant RNA in a cell is ribosomal RNA (rRNA). Nevertheless, methods are available to remove rRNA or enrich mRNA and primary transcripts from a total RNA pool. In addition, recent advancements in sequencing technology may allow sequencing without removing rRNA.

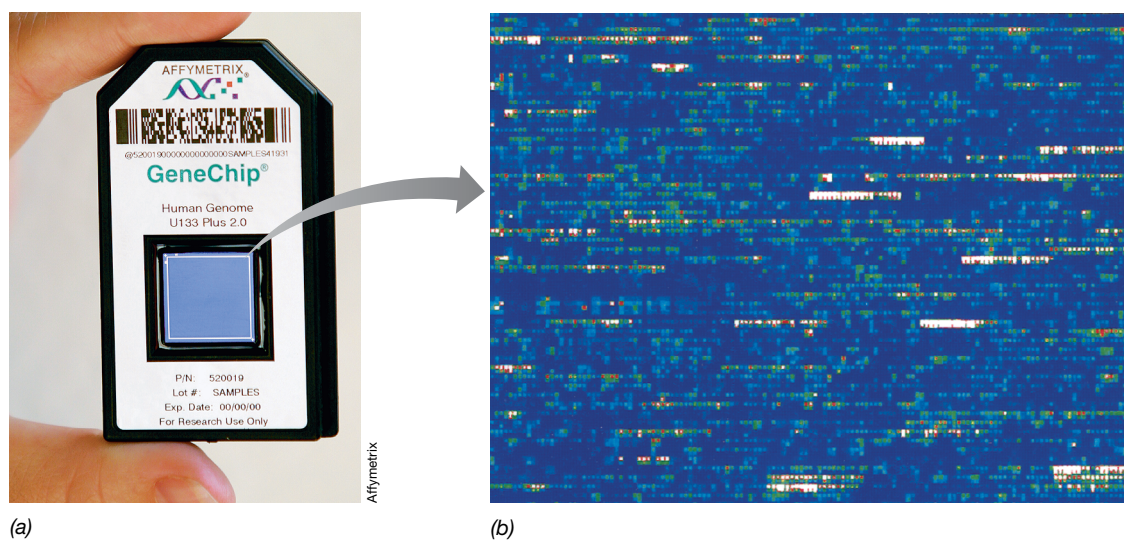


Figure 9.24 Using gene chips to assay gene expression. (a) The human genome chip contains over 47,000 gene fragments. Blowup from part a to part b indicates location of actual microarray. (b) A hybridized yeast chip shows fragments from a quarter of the genome of baker's yeast, *Saccharomyces cerevisiae*. Each gene is present in several copies and has been probed with fluorescently labeled cDNA (derived from mRNA) from yeast cells grown under a specific condition. The background of the chip is blue. Locations where the cDNA has hybridized are indicated by a gradation of colors up to a maximum number of hybridizations, which shows as white. Because the location of each gene on the chip is known, when the chip is scanned, it reveals which genes were expressed.

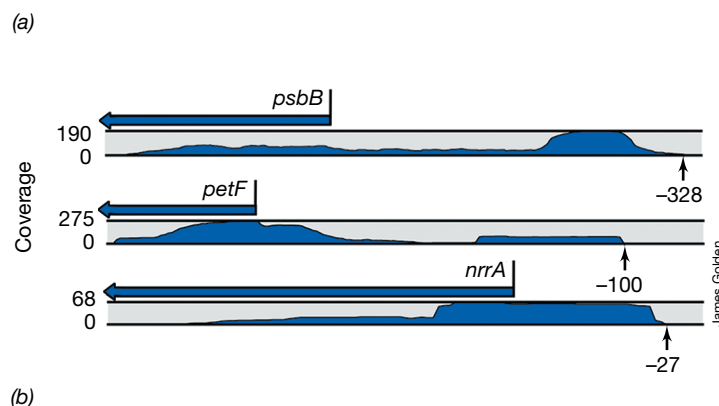
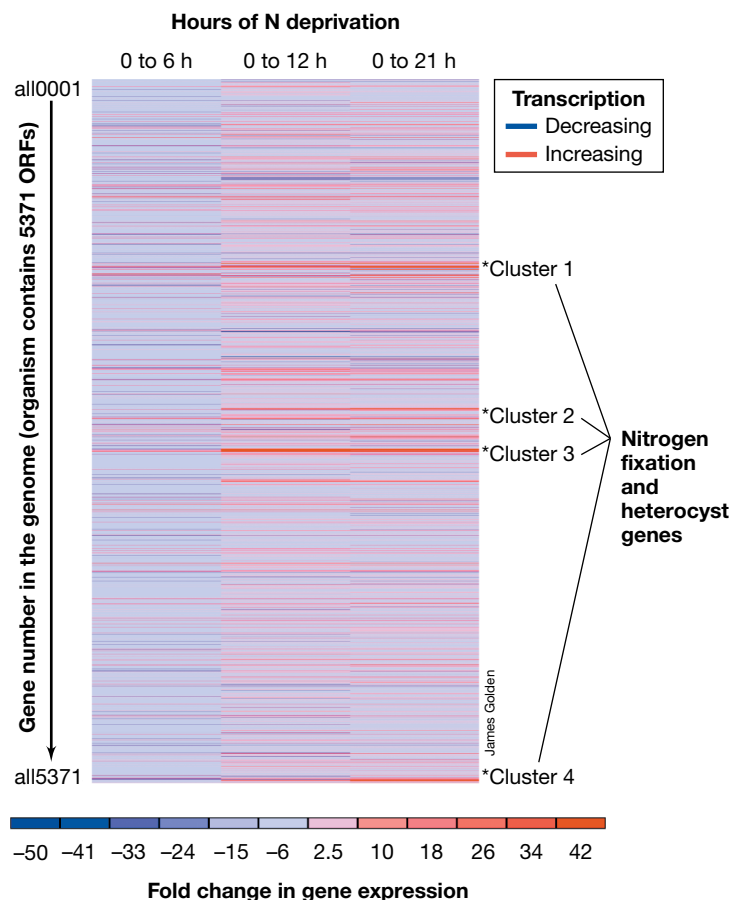


Figure 9.25 RNA-Seq analysis of the heterocyst-forming cyanobacterium *Anabaena* during nitrogen starvation. Cyanobacteria are oxygenic phototrophs (⚡ Section 14.4) and only some species, such as those that form heterocysts, can fix nitrogen under fully oxic conditions. (a) Heat map of gene expression 6, 12, and 21 h after cells were starved for fixed nitrogen. Genes that display increased expression are in red, whereas those that display decreased expression are in blue. Gene clusters 1–4 all encode proteins linked to nitrogen fixation. (b) Mapping of RNA-Seq reads. The arrows indicate the annotated open reading frames, and the plots underneath correspond to the number of sequencing reads detected for each chromosomal nucleotide position. Note that the negative numbers represent chromosomal positions upstream of the predicted start codon for the genes. The genes *psbB* and *petF* encode key proteins of photosynthesis, and *nrrA* encodes a protein that regulates heterocyst formation (the heterocyst is the site of N_2 fixation, ⚡ Section 14.6). Parts a and b modified from Flaherty, B.L., F. van Nieuwerburgh, S.R. Head, and J.W. Golden. 2011. *BMC Genomics* 12: 332.

RNA-Seq has overtaken microarray analysis as the method of choice for global studies of gene expression. The data from transcriptomic experiments can be presented in the form of a *heat map*, which uses different colors to show the level of gene expression. For example, **Figure 9.25a** identifies gene clusters that are upregulated (more intensely transcribed) during nitrogen deprivation in the heterocyst-forming *Anabaena*, a cyanobacterium that can use N_2 as its nitrogen source (nitrogen fixation, ⚡ Section 14.6). Gene clusters 1–4 represent increased expression of nitrogen fixation and heterocyst formation genes (heterocysts are the site of N_2 fixation) as the time of nitrogen deprivation increases (Figure 9.25a). RNA-Seq data can also be used for transcript mapping by plotting the sequencing reads against the genome annotation. Figure 9.25b illustrates the sequencing coverage at each base along the open reading frame regions for *psbB*, *petF*, and *nrrA*, genes that encode two key proteins needed for photosynthesis and a regulator protein for heterocyst formation in *Anabaena*, respectively (⚡ Section 7.8). These plots demonstrate long 5' untranslated regions present in these transcripts, which may be associated with regulation.

Transcript abundance under different culture conditions can also be analyzed using RNA-seq data, as indicated by a comparison of cultures of a *Clostridium* species in exponential and stationary phase (**Figure 9.26**). Clostridia are gram-positive bacteria that produce endospores, the highly resistant and dormant stage of the cell's life cycle (⚡ Section 2.10). As one might predict, transcription of genes of the glycolytic pathway (the major means by which the organism makes ATP) is elevated during exponential growth, whereas expression of sporulation genes increases in stationary phase, when nutrients become limiting. RNA-Seq is also being used for microbial community analysis and can provide information on relative transcription levels when a genome sequence is not available for comparison. In this case the sequences detected must be identified by homology with sequences present in databanks.

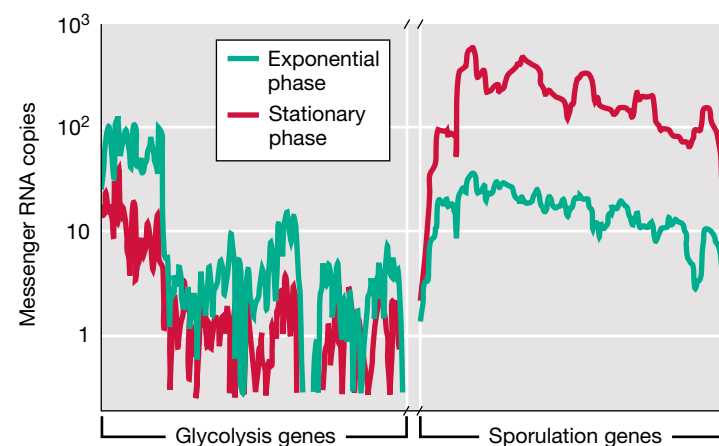


Figure 9.26 Transcriptomic analysis of sporulation genes in *Clostridium*. RNA-Seq analysis of cultures of a *Clostridium* species grown for 4.5 h (cells in exponential phase) or 14 h (cells in stationary phase). Two genomic regions are shown: (left) ~5.4-kb segment surrounding the *gap-pgk-tpi* glycolytic operon, and (right) ~1.2-kb segment surrounding the *cotC-cotB* sporulation operon. Production of endospores is triggered by nutrient starvation (⚡ Section 2.10). Data from Wang, Y., X. Li, Y. Mao, and H.P. Blaschek. 2011. *BMC Genomics* 12: 479–489.

As discussed in Section 9.8, metagenomics is the genomic analysis of pooled DNA or RNA obtained from organisms in an environment. Metagenomic analysis using RNA-Seq has been exploited for culturing a bacterium from nature that had previously resisted laboratory culture. This was accomplished by using RNA-Seq to identify highly transcribed genes in the microbial community that contained the bacterium, followed by sequence analyses to identify the proteins encoded by the highly transcribed genes. These data allowed the researchers to deduce which nutrients the bacterium was likely to be using given the predicted enzymatic activities of these proteins. Culture media were then devised using this information as a guide and the previously uncultured bacterium was successfully cultured.

MINIQUIZ

- Why is it useful to survey expression of the entire genome under particular conditions?
- What do microarrays tell you that studying gene expression by assaying individual enzymes cannot?
- What technological advances does RNA-Seq depend on?

9.10 Proteomics and the Interactome

The genome-wide study of the structure, function, and activity of an organism's *proteins* is called **proteomics**. The number and types of proteins in a cell change in response to the cell's environment or to other factors, such as developmental cycles. As a result, the term **proteome** has unfortunately become ambiguous. In its wider sense, a proteome refers to *all* the proteins encoded by an organism's genome. In its narrower sense, however, it refers to those proteins present in a cell *at any given time*. The term *translatome* has been used to describe the latter; that is, it refers to every protein made by a cell under specific conditions.

Methods in Proteomics

Modern proteomics relies on some form of *mass spectrometry* to characterize the proteome, a method that can also be used to identify metabolites (Section 9.11). The mass of ^{12}C is defined as exactly 12 molecular mass units (daltons). However, the masses of other atoms, such as ^{14}N or ^{16}O , are not exact integers. Mass spectrometry using extremely high mass resolution techniques, which can distinguish between slightly different mass-to-charge ratios, allows the unambiguous determination of the molecular formula of any small molecule. Thus mass spectrometry can be used to identify several peptides in a sample. The amino acid sequence of these peptides can then be searched against the translation of a genome to identify the presence of specific proteins. To increase sensitivity, liquid chromatography is increasingly used to separate protein mixtures. In high-pressure liquid chromatography (HPLC), a protein sample is dissolved in a suitable liquid and forced under pressure through a special column that separates proteins by differences in their chemical properties, such as size, charge, or hydrophobicity. Fractions are collected at the end of the column, the proteins in each fraction are digested by proteases, and the peptides are identified by mass spectrometry.

MALDI (*matrix-assisted laser desorption ionization*) is an advanced version of mass spectrometry that does not require that the proteins be separated or digested. Instead the sample is affixed to a matrix and then ionized and vaporized by a laser (Figure 9.27). The ions generated are accelerated along the column toward the detector by an electric field. The time of flight (TOF) for each ion depends on its mass/charge ratio—the smaller this ratio, the faster the ion moves. The detector measures the TOF for each ion and the computer calculates the mass and hence the molecular formula. The combination of these two techniques is known as *MALDI-TOF*.

Utility of Proteomics: An Environmental Case Study

While proteomic analyses can be performed on pure cultures of specific microbes, *metaproteomics* of environmental samples are

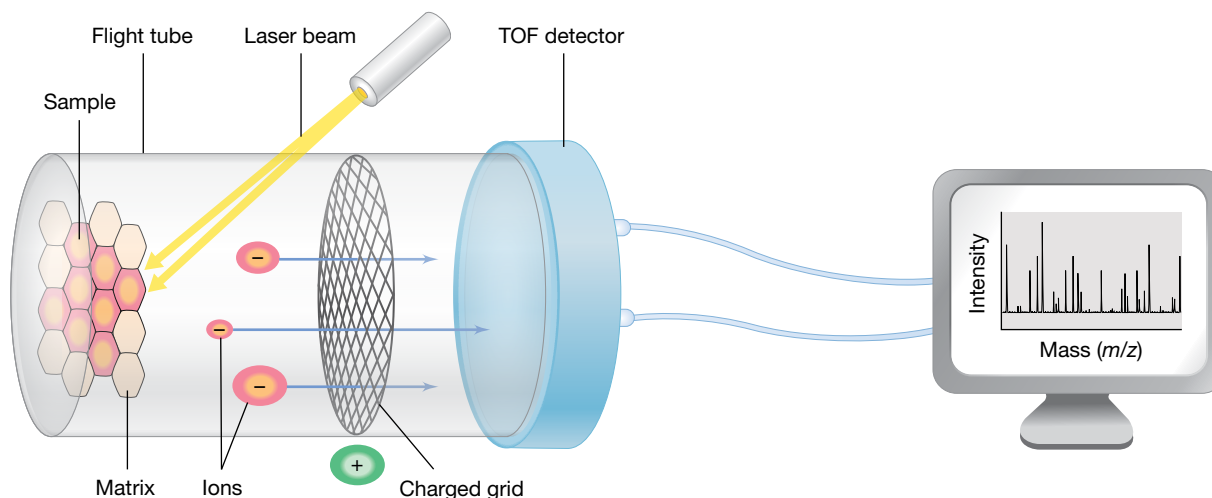


Figure 9.27 MALDI-TOF mass spectrometry. In matrix-assisted laser desorption ionization (MALDI) spectroscopy, the sample is ionized by a laser and the ions travel down the tube to the detector. The time of flight (TOF) depends on the mass/charge (m/z) ratio of the ion. The computer identifies the ions based on their time of flight, that is, the time it takes to reach the detector.

particularly insightful for revealing the collective metabolic potential of a microbial community. For example, permafrost covers an estimated 20% of the land on Earth and sequesters large amounts of organic carbon. Because climate change could lead to the release of the “greenhouse gases” CO₂ and CH₄ from permafrost, identifying the microbes inhabiting actively melting permafrost along with their metabolic potential is essential for climate predictions. Using a combination of metagenomics (Section 9.8) and metaproteomics, a revealing snapshot of the genes and proteins of the microbial community present in a permafrost meltwater bog has been obtained (Figure 9.28).

During the metagenomic analysis of the bog, DNA sequences were assembled and annotated into complete and partial genomes (Figure 9.28a). This analysis indicated that *Bacteria* of the phyla *Proteobacteria*, *Actinobacteria*, and *Chloroflexi* predominated, while methanogens (*Euryarchaeota*) were the dominant *Archaea*. From the large amount of sequence data obtained, the genomes of three novel (and as yet uncultured) methanogens could be assembled to the draft stage without any cultures of the organisms having been obtained. This indicated that unique methanogens resided in the permafrost—most likely species that function well in the cold (psychrophiles)—and therefore that the potential for increasing rates of methanogenesis as the permafrost melted was high. Using the annotated DNA sequences and the output from the mass spectrometry detection of peptides, the identity of proteins extracted from the bog sample was determined along with their microbial sources (Figure 9.28b). The proteomics detected an abundance of functionally diverse proteins including those that participate in cellular housekeeping, transport, and organic carbon respiration. In agreement with the metagenomics data, several proteins associated with C₁ metabolism, including those necessary for both making methane (methanogenesis) and oxidizing methane (methanotrophy) as well as oxidizing and reducing carbon monoxide, were also detected in the bog microbial community.

This combined metagenomic and metaproteomic snapshot of a major microbial community highlights the power of omics for resolving the “who” and “how” of complex microbial communities and for using the results to predict the response(s) of these communities to environmental changes. In the case of permafrost melting, the potential for the release of large amounts of CO₂ (due to increased respiration of organic carbon) and CH₄ (due to increased activities of methanogens) from permafrost as climate change progresses was clearly apparent (Figure 9.28).

The Interactome

By analogy with the terms genome and proteome, the **interactome** is the complete set of *interactions* among the macromolecules within a cell (Figure 9.29). Originally, the word *interactome* was applied only to interactions between proteins, many of which assemble into complexes. However, it is also possible to consider interactions between different classes of macromolecules, such as between protein and RNA (see the protein–RNA interactome in Figure 9.38).

Interactome data are typically expressed in the form of network diagrams, with each node representing a protein and the connecting lines representing the interactions. Diagrams of whole interactomes can be extremely complex (see Figure 9.35a) and thus more

focused interactomes, such as the motility protein network from the bacterium *Campylobacter jejuni* (Figure 9.29), are more instructive. This figure shows the core interactions between well-known components of the chemotaxis system (↔ Sections 2.13 and 6.7), including all other proteins that are known to interact with these.

MINIQUIZ

- Why is the term “proteome” ambiguous, whereas the term “genome” is not?
- What are the most common experimental methods used to survey the proteome?
- What is the interactome?

9.11 Metabolomics

The metabolome is the complete set of *metabolic intermediates* and small molecules produced by an organism. Thus the metabolome reflects the enzymatic pathways encoded by the genome. While gene expression and the presence of corresponding proteins suggest the activity of specific pathways, the metabolomic data confirm that these potential reactions actually occurred in the cell in a given physiological state.

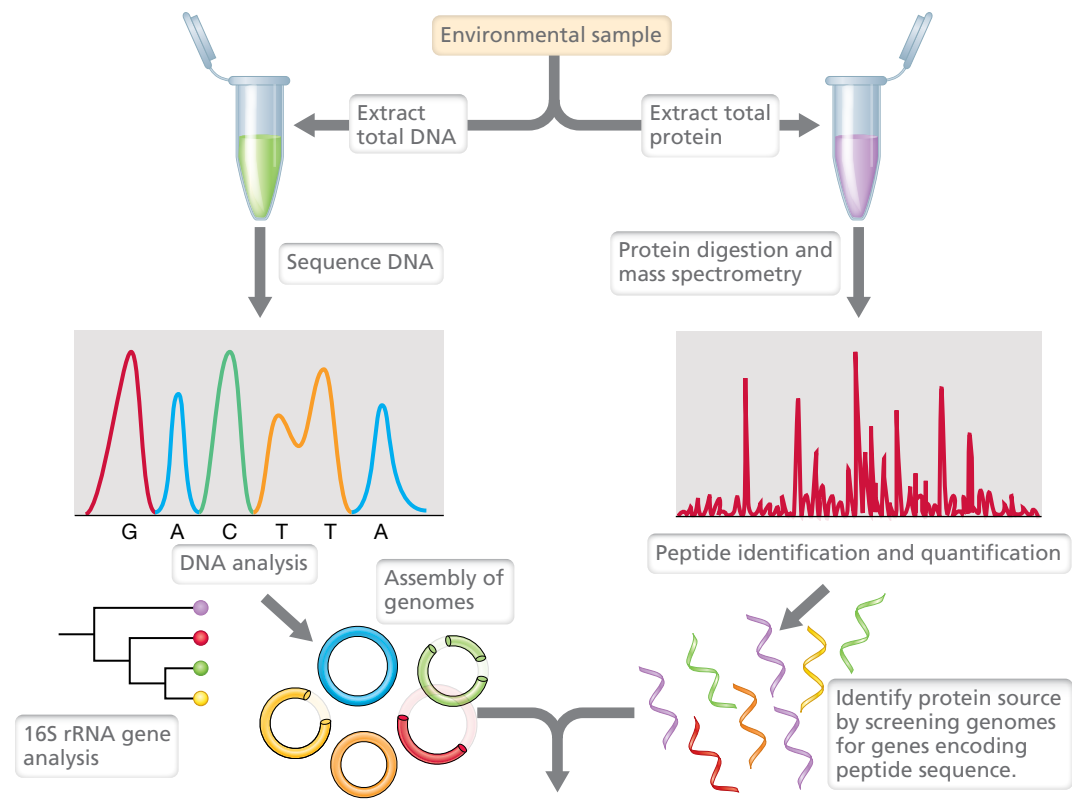
Advances in Metabolomic Techniques: NIMS

Metabolomics has lagged behind other omics due largely to the immense chemical diversity of small metabolites present in cells. This makes systematic metabolomic screening technically challenging. Early attempts used nuclear magnetic resonance (NMR) analysis of extracts from cells labeled with ¹³C-glucose (¹³C is a heavy isotope of carbon, most of which is ¹²C). However, this method is limited in sensitivity, and the number of compounds that can be identified in a mixture simultaneously is too low for resolution of complete cell extracts.

While MALDI-TOF mass spectrometry (Section 9.10) can be used to detect and identify metabolites, *nanostructure-initiator mass spectrometry* (NIMS) is a more useful technique that can directly analyze biological samples without the need for special analytical preparation (Figure 9.30). Thus biofluids, tissues, or even individual cells can be analyzed. A laser is used to ionize the sample in NIMS, just as in MALDI-TOF, but the silicon-coated surface used in NIMS does not generate the background interference seen during ionization of a matrix by MALDI-TOF. This allows for the accurate identification of small metabolites present at low concentrations and increased spatial resolution during tissue imaging. Modifications to the initiator surface can also be made to detect notoriously difficult molecules such as structurally similar carbohydrates or steroids. These traits also make NIMS more sensitive than MALDI, which is illustrated by the ability of NIMS to detect drugs such as the heart drug propafenone in a single heart cell in the yoctomole (10⁻²⁴) range (Figure 9.30).

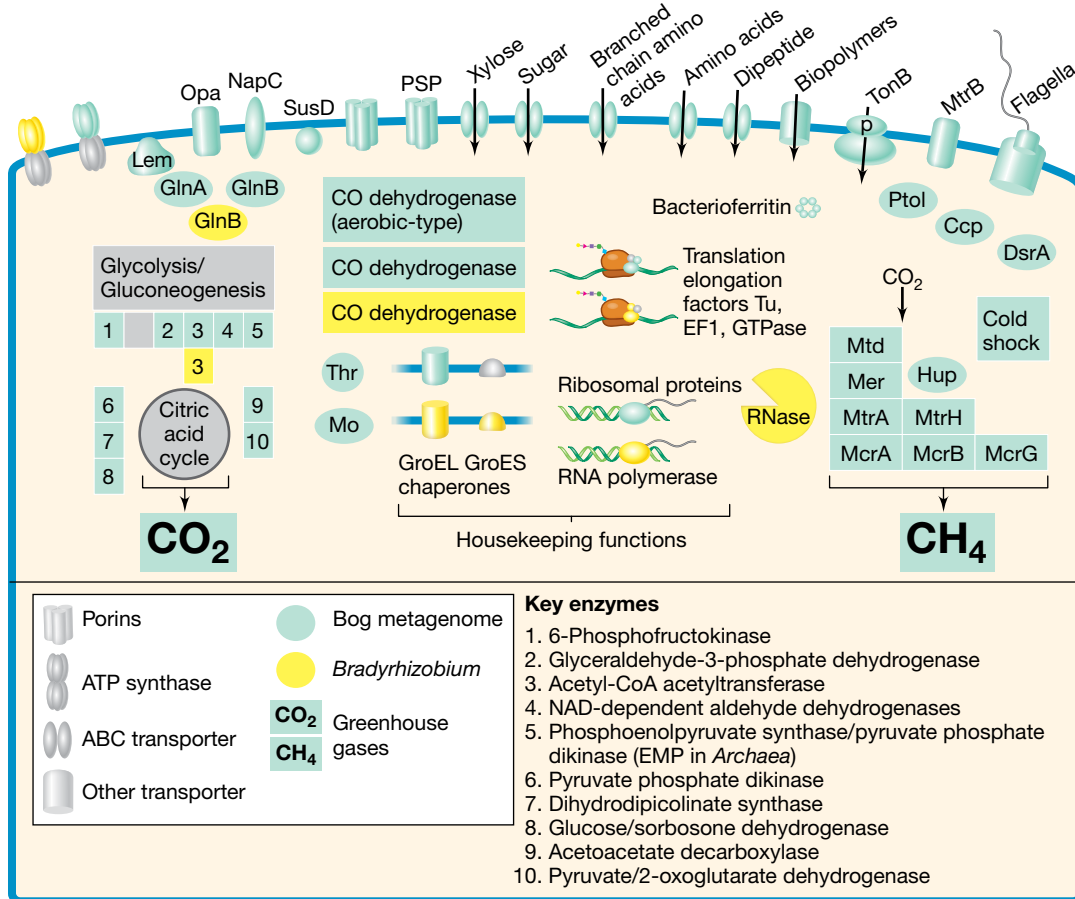
Utility of Metabolomics

Metabolome analysis has been particularly useful for the study of plant biochemistry, since plants produce several thousand different metabolites—more than most other types of organism.



(a)

Figure 9.28 Meta-omics of permafrost. (a) Metagenomic and metaproteomic analysis of an environmental sample. Total DNA and protein are extracted from a sample and analyzed to identify microorganisms and their associated proteins. Following sequencing, the DNA is assembled into partial and complete individual genomes. The 16S rRNA gene can also be profiled to determine the phylogenetic affiliation of *Bacteria* and *Archaea* present in the sample. After identifying the digested proteins by mass spectrometry, their microbial source can be determined by searching for corresponding DNA sequences in the metagenomic data. (b) Visualization of a subset of the proteins identified from metagenomic and metaproteomic analysis of a bog sample. Ccp: cytochrome C peroxidase, DsrA: sulfite reductase subunit A, GlnA/GlnB: nitrogen storage, Hup: heterosulfide reductase, Lem: peptidoglycan-associated lipoprotein, Mtd/MtrA/MtrB/MtrH/McrA/McrB/McrG/Mer/Mo: methane metabolism, NapC: NapC/NiR cytochrome C, Opa: opacity protein, PSP: phosphate-selective porin, Ptol: periplasmic component of the Tol biopolymer transport system, Thr: thermosome, SusD: sulfate ABC transport system ATP-binding protein, TonB: periplasmic protein. Data adapted from Hultman, J., et al. 2015. *Nature* 521: 208–212.



(b)

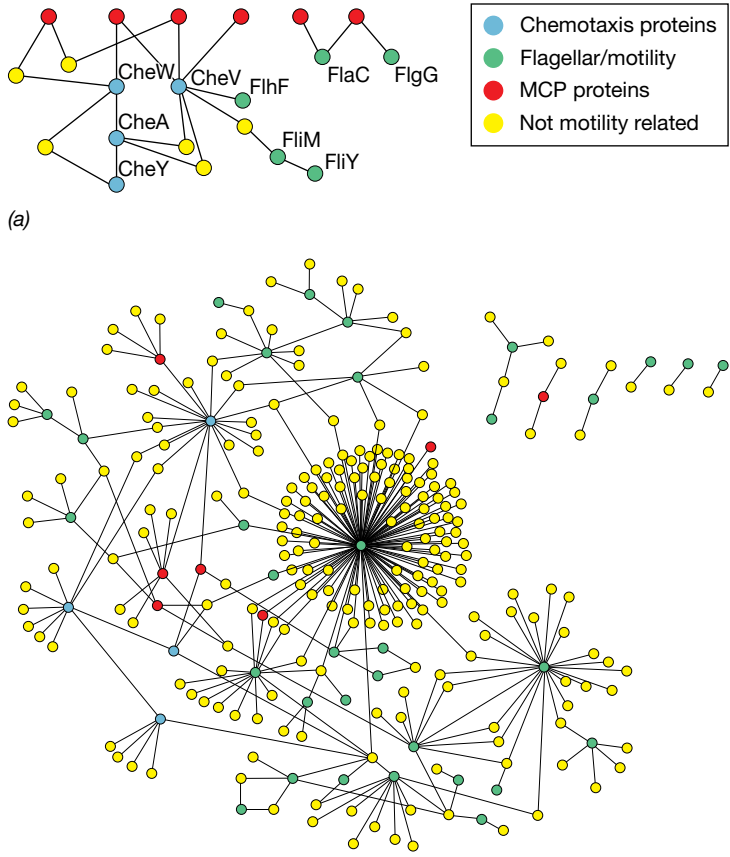


Figure 9.29 Motility protein interactome for *Campylobacter jejuni*. This network illustrates the way in which interactome data are depicted. (a) A subsection of the network highlighting the well-known proteins of the chemotaxis signal transduction pathway (CheW, CheA, and CheY) and their partners. MCP, methyl-accepting chemotaxis proteins (↔ Section 6.7). (b) High-confidence interactions between all proteins known to have roles in motility. Note the six small networks that fall outside the single large network.

These compounds include many so-called *secondary metabolites*, chemicals such as scents, flavors, alkaloids, and pigments, many of which are commercially important. Metabolomic investigations have monitored the levels of several hundred metabolites in the model plant *Arabidopsis*, and significant changes were observed in the levels of many of these metabolites in response to changes in temperature, a hint that climate change will likely alter plant metabolism in major ways. Metabolomics can also be done on microbial cultures as well as natural microbial communities such as biofilms. For example, microbiologists have detected over 3500 different metabolites in a relatively simple microbial biofilm growing in extremely acidic (pH ~0.9) and heavy-metal-rich water draining from an abandoned mine site in northern California (USA). Many of these metabolites were suspected of being osmolytes and other protective molecules for combating the osmotic and other life stressors in this extreme environment.

Metabolomics has also been deployed to help characterize the human microbiome (Chapter 24). For instance, our skin epidermis is composed not only of cells but also of microbes that contribute to epithelial health. Thus metabolites corresponding to our skin

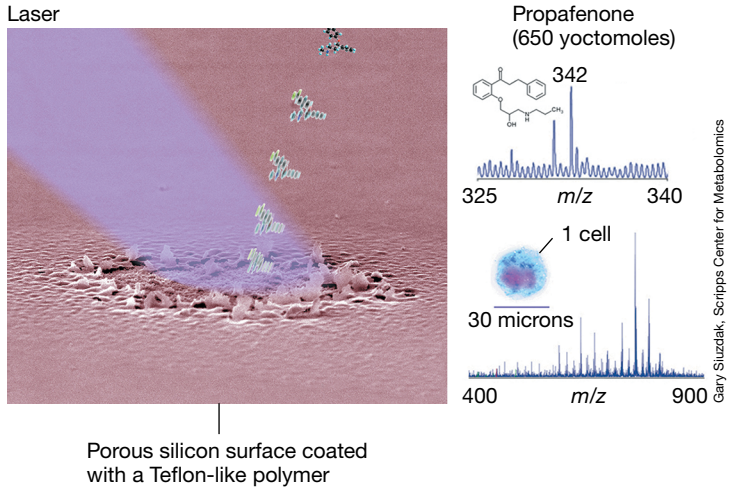


Figure 9.30 NIMS identification of metabolites. In nanostructure-initiator mass spectrometry (NIMS), a cell is placed on the silicon initiator surface and vaporized using a laser. The resulting ionized metabolites within a cell are then detected using mass spectrometry. Because NIMS lacks a matrix, it has extremely high sensitivity and resolution. Ionized metabolites are represented rising from the surface.

cells, to associated microorganisms, and to personal hygiene products are present. **Figure 9.31** shows results of a study on the pattern of metabolites and microbial diversity on the skin of two human subjects plotted simultaneously on a three-dimensional

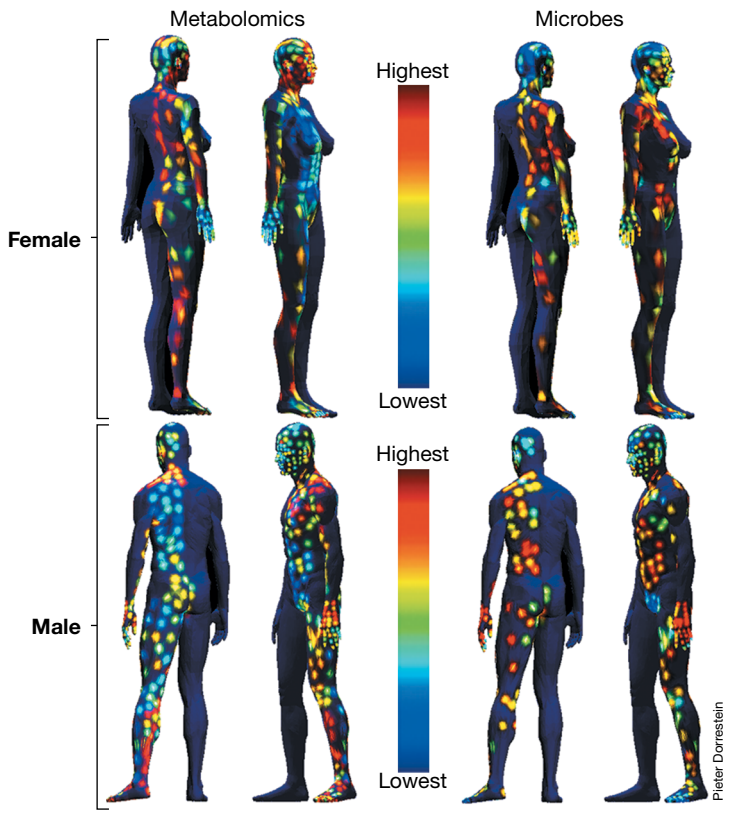


Figure 9.31 Metabolomic and metagenomic mapping of skin. Three-dimensional heat maps represent the diversity of metabolites and microorganisms detected on various areas of the skin on a male and a female subject. Red indicates the highest level of diversity and purple indicates the lowest level of diversity.

Pieter Dorrestein

topographical heat map of the human body. One of the goals of this type of research is to determine if the presence of particular metabolites can be linked to the presence of particular microbial species. Such a picture of the human skin could offer microbiologists as well as medical clinicians a better understanding of the diversity and ecology of the skin microbiome and pave the way for the future use of studies of this type in the diagnosis of skin health or disease.

As expected, the omics study of the skin (Figure 9.31) revealed many microbes known from previous studies of the skin microbiome (↔ Section 24.5). But the study also discovered a diverse mixture of metabolites. While common chemicals produced by human skin cells such as triacylglycerides and diacylglycerides were readily detected, various metabolites resulting from microbial processing of these compounds were also detected. Overall, the level of metabolite diversity detected on specific areas of the body did not correlate well with microbial diversity. Instead, the complexity of the human skin metabolome significantly exceeded the diversity of the microbial profile, and this indicates that each species is likely producing several distinct metabolites. Interestingly, and somewhat surprisingly, the results also showed that personal hygiene products are the major source of metabolites on human skin.

Throughout this chapter we have discussed various omics and their applications as more or less individual entities. We now shift our focus to integrating multiple omics to better understand the entire organism—the major goal of systems biology.

MINIQUIZ

- What techniques are used to monitor the metabolome?
- What is a secondary metabolite?

IV • The Utility of Systems Biology

The basic strategy of systems biology is to generate comprehensive models for predicting the behavior or properties of an organism that were not obvious from pre-omics era observations. These are referred to as the *emergent properties* of the organism. Understanding an organism's emergent properties provides a deeper insight into its overall biology than can any single omics study by itself. The goal is to integrate the numerous omic datasets to create meaningful models in systems biology (Figure 9.32). We begin by seeing how all of these can come together in ecological studies of individual cells in a microbial community.

9.12 Single-Cell Genomics

Besides sequencing total environmental DNA as described for metagenomics (Section 9.8), the genomes of individual cells can also be sequenced—a technique called *single-cell genomics* (SCG) (Figure 9.33). This is now possible because of the ability to amplify tiny amounts of DNA. Single-cell genomics is critical for studying the metabolic potential of microorganisms in natural microbial communities. While metagenomic analysis of a microbial community can detect the presence of genes specific to certain

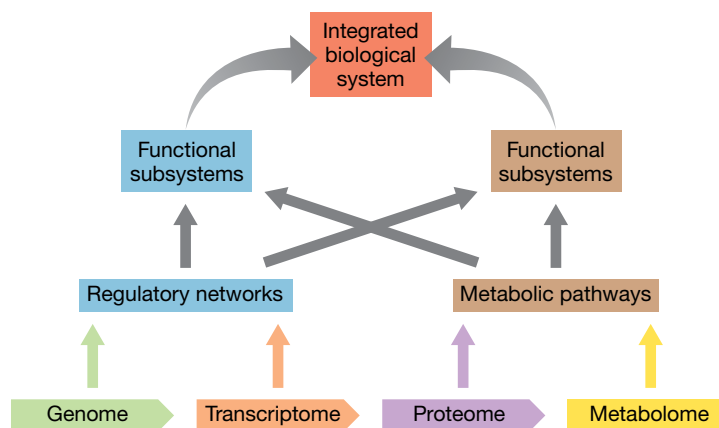


Figure 9.32 The components of systems biology. The results of various “omics” analyses are combined and successively integrated into higher-level views of the entire biology of an organism.

pathways, it is difficult to discern whether the pathways are contained within the same organism. Besides genome sequencing, transcriptome and proteome analyses can also be performed on individual cells, leading to a comprehensive omics study on one component of a microbial ecosystem.

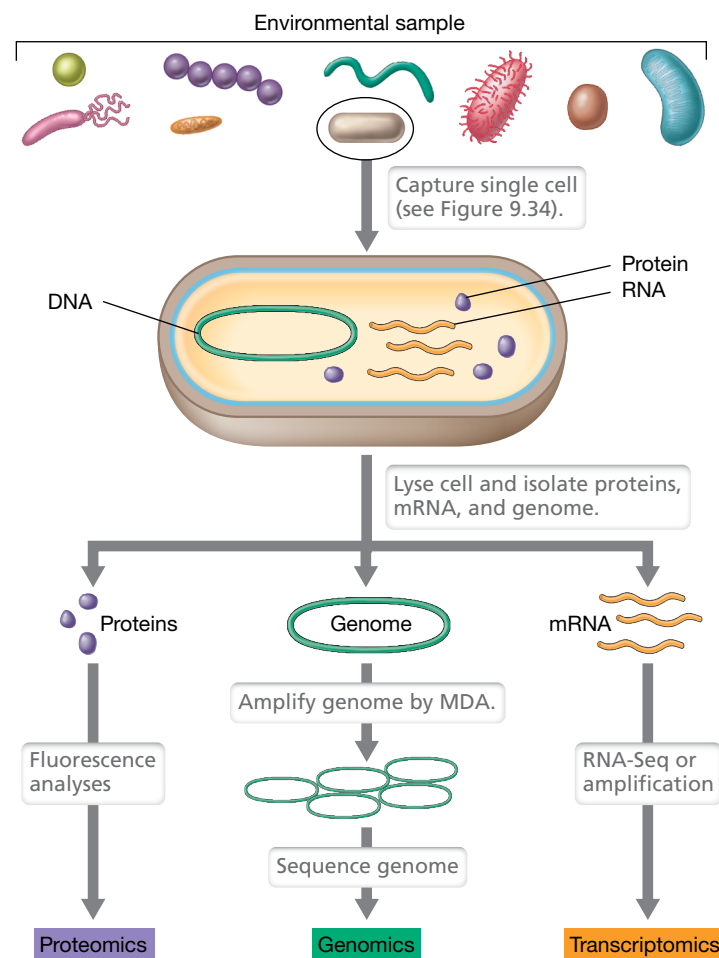


Figure 9.33 Single-cell genomics. A single cell isolated from an environmental sample can be the source of a diversified omics study.

Cell Isolation and Sample Preparation

The ability to isolate cells is obviously essential for single-cell genomics, and various physical techniques including dilution in microwells (↗ Section 19.3), encapsulation, and fluorescence-activated cell sorting (FACS) have been used to do so. For encapsulation, the sample is diluted and added to sterile oil to form microdroplets. Approximately 30% of the resulting droplets from this technique will contain only one cell (Figure 9.34). Combining droplet encapsulation with FACS, which is able to optically detect single cells, allows droplets that do not contain a cell to be rejected. This method of single-cell isolation has also been shown effective in isolating single virions for omics analyses.

Sequencing DNA from single cells relies on a modified version of the polymerase chain reaction (PCR, ↗ Section 12.1) called *multiple displacement amplification* (MDA) (↗ Section 19.12 and Figure 19.37). This PCR technique uses a special viral DNA polymerase to amplify the femtogram (10^{-15} g) quantities of DNA present in a single bacterial cell into the micrograms (10^{-6} g) of DNA required for sequencing (a billionfold amplification). However, because of the sensitivity of MDA, contamination is one of its biggest problems. Contaminating DNA can originate from the sample itself or from the laboratory equipment and reagents. If great care is not taken in isolating the cell and amplifying its genome, contaminating DNA can make up half or more of the reaction products and create major problems for genome assembly and further genomic analyses. In addition to a cell's genome, its RNA can also be analyzed using RNA-Seq following amplification to form cDNA by a modified version of PCR (Section 9.9). Single-cell proteomic analyses are trickier than nucleic acid studies because amplification is not employed, but analyses using extremely sensitive fluorescence methods are available for this purpose (Figure 9.33).

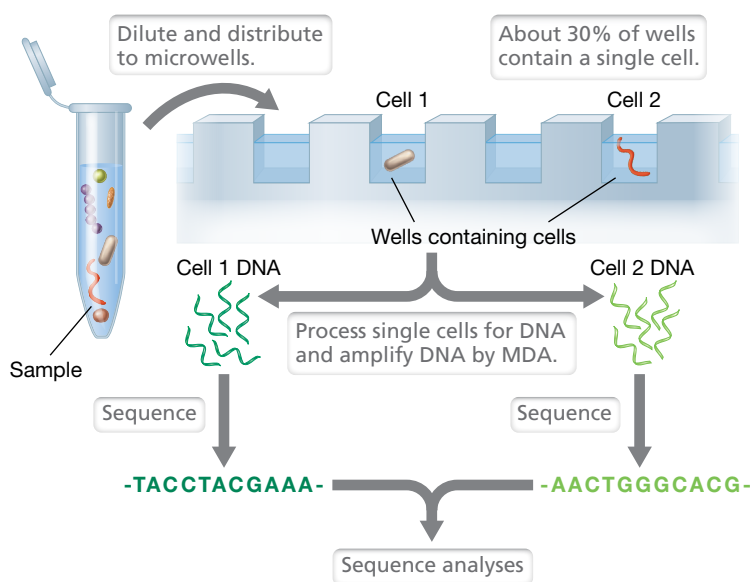


Figure 9.34 Isolation and sequencing of single cells. Droplets of a diluted sample are added to microfluidic wells prior to multiple displacement amplification for DNA sequencing.

Applications of Single-Cell Omics

Single-cell omics have the unique power to probe several facets of an organism's biology in an individual cell rather than on a cell population basis. Using SCG, metabolic genes present in an environment can be not only identified but assigned to particular species; this reveals which particular organisms are degrading which particular nutrients. For example, single-cell genomics has been used to analyze hydrocarbon degradation by bacteria in polluted environments, leading to a better understanding of which organisms are doing what in the overall process. Similarly, plasmids and viruses can be allocated to their correct host when the genome of a single cell is sequenced.

Single-cell genomics has also been used to explore the genomes of candidate phyla of *Bacteria* and *Archaea* that are detected in environmental 16S rRNA gene surveys but for which no laboratory cultures are available. These yet-to-be-cultivated microbes have been called *microbial dark matter*. One dark matter study applied SCG to over 200 uncultivated archaeal and bacterial cells from nine different environments. The results revealed several surprising findings including the following: Some archaeal RNA polymerase sigma factors are similar to their bacterial counterparts (↗ Sections 4.5 and 4.6); some stop codons (↗ Section 4.9 and Table 4.4) have been reassigned to incorporate actual tRNAs; the genomes of some bacterial cells encode an oxidoreductase enzyme previously seen only in eukaryotes; some *Archaea* contain genes encoding a bacterial-like stringent response (↗ Section 6.9); and some *Archaea* produce an enzyme that functions in peptidoglycan synthesis (recall that peptidoglycan is a “signature molecule” of *Bacteria* and is not known from any *Archaea*, ↗ Section 2.4).

Single-cell genomics is an excellent example of serendipity, the “pleasant surprise” of finding one thing while working on another. In this case, the serendipity occurred when methods designed with one goal in mind (the genomic analysis of a population of cells) were refined to probe the biology of a single cell in ways never before thought possible. Single-cell genomics is poised to complement the *Earth Microbiome Project*, a sequencing endeavor to archive the genome sequence of all cultured bacterial and archaeal type strains (<http://www.earthmicrobiome.org/>). For those species of *Bacteria* and *Archaea* that reside in the microbial dark matter, SCG offers a way to include these species in the archive while simultaneously revealing valuable clues that may help bring these organisms into laboratory culture.

MINIQUIZ

- How are single cells isolated from a mixed population?
- What must be done before minute amounts of DNA can be sequenced?

9.13 Integrating *Mycobacterium tuberculosis* Omics

Mycobacterium tuberculosis is a pathogen that infects one-third of the world's population and kills approximately 2 million people every year. Multidrug resistance and the ability to temporarily enter dormancy in response to stress (↗ Section 7.11) are two

characteristics that contribute to the persistence of *M. tuberculosis*. Thus the identification of new treatment methods is critical for combating *M. tuberculosis*. This challenge is being tackled using a systems biology approach for understanding how *M. tuberculosis*—an intracellular pathogen—adapts to the oxygen deficiency (hypoxia, a condition that develops in tuberculosis) in host cells and identifying potential drug targets for therapeutic design.

Tuberculosis Gene Expression and Regulatory Networks

As we discussed in Section 9.9, RNA-Seq can be used to characterize an organism's complete expression profile. A modification of this technique termed *dual RNA-Seq* can be used to simultaneously profile the transcriptomes of a pathogen *and* its host cell. This approach allows for the responses of both the pathogen and host to be captured during the infection process and is especially useful for understanding how *M. tuberculosis* evades the host's defense systems. Data from dual RNA-Seq and the integration of over 600 separate expression experiments have facilitated the construction of gene expression models. These models have identified the primary energy source of phagocytosed *M. tuberculosis* cells as host cholesterol, and they have shown that the amino acid aspartate produced by the host is used by *M. tuberculosis* not only as a nitrogen source but also as protection against reactive oxygen species (Section 5.14) produced by macrophages trying to kill *M. tuberculosis* cells.

The *M. tuberculosis* research has integrated transcriptome and other omics datasets including ChIP-Seq, a method where an antibody to a specific DNA-binding protein is used to trap the protein bound to its DNA, after which the DNA is removed and sequenced. From ChIP-Seq analyses, several transcription factors

and regulatory networks critical to the pathogenesis of *M. tuberculosis* have been identified. These include regulation by the major bacterial cell regulator LexA (Section 11.4) of certain genes for DNA damage response, and control by members of the DevR regulon of the entry of *M. tuberculosis* into dormancy (Figure 9.35). Additional potential drug targets have also been identified through the screening of over 1000 different mutant strains of *M. tuberculosis*. This analysis mapped 18 previously uncharacterized genes to the persistence response of *M. tuberculosis* (Section 7.11), illustrating the power of integrating omics and mutant analyses to identify important pathogenicity genes in *M. tuberculosis*.

Tuberculosis Proteomics and Metabolomics

Proteomics has been used to identify proteins essential to the *M. tuberculosis* hypoxia response. While expected proteins that combat reactive oxygen species such as superoxide dismutase were detected, so were toxin–antitoxin systems (Section 7.11) and proteins that participate in the biosynthesis of the unique lipoproteins of *M. tuberculosis*. Production of nitrate and nitrite transporters also increased as *M. tuberculosis* switched to anaerobic respiration (Section 3.12) for energy metabolism. Moreover, at least 160 different uncharacterized proteins containing the N-terminal amino acid sequence “Pro-Glu/Pro-Pro-Glu” were detected in the pathogen. If drugs could be developed that target this sequence, they might be effective treatments for tuberculosis.

Online fundraising and social media have even been combined for the common goal of identifying new drugs to treat tuberculosis. The *Connect to Decode* initiative (<http://c2d.osdd.net>) is composed of more than 800 researchers whose collective goal is to identify new tuberculosis drug targets by mapping the entire interactome of *M. tuberculosis*. By analyzing over 10,000 *M. tuberculosis*

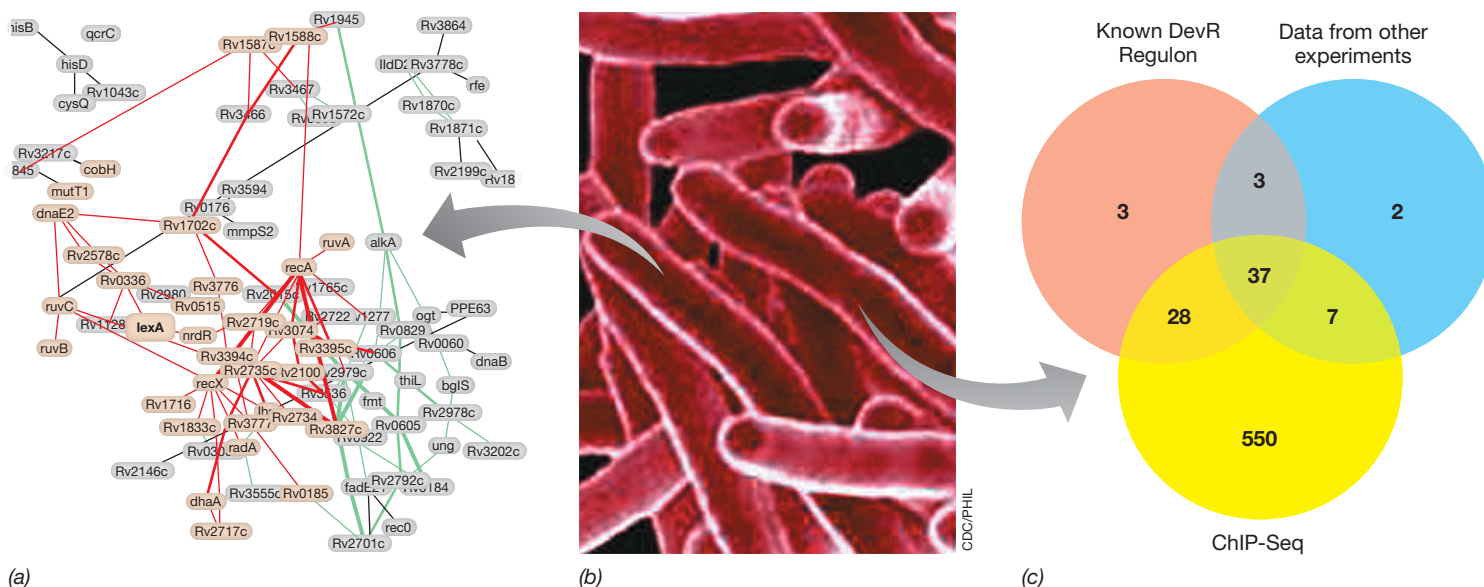


Figure 9.35 Genes controlled by the regulators LexA and DevR in *Mycobacterium tuberculosis*. (a) A snapshot of the interactome of some of the gene clusters that participate in DNA repair as identified from their expression pattern. Members of the LexA regulon are in red, while green connectors indicate interactions

of other DNA repair genes. (b) Colorized scanning electron micrograph of cells of *M. tuberculosis*. (c) Venn diagram of results of various studies of the DevR regulon. Three methods of assessing genes in the regulon were employed and the number of genes identified from each detection method is

indicated. A total of 662 genes were identified by ChIP-Seq analysis; of these, only 37 were identified by every method. Data adapted from van Dam, J.C.J., et al. 2014. *BMC Biology Systems* 8: 111.

experimental datasets, 87% of the proteins encoded by the genome have now been annotated. This is a vast improvement over the 52% of genes annotated in the original genome sequence. This collective effort has also resulted in a map of the complete *M. tuberculosis* interactome, with over 1400 proteins connected by over 2500 functional relationships (Figure 9.36). So far, 17 potential drug targets and their interactomes have been identified by this initiative. These targets lack homology to human proteins or to proteins of the human oral and gastrointestinal microbiome, thus increasing the probability of identifying drugs with the least side effects.

While tuberculosis metabolomics is in its infancy, studies comparing virulent versus attenuated *Mycobacterium* strains have identified over 1000 different lipids—many of them unique lipoproteins—that are likely quite important to the biology of *M. tuberculosis*. Thus, the *M. tuberculosis* lipid metabolism proteome may contain several potential therapeutic targets for tuberculosis. Finally, by profiling the *secretome* (an inventory of metabolites secreted) of *M. tuberculosis*, molecules unique to virulent strains (such as the nucleotide 1-tuberculosinyladenosine) have been discovered. This modified nucleotide is present in the urine of humans infected with *M. tuberculosis*, and its discovery illustrates the power of omics to link a particular pathogen to a specific molecule—be it a gene, protein, or metabolite—and yield new disease markers for use in clinical diagnostics. Moreover, if this modified nucleotide is essential to *M. tuberculosis*, drugs that interfere with its synthesis or activity are potential anti-tuberculosis drugs.



Vashisht, R. et al. 2012. *PLoS One* 7 (7): e39809

Figure 9.36 *Mycobacterium tuberculosis* protein interactome. Node color indicates proteins of the same category, connecting lines indicate interactions, and node size indicates relative number of interactions. Adapted from Vashisht, R., et al. 2012. *PLoS One* 7(7): e39809.

MINIQUIZ

- How does dual RNA-Seq differ from traditional RNA-Seq?
- How has systems biology provided new approaches to treating the disease tuberculosis?

9.14 Systems Biology and Human Health

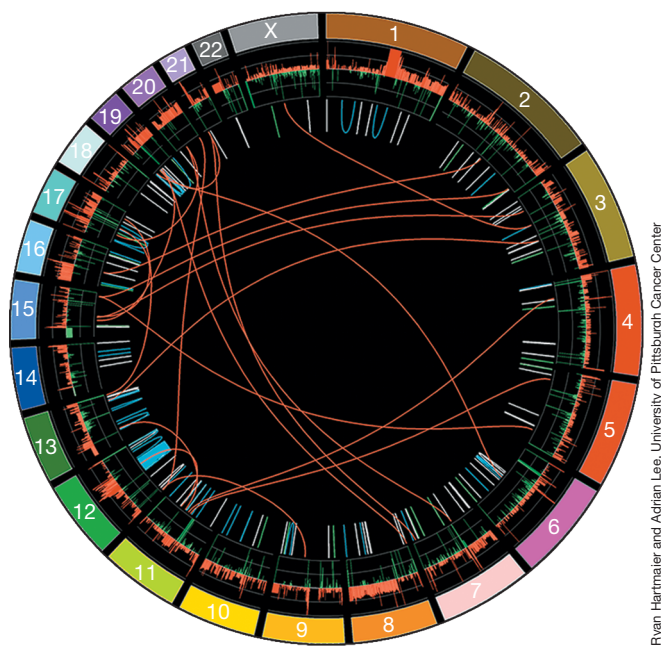
In 2003 the sequence of the 3-billion-base-pair human genome was completed and released in an international effort that cost nearly \$3 billion. Scientists initially estimated that the human genome contained about 100,000 protein-encoding genes; however, we now know this number is much smaller, around 20,000. Since much more DNA is present in the human genome than the DNA that encodes proteins, what does the remaining DNA do? Is the bulk of the human genome simply “junk DNA”? Systems biology has attacked this question and revealed new and important information.

The Human Genome and ENCODE

To help unravel the mystery of the human genome and discover the role of the predicted genes in the human body, an international collaborative project termed the *Encyclopedia of DNA Elements* (ENCODE) project was initiated (www.encodeproject.org). The goal of this project is to create a catalog of functional elements in the human genome by measuring RNA expression (transcriptomics), identifying proteins that interact with DNA and RNA, and measuring DNA methylation to assess epigenetic interactions (changes to DNA that are not sequence based) in specific cells under different conditions. Thus far, ENCODE has revealed the surprising result that 80% of the human genome is *functional* in one cell type or another! If not encoding protein, this DNA functions as binding sites for proteins that influence gene activity, or functions as sites where chemical modifications lead to gene silencing, or encodes regulatory RNA. Hence, much of what was previously thought to be junk DNA in the human genome (stretches of DNA that lacked open reading frames) actually has a function, mainly in regulatory roles. ENCODE experiments have also been instrumental in revealing nucleotide polymorphisms (slight sequence differences in the same gene from different people) that correlate with certain genetic diseases. Thus, insights provided by ENCODE regarding how human DNA works are already bearing fruit for diagnostic medicine.

Personalized Omics Profiles and Medicine

How can the human genome and systems biology be applied to health? With the advent of next-generation sequencing, the cost to sequence a human genome has fallen under \$1000, triggering an onslaught of human genome sequence data. By comparing genome sequences in over 2500 people from different continents, scientists have already identified over 88 million genome sites subject to variation. These genomic variants include single nucleotide differences, insertions and deletions, and rearrangements, each of which may turn out to be harmless, to be beneficial, to contribute to noninfectious “lifestyle” diseases such as obesity, diabetes, and heart disease, or to govern susceptibility to certain



Ryan Hartmaier and Adrian Lee, University of Pittsburgh Cancer Center

Figure 9.37 Breast cancer tumor genome. Numbers on the outside ring represent individual chromosomes. The next ring represents increased (red) and decreased (green) copy number of DNA loci compared to a nontumor genome. The inner circle represents structural variations: red, translocations; blue, inversions; black, deletions; green, segmental duplications.

cancers. Understanding if and how genomic variants contribute to disease can be used to improve diagnostics, treatments, and prevention.

The genomics age has also benefited routine tumor genome sequencing. Data from these analyses are deposited in the Cancer Genome Atlas (<http://cancergenome.nih.gov/>) and have shown that each form of cancer contains a unique set of somatic (non-germ-line) mutations. For example, **Figure 9.37** illustrates copy number variations and genome rearrangements in a metastatic breast cancer sample. Patients at risk can have regions of specific genes sequenced (called *genotyping*) to determine if they are more susceptible to a certain cancer according to a comparison with these datasets. One example of this is genotyping the tumor suppressor encoding the genes *BRCA1* and *BRCA2*; certain mutations in these genes are known to increase the risk of developing breast cancer.

In addition to cancer diagnostics, omics has opened the era of *personalized medicine*, where genomic, transcriptomic, proteomic, metabolomic, and pharmacogenomic (omics responses to drugs) data are exploited to generate a snapshot of normal and disease states along with immune processes that occur in between. The first step in personalized medicine is the generation of an *integrative Personal Omics Profile (iPOP)*. Tracking changes in this profile during healthy and disease states can aid in assessing medical risks and diagnosing and treating patients. The utility of personalized medicine can be illustrated by a study where blood samples taken from a male subject 20 times over a period of two years were subjected to proteomic and

metabolic profiling of about 5000 proteins and 4000 metabolites (**Figure 9.38**). The results suggested that the subject was at risk for coronary artery disease—which was not surprising based on his medical history and familial incidence—but also indicated that he was at high risk for type 2 diabetes. While this was a surprising finding considering his overall condition and familial health history, the prediction was based on profiling the subject’s immune response to a viral infection. During these infections, both an increase in autoantibodies targeting an insulin receptor-binding protein and changes in gene expression related to the insulin response were detected. Just as this subject’s personalized medicine profile predicted, he developed diabetes over the course of the two-year study (**Figure 9.38**). By tracking his iPOP, the onset of diabetes was revealed by the increasing production of RNAs, proteins, and metabolites, such as hemoglobin A1c and lauric acid, related to glucose metabolism (**Figure 9.38**). This case study illustrates the potential of iPOPs in monitoring the immune response and disease states. However, problems such as error rates, the analysis and storage of such large datasets, assessment of complicating factors

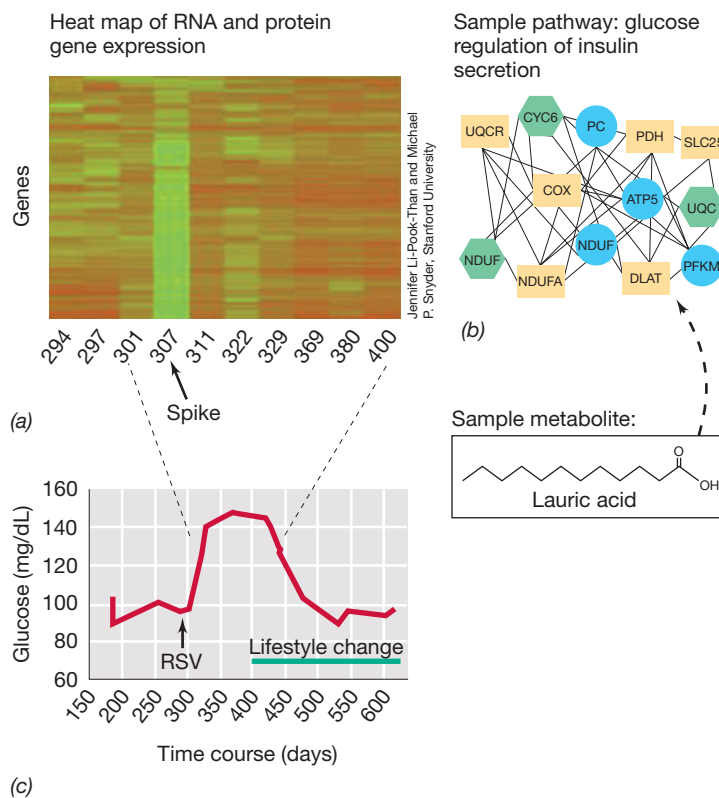


Figure 9.38 A personal integrated omics profile (iPOP). (a) Partial heat map of RNA and protein expression. Data plotted are from days 294–400 of the two-year study, with an increase in RNAs and proteins related to diabetes spiking on day 307. (b) Glucose regulation interactome. RNAs are indicated with blue circles, proteins with yellow squares, and RNA and its corresponding protein with green hexagons. A metabolite from the interactome is also shown. (c) Blood glucose levels during the time course of the study. The time of contracting a respiratory virus (RSV) is indicated as well as the time frame in which the patient made lifestyle and diet changes. Parts b and c adapted from Jennifer Li-Pook-Than and Michael P. Snyder, Stanford University.

(such as additional diseases during the course of the study, Figure 9.38), and ethical issues need to be addressed before iPOPs become common tools in clinical medicine.

Microbiology as well as biology in general has been forever changed by the dawn of genomics. Our journey through this chapter has only scratched the surface in describing how omics can be used to address previously intractable scientific questions. Much more is almost certainly in store, as major leaps forward in the omics field so far have typically been only one technological

advancement away. In Chapter 10 we continue our genomics theme but change our focus from cells to viruses—the most abundant and genetically diverse microbes on Earth.

MINIQUIZ

- What is the ENCODE project?
- What are genomic variants?

MasteringMicrobiology®

Visualize, explore, and think critically with Interactive Microbiology, MicroLab Tutors, MicroCareers case studies, and more. MasteringMicrobiology offers practice quizzes, helpful animations, and other study tools for lecture and lab to help you master microbiology.

Chapter Review

I • Genomics

9.1 Small viruses were the first organisms whose genomes were sequenced, but now many prokaryotic and eukaryotic cellular genomes have been sequenced.

Q What is one discovery resulting from the availability of a microbial genome?

9.2 DNA sequencing technology is advancing quickly. These advances have greatly increased the speed of DNA sequencing. Computer analysis of resulting sequencing data is also a vital part of genomics. Computational tools are used not only to annotate genomes but also to analyze sequences and the structures of biological macromolecules.

Q How can protein homology assist in genome annotation?

9.3 Sequenced genomes of *Bacteria* and *Archaea* range in size from 0.14 to 14.7 Mbp. The smallest are smaller than those of the largest viruses, whereas the largest have more genes than some eukaryotes. Gene content in prokaryotic cells is typically proportional to genome size. Many genes can be identified by their sequence similarity to genes found in other organisms. However, a significant percentage of sequenced genes are of unknown function.

Q Approximately how many genes are necessary for a microbial cell to have a free-living existence?

9.4 Virtually all eukaryotic cells contain mitochondria, and in addition, plant cells contain chloroplasts. Although the genomes of organelles are independent of the nuclear genome, the organelles themselves are not. Many genes in the nucleus encode proteins required for organelle function. The complete genomic sequence of many microbial eukaryotes has also been determined, and the number of genes ranges from 1000 (less than many bacteria) to 60,000 (more than twice as many as humans).

Q Which genomes are larger, those of chloroplasts or those of mitochondria? How does your genome compare with that of yeast in overall size and gene number?

II • The Evolution of Genomes

9.5 Genomics can be used to study the evolutionary history of an organism. Organisms contain gene families, genes with related sequences. If these arose because of gene duplication, the genes are said to be paralogs; if they arose by speciation, they are called orthologs.

Q What is the major difference in how duplications have contributed to the evolution of the genomes of prokaryotic and eukaryotic cells?

9.6 Organisms may acquire genes from other organisms in their environment by horizontal gene transfer, and such a transfer may even cross phylogenetic domain boundaries. The mobilome, which includes transposons, integrons, and viruses, is important in genome evolution and often carries genes encoding antibiotic resistance or virulence factors.

Q How can comparative genomics help identify horizontal gene transfer?

9.7 Comparison of the genomes of multiple strains of the same bacterial species shows a conserved component (the core genome) plus many variable genetic modules only present in certain strains of the species (combined with the core genome, this constitutes the pan genome). Many *Bacteria* contain relatively large inserts of foreign origin known as chromosomal islands. These contain clusters of genes that encode specialized metabolic functions or pathogenesis and virulence factors (pathogenicity islands).

Q What are chromosomal islands? Why are they considered to be of foreign origin?

III • Functional Omics

9.8 Most microorganisms in the environment have never been cultured. Nonetheless, analysis of DNA samples has revealed enormous sequence diversity in most habitats. The concept of the metagenome embraces the total genetic content of all the organisms in a particular habitat.

Q How do the human microbiome and mycobiome differ?

9.9 Microarrays consist of oligonucleotide probes corresponding to genes or gene fragments attached to a solid support in a known pattern; mRNA, cDNA, or DNA can then be labeled and hybridized to the gene chip to determine patterns of gene expression or the presence or absence of specific organisms. RNA-Seq combined with sequencing of cDNA can be used to profile the entire transcriptome of an organism.

Q Besides gene expression, what else can be assayed using gene chips?

9.10 Proteomics is the analysis of all the proteins present in an organism. The ultimate aim of proteomics is to understand the structure, function, and regulation of these proteins. The interactome is the total set of interactions between macromolecules inside the cell.

Q How does metaproteomics differ from proteomics?

9.11 Metabolomics profiles the complete set of metabolic intermediates produced by an organism. This analysis can determine active metabolic pathways and potential

cross-feeding (one organism supplies a nutrient for another organism) in community samples.

Q Why is investigation of the metabolome lagging behind that of the proteome?

IV • The Utility of Systems Biology

9.12 With advances in molecular techniques, the genomes of single cells can be sequenced. Expression and protein profiles of single cells can also be determined. These techniques are instrumental for studying as yet uncultured microbes.

Q How can single-cell genomics be used to address microbial dark matter?

9.13 By integrating multiple omics datasets in a systems biology approach, computer models predicting molecular activities and interactions in cells can be generated. For example, potential drug targets for the treatment of *Mycobacterium tuberculosis* have been identified using systems biology.

Q How can systems biology be used to discover new diagnostic markers for disease?

9.14 Systems biology can also be applied to personal medicine. Besides detecting genetic variants, disease risks can be predicted by profiling a person's genome, transcriptome, proteome, and metabolome.

Q What percentage of the human genome is now predicted to have functionality in at least one cell type?

Application Questions

1. Apart from genome size, what factors make complete assembly of a eukaryotic genome more difficult than assembly of a genome from a species of *Bacteria* or *Archaea*?
2. Describe how one might determine which proteins in *Escherichia coli* are repressed when a culture is shifted from a minimal medium (containing only a single carbon source) to a rich medium containing many amino acids, bases, and vitamins. How might one study which genes are expressed during each growth condition?
3. The gene encoding the beta subunit of RNA polymerase from *Escherichia coli* is said to be orthologous to the *rpoB* gene of *Bacillus subtilis*. What does that mean about the relationship between the two genes? What protein do you suppose the *rpoB* gene of *B. subtilis* encodes? The genes for the different sigma factors of *E. coli* are paralogous. What does that say about the relationship among these genes?
4. Describe how you could use systems biology to discover a new biologically produced antibiotic.

Chapter Glossary

Bioinformatics the use of computational tools to acquire, analyze, store, and access DNA and protein sequences

Chromosomal island a bacterial chromosome region of foreign origin that contains clustered genes for some extra property such as virulence or symbiosis

Codon bias the relative proportions of different codons encoding the same amino acid; it varies in different organisms. Same as codon usage

Core genome the part of a genome shared by all strains of a species

Gene chip small, solid supports to which genes or portions of genes are affixed and

arrayed spatially in a known pattern (also called microarrays)

Gene family genes related in sequence to each other because of common evolutionary origin

Genome the total complement of genetic information of a cell or a virus

Genomics the discipline that maps, sequences, analyzes, and compares genomes

Homologs genes related in sequence to an extent that implies common genetic ancestry; includes both orthologs and paralogs

Horizontal gene transfer the transfer of genetic information between organisms as opposed to transfer from parent to offspring

Hybridization the joining of two single-stranded nucleic acid molecules by complementary base pairing to form a double-stranded hybrid DNA or DNA–RNA molecule

Interactome the total set of interactions between proteins (or other macromolecules) in an organism

Metabolome the total complement of small molecules and metabolic intermediates of a cell or organism

Metagenome the total genetic complement of all the cells present in a particular environment

Metagenomics the genomic analysis of pooled DNA or RNA from an environmental sample containing organisms that have

not been isolated; same as environmental genomics

Microarray small, solid supports to which genes or portions of genes are affixed and arrayed spatially in a known pattern (also called gene chips)

Mobilome the mobile genetic elements in a genome

Nucleic acid probe a strand of nucleic acid that can be used to hybridize with a complementary strand of nucleic acid in a mixture; one of the two strands is labeled.

Open reading frame (ORF) a sequence of DNA or RNA that could be translated to give a polypeptide

Ortholog a gene in one organism that is similar to a gene in another organism because of descent from a common ancestor (see also *paralog*)

Pan genome the totality of the genes present in the different strains of a species

Paralog a gene whose similarity to one or more other genes in the same organism is the result of gene duplication (see also *ortholog*)

Pathogenicity island a bacterial chromosome region of foreign origin that contains clustered genes for virulence

Proteome (1) the total set of proteins encoded by a genome or (2) the total protein complement of an organism under a given set of conditions, also called the translome

Proteomics the genome-wide study of the structure, function, and regulation of the proteins of an organism

Sequencing deducing the order of nucleotides in a DNA or RNA molecule by a series of chemical reactions

Systems biology the integration of data from genomics and other “omics” areas to build an overall picture of a biological system

Transcriptome the complement of all RNA produced in an organism under a specific set of conditions

Translatome the total set of proteins produced by an organism under a specific set of conditions

10

Viral Genomics, Diversity, and Ecology

microbiologynow

Viral Imaging to the Rescue: Structural Blueprint of Zika


Prior to 2016, Zika was known as a relatively benign mosquito-transmitted viral disease. Little research had been performed on the virus because most Zika infections caused only mild flulike symptoms lasting a few days. However, in recent outbreaks in South America, Central America, and Mexico, Zika infections in pregnant women have been linked to a birth defect called microcephaly, a severe brain malformation in which a baby is born with an abnormally small head. As the virus reached epidemic status, unexpected reports of Zika transmission from sexual contact emerged that resulted in the World Health Organization declaring the Zika virus “a public health emergency of international concern.”

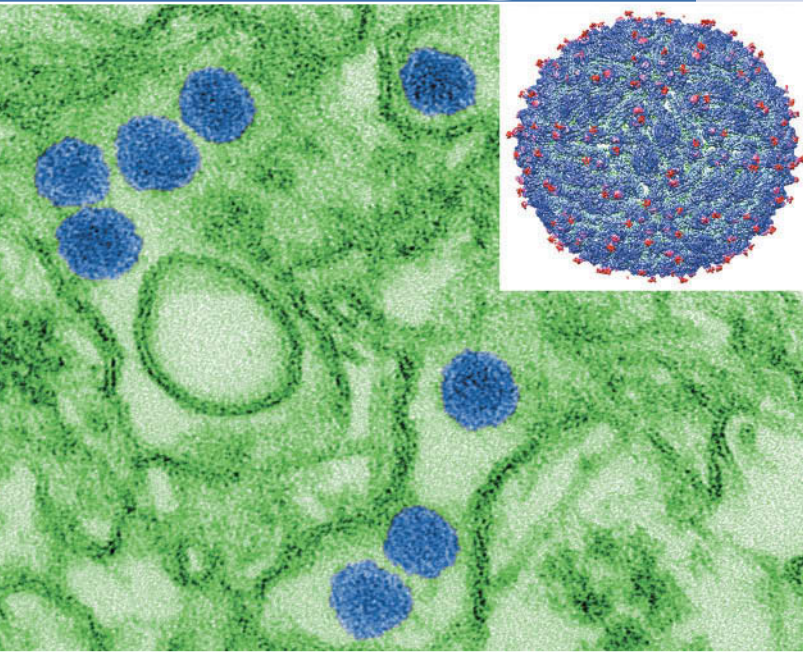
What can be done to combat the Zika epidemic? Zika is an enveloped icosahedral virus with a single-stranded plus-sense RNA genome. As a member of the *Flaviviridae* virus family, Zika is similar in many respects to the mosquito-transmitted dengue, yellow fever, and

West Nile viruses. However, Zika differs from these other viruses in its ability to enter the central nervous system and potentially cross the placenta, and such capabilities likely set the stage for the virus to cause birth defects.

To better understand the Zika virus, scientists have zeroed in on the structure of the virion using advanced imaging techniques. The image here is a colored transmission electron micrograph of Zika virions (blue, 40 nm in diameter) and a three-dimensional (3D) cryo-electron micrograph reconstruction of a mature Zika virion at near-atomic resolution (inset). The 3D analysis has revealed a striking difference between Zika and its close relatives—a unique carbohydrate group is associated with each of the 180 proteins that make up the icosahedral capsid of the Zika virus. Scientists believe that this molecule may be the key to Zika’s ability to bind to nerve cells. If so, the structure could make an excellent antiviral target and offer opportunities for the development of both therapeutic and preventive drugs.

Besides answering questions regarding Zika’s ability to attack nerve cells, the structural blueprint generated by these advanced microscopic techniques could be instrumental in future vaccine development, scientists say. Immediately, however, this work has provided a Zika virus fingerprint for its diagnosis and detection, and illustrates how basic science often contributes to improving human health.

 **Source:** Sirohi, D., et al. 2016. The 3.8 Å resolution cryo-EM structure of Zika virus. *Science* 352: 467–470.



- I Viral Genomes and Evolution 311
- II DNA Viruses 315
- III Viruses with RNA Genomes 324
- IV Viral Ecology 331
- V Subviral Agents 336

Viruses infect all organisms and are the greatest repository of genetic diversity on Earth. In this chapter we explore viral diversity from both genomic and ecological perspectives and revisit and reinforce many of the basic concepts of virology developed in Chapter 8.

I • Viral Genomes and Evolution

Viruses have DNA or RNA genomes that can be either single-stranded or double-stranded (Chapter 8). Compared with cells, viral genomes can create some unusual challenges for genetic information flow. We begin our coverage by grouping viruses by their genome structure rather than by the hosts they infect, because viruses with the same genome structure face common problems in genetic information flow. We then consider which viruses infect cells in each of the domains of life and conclude with some hypotheses for how viruses may have first appeared and how viruses may have found a home on the universal tree of life.

10.1 Size and Structure of Viral Genomes

Viral genomes vary almost a thousandfold in size from smallest to largest. DNA viruses exist along this entire gradient from the tiny circovirus, whose 1.75-kilobase single-stranded genome pales in comparison to that of the recently discovered 2.5-megabase-pair double-stranded DNA genome of Pandoravirus (Figure 10.1). The genome of the latter is over twice that of the previously known largest virus (Mimivirus, see Figure 10.5a) and is larger than the genomes of several species of *Bacteria* and *Archaea* (Table 9.1). Pandoravirus infects certain marine amoebae, and with dimensions of $1\ \mu\text{m} \times 0.5\ \mu\text{m}$, is also larger than some bacterial cells (Figure 10.1).

RNA genomes, whether single- or double-stranded, are typically smaller than DNA viruses. Although some viral genomes are larger than those of some prokaryotic cells, genomes of *Bacteria* and *Archaea* are typically much larger than those of viruses (Figure 10.1), and genomes of eukaryotes are the largest of all. Viroids, naked infectious RNAs that cause certain plant diseases (Section 10.15), have the smallest genomes of all microbes (Figure 10.1).

Whether a viral genome is large or small, once a virus has infected its host, transcription of viral genes must occur and new copies of the viral genome must be made. Only later, once viral proteins begin to appear from the translation of viral transcripts, can viral assembly begin. For certain RNA viruses, the genome is also the mRNA. For most viruses, however, viral mRNA must first be made by transcription off of the DNA or RNA genome, and we now consider the variations on how this occurs.

The Baltimore Scheme: DNA Viruses

The American virologist David Baltimore, who shared with the American Howard Temin and Italian American Renato Dulbecco the Nobel Prize for Physiology or Medicine in 1975 for the discovery of retroviruses and their key enzyme, reverse transcriptase, developed a classification scheme for viruses. The scheme is based on the relationship of the viral genome to its mRNA and recognizes seven classes of viruses (Figure 10.2). By convention in virology, viral mRNA is always considered to be of the *plus*

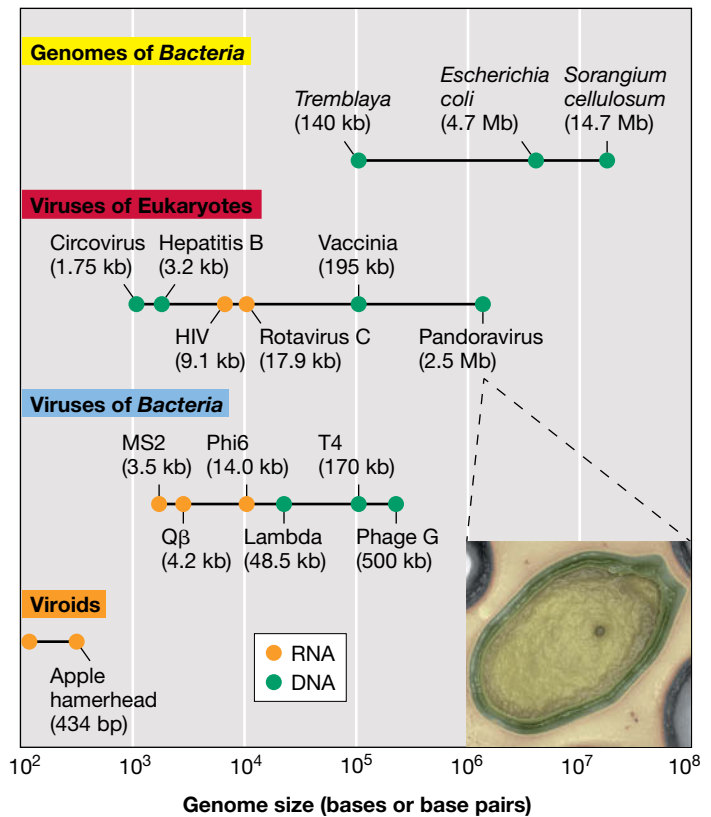


Figure 10.1 Comparative genomics. Genome sizes of select viroids, viruses, and prokaryotic cells. Inset: Micrograph of Pandoravirus, $\sim 1\ \mu\text{m}$ in length. Image courtesy of Chantal Abergel, IGS, UMR7256 CNRS-AMU. Bacteriophage phi6 and phage G infect *Pseudomonas* and *Bacillus* species, respectively; other bacterial viruses infect *Escherichia coli*.

configuration. Thus, to understand the molecular biology of a particular class of virus, one must know the nature of the viral genome and what steps are necessary to produce plus complementarity mRNA from it (Figure 10.2).

Double-stranded DNA viruses are in Baltimore class I. The mechanism of mRNA production and genome replication of class I viruses is the same as that used by the host cell, and we saw this with bacteriophage T4, a typical class I virus (Section 8.6). A virus containing a single-stranded genome may be either a **positive-strand** virus (also called a “plus-strand virus”) or a **negative-strand** virus (also called a “minus-strand virus”). Class II viruses contain single-stranded plus-strand DNA genomes. Transcription of such a genome would yield a message of the minus sense. Therefore, before mRNA can be produced from class II viruses, a complementary DNA strand must first be made to form a double-stranded DNA intermediate; this is called the **replicative form**. The latter is used for transcription and as the source of new genome copies, the plus strand becoming the genome while the minus strand is discarded (Figure 10.2). With only one known exception, all single-stranded DNA viruses are positive-strand viruses.

The Baltimore Scheme: RNA Viruses

The production of mRNA and genome replication will obviously be different for RNA viruses than for DNA viruses. Cellular RNA

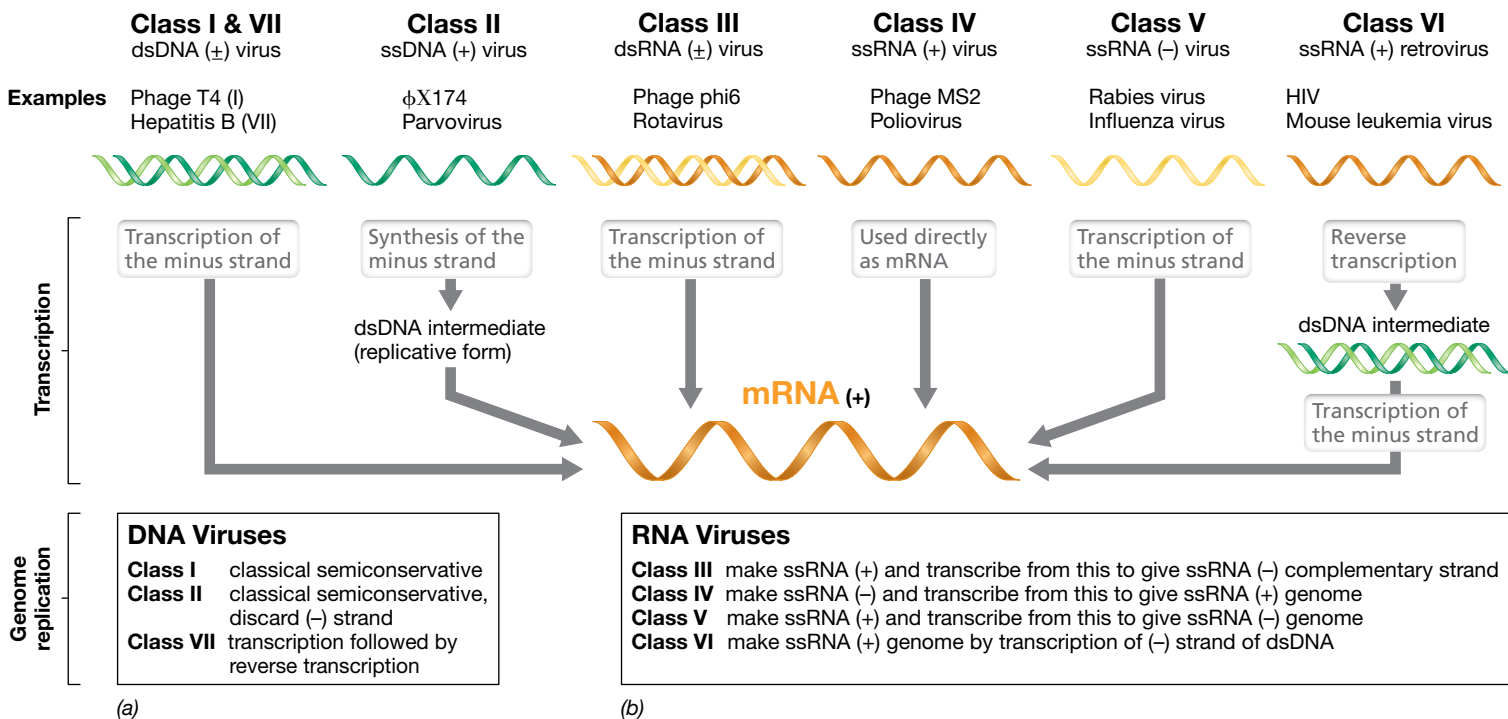


Figure 10.2 The Baltimore classification of viral genomes. Seven classes of viral genomes are known. The genomes can be either (a) DNA or (b) RNA, and either single-stranded (ss) or double-stranded (ds). With the exception of classes V and VI viruses, where the only known examples infect eukaryotic hosts, the top example listed is a bacterial virus and the bottom example an animal virus. The path each viral genome takes to form its mRNA and the strategy each uses for replication is shown.

polymerases do not catalyze the formation of RNA from an RNA template, but instead require a DNA template. Therefore, depending on the virus, RNA viruses must either carry in their virions or encode in their genomes an RNA-dependent RNA polymerase called *RNA replicase* (↔ Section 8.2). With positive-strand RNA viruses (class IV), the genome is also mRNA. But for negative-strand RNA viruses (class V), RNA replicase must synthesize a plus strand of RNA off of the negative-strand template, and the plus strand is then used as mRNA. The plus strand is also used as a template to make more negative-strand genomes (Figure 10.2). RNA viruses of class III face a similar problem but start with double-stranded (+/-) RNA instead of only a positive or negative strand.

Retroviruses are animal viruses whose genomes consist of single-stranded RNA of the plus configuration but which replicate through a double-stranded DNA intermediate (class VI). The process of copying the information found in RNA into DNA is called **reverse transcription** and is catalyzed by an enzyme called *reverse transcriptase*. Human immunodeficiency virus, HIV, is a retrovirus. Finally, class VII viruses are those highly unusual viruses (hepatitis B virus is an example) whose genomes consist of double-stranded DNA but which replicate through an RNA intermediate. As we will see later, these viruses also use reverse transcriptase.

Hosts for Viruses of Each Baltimore Class

Only certain of the Baltimore classes of viruses are known to infect cells of a particular phylogenetic domain. For example, only two

classes of virus are known in *Archaea* and only four in *Bacteria*; only in animals do we find examples of all seven Baltimore classes of viruses (Figure 10.3a). Examples of viruses from each of the Baltimore classes are drawn to scale in Figure 10.3b.

Double-stranded DNA viruses (class I) are the primary viruses infecting prokaryotic cells, while single-stranded plus-sense RNA viruses (class IV) are the major viral predators of eukaryotic cells (Figure 10.3b). As far as is known, fungi are only infected by RNA viruses of classes III and IV, whereas the vast majority of class I viruses that infect eukaryotes replicate in animal hosts rather than plants. By contrast, plants serve as hosts to many more class II viruses than do animals, whereas virtually all class V viruses (viruses with a single-strand minus-sense genome) infect animals rather than plants. And finally, retroviruses (class VI) are known only from animal hosts, while class VII viruses, which like retroviruses depend on reverse transcriptase to replicate their genome, are much more common in plants than in animals (Figure 10.3a). Although the reasons that different genomic classes of virus show specific host preferences is unclear, the fact that *Bacteria* and *Archaea* are hosts for only a relatively small group of viruses suggests that some viral classes may have evolved only later in the timeline of life when more complex eukaryotic hosts were available for infection. However, computational analyses to be discussed in Section 10.2 indicate that this hypothesis is probably incorrect, in that most viral groups, especially the RNA viruses, appear to be of ancient origin.

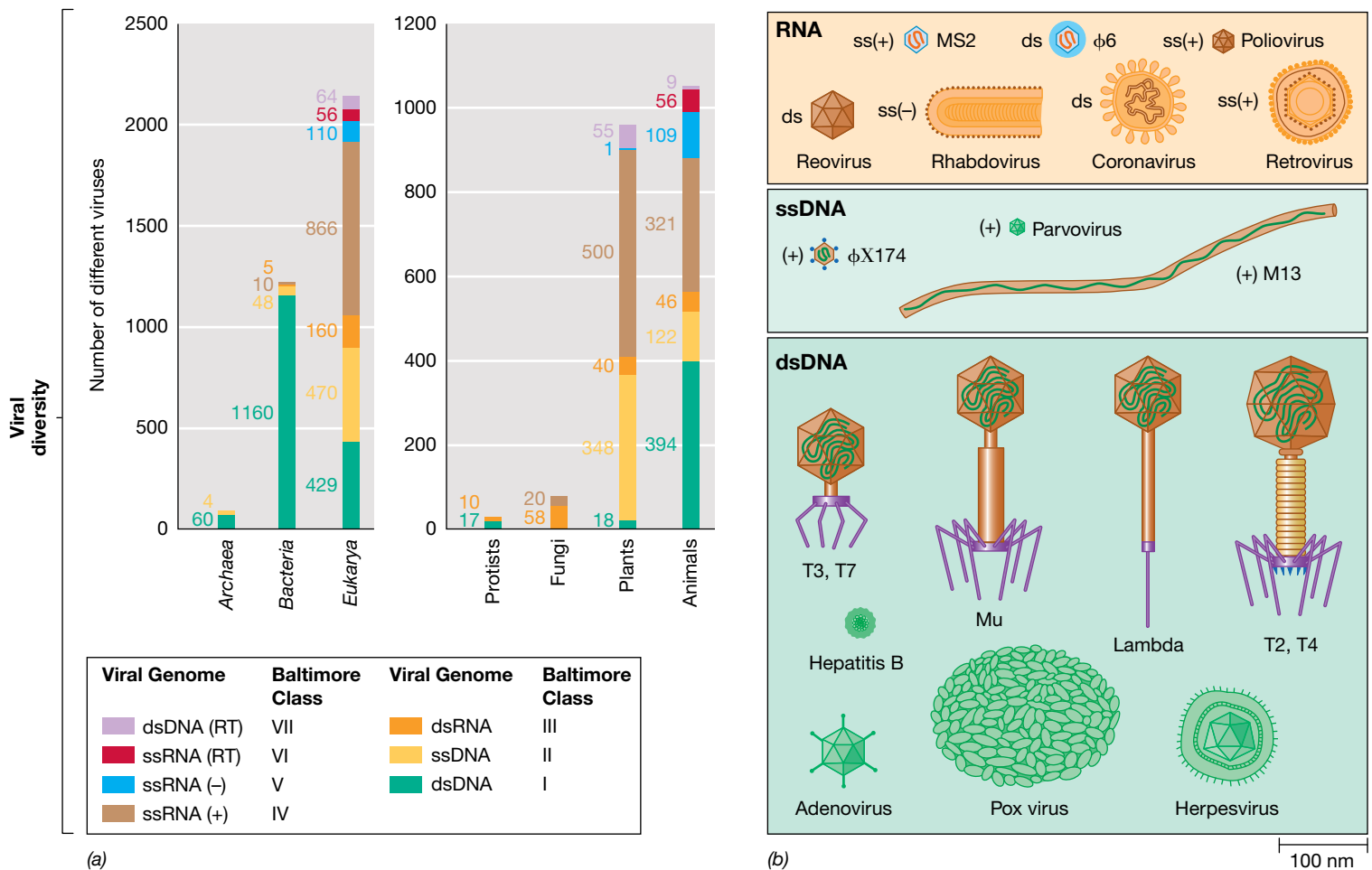


Figure 10.3 Viral hosts and viral diversity. (a) Virus host preferences by cellular domain (left graph) and in different major groups of eukaryotes (right graph). Data adapted from Nasir, A., and G. Caetano-Anollés. 2015. *Sci. Adv.* 2015;1: e1500527. (b) Drawings to scale of several viruses discussed in this chapter.

Viral Protein Synthesis

Once viral mRNA is made (Figure 10.2), viral proteins can be synthesized. In all viruses, these proteins can be grouped into two broad categories: (1) proteins synthesized soon after infection, called *early proteins*, and (2) proteins synthesized later in the infection, called *late proteins*. Both the timing and amount of viral protein synthesis is highly regulated. Early proteins are typically enzymes and are therefore synthesized in relatively small amounts. These include not only nucleic acid polymerases but also proteins that function to shut down host transcription and translation. By contrast, late proteins are typically structural components of the virion and other proteins that are not needed until virion assembly begins, and these are made in much larger amounts (🔗 Section 8.6).

Virus infection upsets the regulatory mechanisms of the host because there is a marked overproduction of viral nucleic acid and protein in the infected cell. Eventually, when the proper proportions of viral genome copies and virion structural components have been synthesized, new virions are assembled—typically spontaneously—and exit the host cell by either lysing and killing it or by a budding process in which the host cell may remain alive.

MINIQUIZ

- Distinguish between a positive-strand RNA virus and a negative-strand RNA virus.
- Contrast mRNA production in the two classes of single-stranded RNA viruses.
- What is unusual about genetic information flow in retroviruses?

10.2 Viral Evolution

When did viruses first appear on Earth and what is their relationship to cells? All known viruses require a host cell for their replication, and this leads to the natural conclusion that viruses evolved at some time *after* cells first appeared on Earth, about 4 billion years ago. Following this line of reasoning, viruses would be remnant cell components that evolved an ability to replicate with assistance from the cell. However, other hypotheses for the origin of viruses have been proposed, including that viruses are relics of the “RNA world,” a period in evolution when RNA is hypothesized to have been the sole carrier of genetic information

(see Section 13.1 and see Figure 10.4), or that viruses were the “first forms of life” on Earth and existed in a precellular era.

Although *how* viruses appeared remains an unanswered question, so is the question of *why* viruses appeared. One likely driver of viral evolution was as a mechanism for cells to quickly move genes about in nature. Because viruses have an extracellular form that protects the nucleic acid inside them, they could have been selected as a means of enriching the genetic diversity (and thus fitness) of their hosts by facilitating gene transfers between them. This function seems especially relevant for prokaryotic cells, where horizontal gene exchange is clearly a major factor in their rapid evolution (see Sections 9.6 and 13.3, Chapter 11). Although many viruses kill their host cell, latent viruses do not, and it is possible that the earliest viruses were primarily latent and evolved lytic capacities only later to more rapidly access new hosts.

Proteomics Support an Early Appearance of Viruses

An experimental analysis of the proteins of a wide variety of viruses has shed new light on how viruses might first have appeared and diversified over time. Because viruses are not cells (and thus do not contain ribosomes), it has been impossible to place viruses on the universal tree of life constructed from comparative ribosomal RNA sequences (see Section 1.13 and Figure 1.36*b*). However, powerful computational methods have recently been deployed to compare the *proteomes* of a large group of viruses with those of cells (the proteome is the total complement of proteins made by a virus or a cell, see Section 9.10). From analyses of proteomic sequence data and protein folding patterns, it has been possible to gain new insights on how viruses first appeared on Earth.

Proteomics point to an origin of viruses from ancient cells that contained segmented RNA genomes and that existed before the last universal common ancestor (LUCA) of modern cells appeared (Figure 10.4*a*). This would have been the era of the “RNA World,” a time preceding the appearance of DNA. Within this line of “virocells,” strong evolutionary pressure for a reduction in both genomic and compartment size eventually eliminated the cellular nature of virocells altogether, leaving only a protein shell to protect the genome from damage. Such reduction would also have triggered a strict dependence in these emerging structures on archaeal, bacterial, or eukaryal cells for replication functions, a characteristic property of viruses. Proteomic analyses point to RNA viruses in general as being older than DNA viruses and, more specifically, to dsRNA viruses (Baltimore class III, Figure 10.2) as being the most ancient of all viruses. Interestingly, many different types of dsRNA viruses (Section 10.10) contain segmented genomes, a possible remnant from their virocell ancestors. Retroviruses also appear to be ancient and may have played a role in the transition from an RNA to a DNA world; we explore this possibility now.

The RNA to DNA Transition

If RNA viruses originated in the scenario just described (Figure 10.4*a*), how did DNA viruses arise? It is thought that some RNA viruses evolved DNA genomes as a mechanism to protect their genomes from cellular ribonucleases—cellular enzymes that destroy foreign RNA. Because DNA is not RNA, these viruses would have had to evolve their own DNA replication machinery to replicate their genomes. It is conceivable that an enzyme like reverse transcriptase

was a key to the conversion of RNA into DNA, just as it is in retroviruses (Baltimore class VI, Figure 10.2) today. It is further hypothesized that DNA viruses then infected the ancestors of the three cellular domains. Gradually, by genetic exchange with DNA viral genomes, each group of cells obtained the machinery necessary to replicate DNA and eventually converted their genomes from RNA-based to DNA-based chemistry.

There are logical reasons for why the transition from RNA to DNA may have occurred. DNA is a more stable molecule than RNA—for example, the spontaneous mutation rate of RNA is much higher than of DNA, and RNA is more susceptible to spontaneous hydrolysis—and this stability would over time have naturally selected for DNA as the genomic repository in cells. This RNA to DNA transition would then have initiated the DNA world we know today (Figure 10.4*b*). The absence of extant cells with RNA

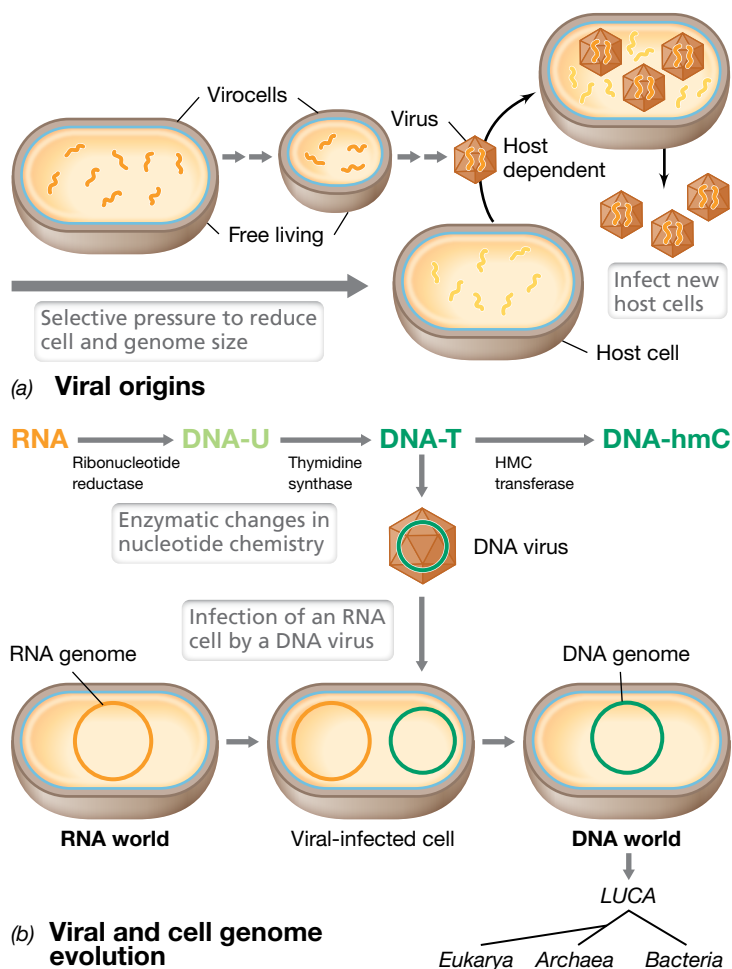


Figure 10.4 Viral origins and the role of viruses in the transition from an RNA world to a DNA world. (a) Viruses are thought to have arisen from primitive “virocells” that contained RNA genomes. Selection for reduced cell size and genomic demands led to the evolution of viruses. (b) The evolution of DNA-specific enzymes would have allowed RNA viruses to become DNA viruses. DNA-U, DNA with uracil (uracil is a base now found mainly in RNA); DNA-T, DNA with thymine (a base found in DNA but not RNA); DNA-hmC, DNA with 5-hydroxycytosine; DNA-U and DNA-hmC are DNA variants known from one virus or another. Infection of an RNA cell by a DNA virus could then have transferred DNA synthetic capacity to the cell, which led to DNA becoming the genomic repository of cells. LUCA, last universal common ancestor.

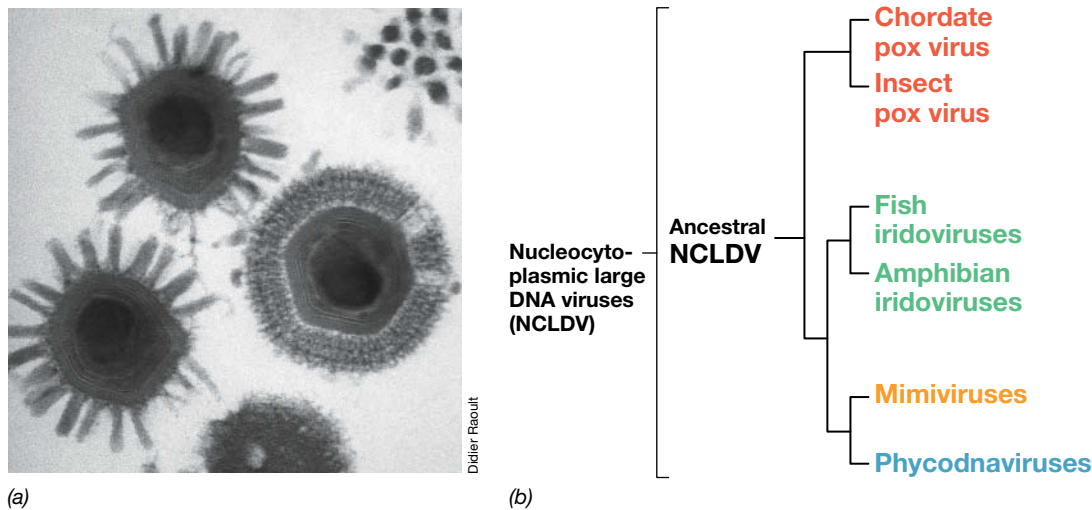


Figure 10.5 Phylogeny of nucleocytoplasmic large DNA viruses (NCLDV). (a) Transmission electron micrograph of Mimivirus, a member of the NCLDV group. A virion is about 0.75 μm in diameter. (b) Phylogeny of major groups of NCLDV based on comparative sequences of several proteins of DNA metabolism. See page 259 for additional coverage of large viruses.

genomes may be because such cells were never infected by DNA viruses and thus never evolved DNA genomes; Darwinian selection would have eventually driven these less fit cells to extinction. However, the fact that some RNA viruses still remain today may actually be a result of their high spontaneous mutation rate because it would allow them to stay one step ahead of evolving host defenses and more quickly adapt to available hosts.

Viral Phylogeny

Using the newly developed proteomic analyses just discussed, it has been possible to place viruses on a universal phylogenetic tree of life constructed from a combination of protein sequences and structural features rather than ribosomal RNA sequences (see Section 1.13). As expected, such trees position viruses at the root of the tree—with RNA viruses preceding DNA viruses—and contain a long branch leading to the three domains of cellular life, with the latter branching out in much the same way as in trees based on ribosomal RNA. The branching order of viruses in the proteomics tree remains a bit fuzzy, but the analyses clearly show RNA viruses to lie basal to DNA viruses in agreement with hypotheses for how DNA viruses arose and how DNA replaced RNA in cellular genomes (Figure 10.4).

In only a few groups of viruses has it been possible to reliably trace phylogenies more precisely, and in these cases, trees have been assembled from sequences of a group of genes or proteins shared in common among the group. One such example is Mimivirus and its relatives, one of the larger known viruses (Figure 10.5). Mimivirus capsids are multilayered and icosahedral. The virion is surrounded by spikes and is nearly 0.75 μm in diameter, larger than some prokaryotic cells (Figure 10.5a). Mimivirus contains a 1.2-megabase-pair genome consisting of double-stranded DNA. Mimivirus infects the protozoan *Acanthamoeba* and belongs to a group of giant viruses with large genomes called *nucleocytoplasmic large DNA viruses* (NCLDV) (Figure 10.5b). The NCLDV comprise several virus families, including pox viruses (Section 10.6), iridoviruses, and certain plant viruses. These viruses share a set of highly homologous

proteins, most of which function in DNA metabolism. A phylogenetic tree of these viruses constructed from DNA sequences encoding these proteins shows how they have diverged from a common ancestor (Figure 10.5b). It is thus possible to track the phylogeny of particular viral groups with some confidence, but to do so, one needs to start with a group that is already known to share a number of properties in common.

It is clear that diversity in the viral world is enormous and that obtaining a detailed phylogenetic tree of all viruses will remain a challenge. The continual isolation of highly unusual new viruses makes this difficult task even more challenging. For example, only about 7% of the genes of

Pandoravirus (Figure 10.1) have gene homologs in existing genomic databases. What this means is that over 90% of the genome of this giant virus will likely be new to biology—a striking example of what awaits discovery in the fascinating world of viruses.

MINIQUIZ

- How could viruses have accelerated the evolution of cells?
- Explain how viruses could have “invented” the genetic material found in all cells.
- Lacking ribosomes, how can viruses be placed on the universal tree of life?

II • DNA Viruses

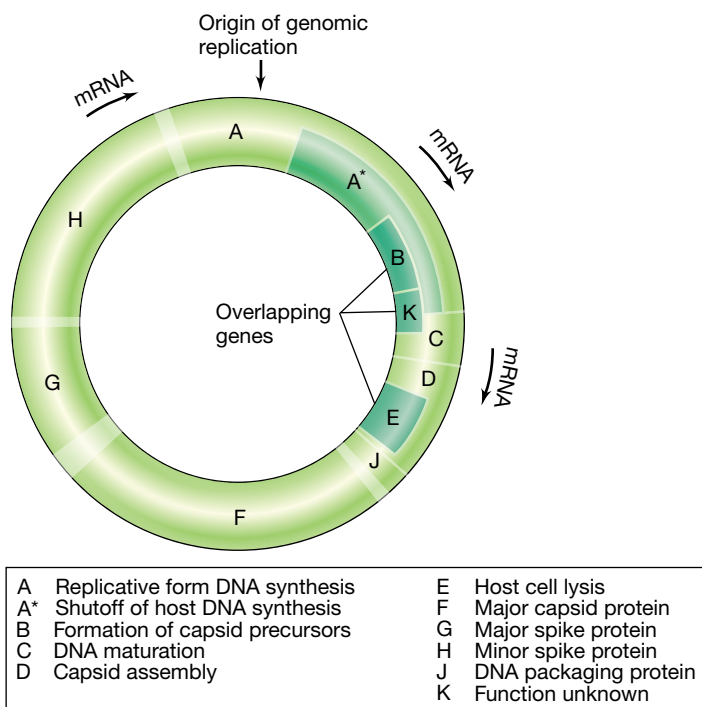
Although DNA viruses likely appeared later in evolutionary times than RNA viruses (Section 10.2), DNA viruses today infect a wide variety of organisms, in particular, species of *Bacteria* and *Archaea*. In fact, the majority of viruses that infect prokaryotic cells are DNA viruses, mainly of the double-stranded variety (Figure 10.3). We examine several of these here along with some DNA viruses that infect eukaryotes, and keep our focus on the processes involved in transcription and genome replication in DNA viruses of different genomic makeups.

10.3 Single-Stranded DNA Bacteriophages: ϕ X174 and M13

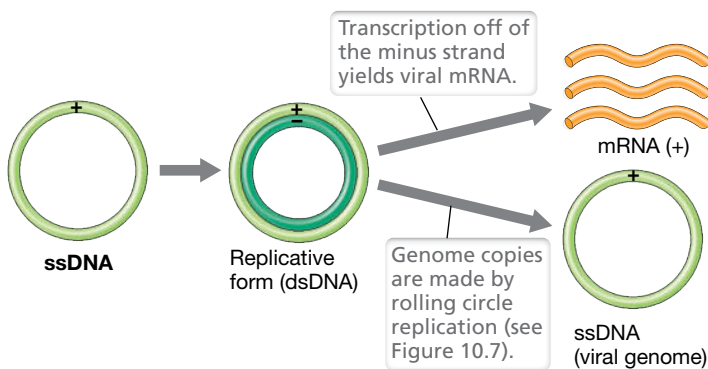
In this section we discuss two single-stranded DNA bacteriophages, ϕ X174 and M13. Many single-stranded DNA plant and animal viruses are also known, and because these share with bacterial viruses the fact that their genomes are of the plus complementarity (Baltimore class II, Figure 10.2a), many molecular events are similar. Hence, our focus here will be on the phages.

Bacteriophage ϕ X174

Bacteriophage ϕ X174 contains a circular genome of 5386 nucleotides inside a tiny icosahedral virion, about 25 nm in diameter. Phage ϕ X174 has only a few genes and shows the phenomenon of **overlapping genes**, a condition in which there is insufficient DNA to encode all viral-specific proteins unless parts of the genome are read more than once in different reading frames. For example, in the ϕ X174 genome, gene B resides within gene A, and gene K resides within both genes A and C (Figure 10.6a). Genes D and E also overlap, gene E being contained completely within gene D. Also, the termination codon of gene D overlaps the initiation codon of gene J (Figure 10.6a).



(a) Genetic map of ϕ X174



(b) Flow of events during ϕ X174 replication

Figure 10.6 Bacteriophage ϕ X174, a single-stranded DNA phage. (a) Genetic map. Note regions of gene overlap. Protein A* is formed using only part of the coding sequence of gene A by reinitiation of translation. The key indicates the functions of the proteins encoded by each gene. Unlabeled parts of the chromosome are regions of noncoding DNA. (b) Genetic information flow in ϕ X174. Progeny single-stranded DNA is produced from the replicative form by rolling circle replication (see Figure 10.7).

The distinct gene products from overlapping genes are made by reinitiating transcription *in a different reading frame* within a gene to yield a second (and distinct) transcript. In addition to overlapping genes, a small protein in ϕ X174 called A*, which functions to shut down host DNA synthesis, is synthesized by the reinitiation of *translation* (not transcription) within the mRNA for gene A. The A* protein is read from the same mRNA reading frame as A protein but has a different in-frame start codon and is thus a shorter protein.

Before a single-stranded DNA genome can be transcribed, a complementary strand of DNA must be synthesized, forming a double-stranded molecule called the replicative form. This can then be used as a source of both mRNA and genome copies. Upon infection of an *Escherichia coli* cell by ϕ X174, the viral DNA is separated from the protein coat and the genome is converted into the replicative form by host enzymes. From this, several copies are made by semiconservative replication, and phage-specific transcripts are made by transcription off of the negative strand of the replicative form (Figure 10.6b). The replicative form is also the starting point for making copies of the phage genome by a mechanism we have already seen used in phage lambda (Section 8.7): **rolling circle replication** (Figure 10.7).

In the synthesis of the ϕ X174 genome, the rolling circle facilitates the continuous production of positive strands from the replicative form. To do this, the positive strand of the latter is nicked and the 3' end of the exposed DNA is used to prime synthesis of a new strand (Figure 10.7). Cutting of the plus strand is accomplished by the A protein (Figure 10.6a). Continued rotation of the circle leads to the synthesis of a linear ϕ X174 genome. Note that rolling circle synthesis differs from semiconservative replication (Section 4.3) because only the negative strand serves as a template.

When the growing viral strand reaches unit length (5386 residues for ϕ X174), the A protein cleaves it and then ligates the two ends of the newly synthesized single strand to give a single-stranded DNA circle. Ultimately, assembly of mature ϕ X174 virions occurs and cell lysis follows. The E protein (Figure 10.6a) promotes cell lysis by inhibiting the activity of an enzyme in peptidoglycan synthesis (Section 7.5) in the host cell. Because of the resulting weakness in newly synthesized cell wall material, the host cell ruptures, releasing the phage virions.

Bacteriophage M13

Bacteriophage M13 is a filamentous virus with helical symmetry; the virion is long and thin and attaches to the pilus of its host cell (Section 8.5). Filamentous phages such as M13 have the unusual property of being released from the host cell without the cell undergoing lysis; infected cells continue to grow, and typical viral plaques (Section 8.4) are not observed. To facilitate the nonlytic release, M13 DNA is covered with coat proteins as it exits across the cell envelope. Four minor coat proteins cover the tips of the virion while the major coat protein covers the sides (Figure 10.8). Thus with M13, there is no intracellular accumulation of mature virions as occurs with typical lytic bacteriophages. Instead, these filamentous bacteriophages cause chronic infections.

Several features of phage M13 have made it useful as a cloning and DNA sequencing vehicle in the past. For example, many aspects of DNA replication in M13 are similar to those of ϕ X174 and the genome is very small; this facilitates sequencing efforts.

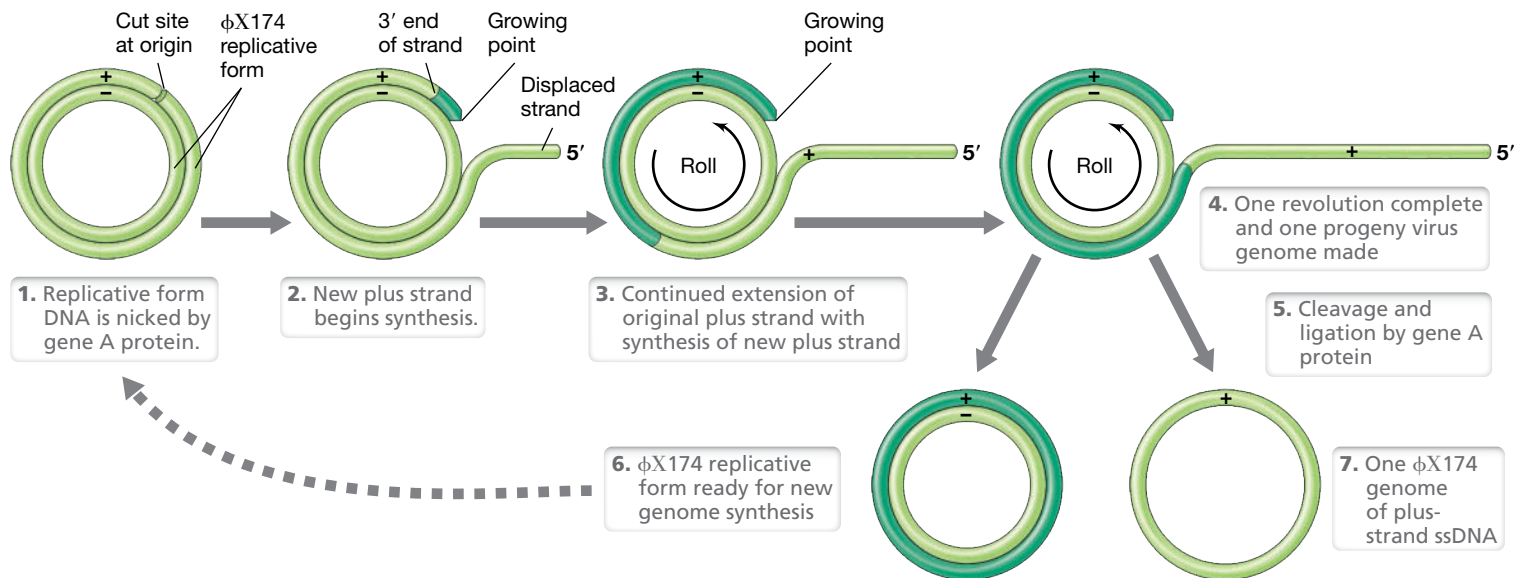


Figure 10.7 Rolling circle replication in phage ϕ X174. Replication begins at the origin of the double-stranded replicative form with the cutting of the plus strand of DNA by gene A protein (both strands of DNA are shown in light green here for simplification). After one new progeny strand has been synthesized (one revolution of the circle), the gene A protein cleaves the new strand and ligates its two ends.

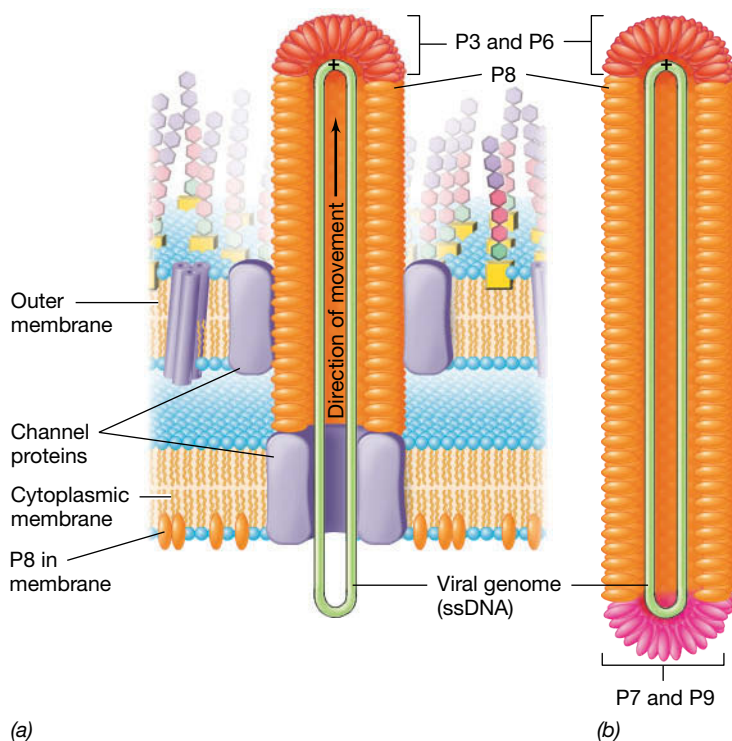


Figure 10.8 Release of phage M13. The virions of phage M13 exit infected cells without lysis. (a) Budding. The virus DNA crosses the cell envelope through a channel constructed from virus-encoded proteins. As this occurs, the DNA is coated with phage proteins that have been embedded in the cytoplasmic membrane. (b) Complete virion. The two ends of the virion are covered with small numbers of the minor coat proteins P3 and P6 (front end) or P7 and P9 (rear end). Because bacteriophage M13 is a single-stranded DNA phage, it was widely used in the past as a tool for molecular cloning and DNA sequencing (Chapters 9 and 12).

Second, a double-stranded form of genomic DNA essential for cloning purposes is produced naturally when M13 produces its replicative form. Third, as long as infected cells are kept growing, phage can be produced indefinitely, yielding a continuous source of the cloned DNA. These and other features of M13 made this phage a workhorse of the genetic engineering field for many years, although today M13 has been replaced for most genetic engineering tasks by a variety of even more convenient and useful tools.

MINIQUIZ

- Why is formation of the replicative form of ϕ X174 necessary in order to make phage-specific mRNA?
- In the ϕ X174 genome, describe the difference between how the gene B and gene A* proteins are made.
- How can M13 virions be released without killing the infected host cell?

10.4 Double-Stranded DNA Bacteriophages: T7 and Mu

The double-stranded DNA (dsDNA) (Baltimore class I, Figure 10.2) bacteriophages are among the best studied of all viruses, and we have already discussed two important ones, T4 and lambda, in Chapter 8. However, because of their importance in molecular biology, gene regulation, and genomics, we consider two more such viruses here, T7 and Mu, each of which has features distinct from those of T4 and lambda.

Bacteriophage T7

Bacteriophage T7 is a relatively small DNA virus that infects *Escherichia coli* and a few related enteric bacteria. The virion has

an icosahedral head and a very short tail, and the T7 genome is a linear double-stranded DNA molecule of about 40 kilobase pairs.

When a T7 virion attaches to a host cell and the DNA is injected, early genes are quickly transcribed by host RNA polymerase and then translated. One of these early proteins inhibits the host restriction system, a mechanism for protecting the cell from foreign DNA (see Section 8.5). This occurs very rapidly, as the T7 anti-restriction protein is made and becomes active before the entire T7 genome has entered the cell. Other early proteins include a T7 RNA polymerase and proteins that inhibit host RNA polymerase activity. T7 RNA polymerase recognizes only T7 gene promoters distributed along the T7 genome. This transcriptional strategy differs from that of phage T4 because T4 uses the host RNA polymerase throughout its replication cycle but modifies the host polymerase such that it recognizes only phage genes (see Section 8.6).

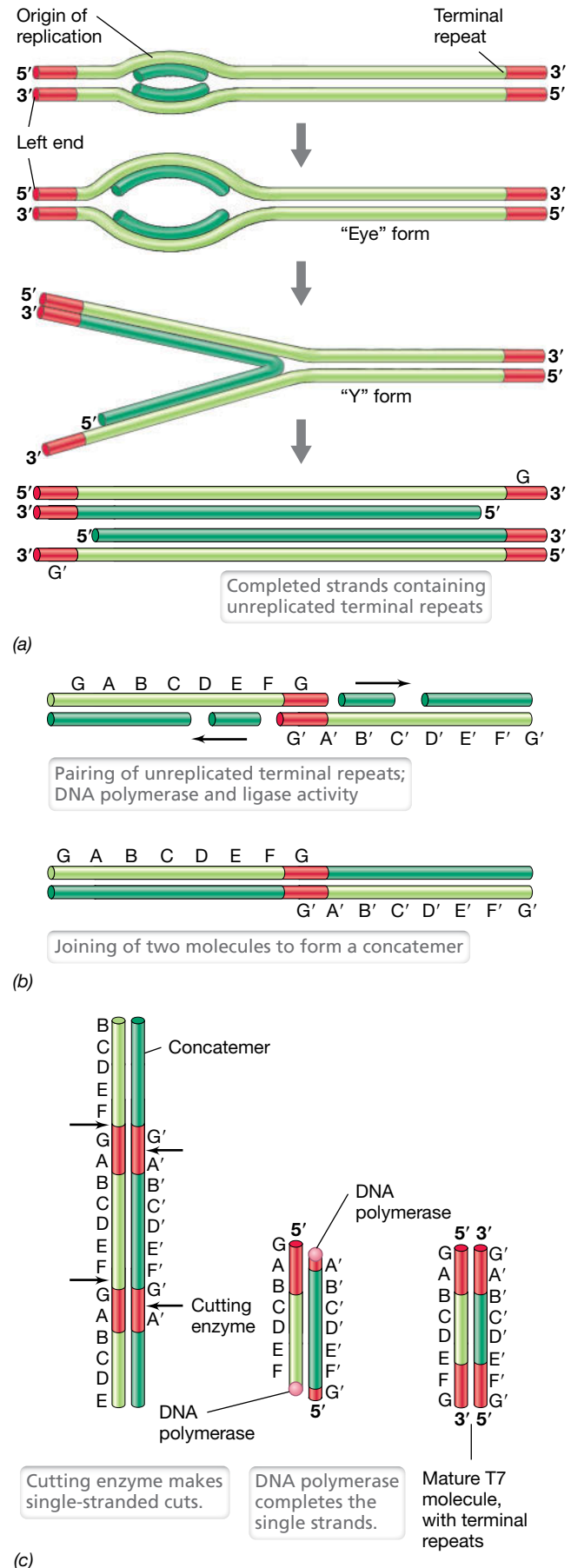
Genome replication in T7 begins at an origin of replication within the molecule and proceeds bidirectionally from this point (Figure 10.9a). Phage T7 uses its own DNA polymerase, which is a composite protein including one polypeptide encoded by the phage and one by the host. As in phage T4, T7 DNA contains terminal repeats at both ends of the molecule and these are eventually used to form *concatemers* (Figure 10.9b). Continued replication and recombination leads to concatemers of considerable length, but ultimately a phage-encoded endonuclease cuts each concatemer at a specific site, resulting in the formation of linear DNA molecules with terminal repeats that are packaged into phage heads (Figure 10.9c). However, because T7 endonuclease cuts the concatemer at specific sequences, the DNA sequence in each T7 virion is identical. This differs from the situation in phage T4, where DNA concatemers are processed using a “headful mechanism” that generates circularly permuted genomes (see Section 8.6).

Bacteriophage Mu

Like bacteriophage lambda (see Section 8.7), bacteriophage Mu is a temperate phage but has the unusual property of replicating by *transposition*. Transposable elements are sequences of DNA that can move within their host genome from one location to another as discrete genetic units (see Section 11.11); transposition is facilitated by an enzyme called **transposase**. Mu was so named because it generates *mutations* when it integrates into the host cell chromosome, and thus it has been useful in bacterial genetics because it can generate mutants easily.

Bacteriophage Mu has an icosahedral head, a helical tail, and several tail fibers (Figure 10.10a). The genome of Mu consists of linear double-stranded DNA, and most Mu genes encode head and tail proteins, other replication factors such as the Mu transposase, and factors that affect host range. Host range is controlled by the kind of tail fibers that are made, with one type allowing only infection of *E. coli* while the other type allows the phage to infect several other enteric bacteria as well.

Figure 10.9 Replication of the bacteriophage T7 genome. (a) The linear, double-stranded DNA undergoes bidirectional replication, giving rise to intermediate “eye” and “Y” forms (for simplicity, both template strands are shown in light green and both newly synthesized strands in dark green). (b) Formation of concatemers by joining DNA molecules at their unreplacated terminal ends. (c) Production of mature viral DNA molecules from T7 concatemers by activity of the cutting enzyme, an endonuclease.



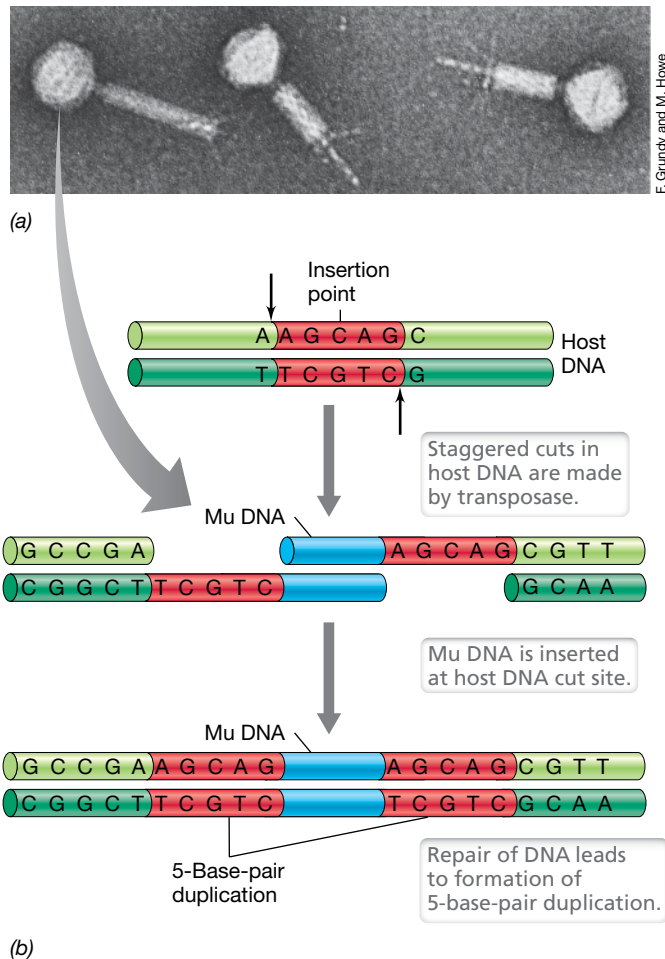


Figure 10.10 Bacteriophage Mu. (a) Electron micrograph of virions of the double-stranded DNA phage Mu. (b) Integration of Mu into the host DNA, showing the generation of a 5-base-pair duplication of host DNA.

Phage Mu replicates in a completely different manner from all other bacteriophages because its genome is replicated as part of a larger DNA molecule (Figure 10.10b). Thus, integration of Mu DNA into the host genome is essential for both lytic and lysogenic development. Integration requires the activity of Mu transposase, and a 5-base-pair fragment of host DNA is duplicated at the target site where Mu DNA is integrated. This host DNA duplication arises because staggered cuts are made at the point in the host genome where Mu DNA is inserted. The resulting single-stranded segments are converted to the double-stranded form as part of the Mu integration process (Figure 10.10b).

Phage Mu can enter the lytic pathway upon initial infection if its repressor is *not* made; alternatively, Mu can form a lysogen if its repressor is made. In either case, Mu DNA is replicated by repeated transposition of Mu to multiple sites on the host genome. If the lytic cycle pathway is triggered, only the early genes of Mu are initially transcribed. Then, following expression of a Mu transcriptional activating protein, Mu head and tail proteins are synthesized. Following self-assembly, the cell is lysed and mature Mu virions are released. The lysogenic state in Mu requires that sufficient Mu repressor protein be present to prevent transcription of integrated Mu DNA.

MINIQUIZ

- In what major way does transcription of phage DNA differ in phages T4 and T7?
- What is unusual about the replication mechanism of the Mu genome?

10.5 Viruses of Archaea

Many bacteriophages and archaeal viruses have been isolated and characterized thus far. For *Bacteria*, these include both DNA and RNA phages, some with single-stranded and others with double-stranded genomes. However, all characterized archaeal viruses have DNA genomes, and with rare exception, double-stranded circular DNA genomes (Figure 10.3a).

DNA Archaeal Viruses

Several DNA viruses have been discovered whose hosts are species of *Archaea*, including representatives of both the *Euryarchaeota* and *Crenarchaeota* phyla (Chapter 17). Most viruses that infect species of *Euryarchaeota*, including both methanogenic and halophilic *Archaea*, are of the “head and tail” type, resembling the structurally complex phages that infect enteric bacteria, such as phage T4. One novel archaeal virus infects a halophile and is unusual because it is both enveloped and contains a single-stranded DNA genome. By contrast, all other characterized archaeal DNA viruses contain double-stranded and typically circular DNA genomes.

The most distinctive archaeal viruses infect hyperthermophilic *Crenarchaeota*. For example, the sulfur chemolithotroph *Sulfolobus* is host to several structurally unusual viruses. One such virus, called SSV, forms spindle-shaped virions that often cluster in rosettes (Figure 10.11a). Such viruses are widespread in acidic hot springs worldwide. Virions of SSV contain a circular DNA genome of about 15 kilobase pairs. A second morphological type of *Sulfolobus* virus forms a rigid, helical rod-shaped structure (Figure 10.11b). Viruses in this class, nicknamed *SIFV*, contain linear DNA genomes about twice the size of that of SSV. Many variations on the spindle- and rod-shaped patterns have been seen in archaeal viral isolation studies, and a few species of *Crenarchaeota* are even infected by filamentous viruses.

A spindle-shaped virus that infects the hyperthermophile *Acidianus* displays a novel behavior. The virion, called ATV, contains a circular genome of about 68 kilobase pairs and is lemon-shaped when first released from the host cells. However, shortly after release from its lysed host cell, the virion produces long, thin tails, one at each end (Figure 10.11d). The tails are actually tubes, and as they form, the virion becomes thinner and its volume is reduced. Remarkably, this is the first example of virus development in the complete absence of host cell contact. It is thought that the extended tails of ATV help the virus in some way survive in its hot (85°C), acidic (pH 1.5) environment. This unusually shaped virus is also lysogenic, a property rarely seen in other archaeal viruses.

A spindle-shaped virus also infects the hyperthermophile *Pyrococcus* (*Euryarchaeota*). This virus, named PAV1, resembles SSV but is larger and contains a very short tail (Figure 10.11c). PAV1 has a small circular DNA genome and is released from host cells without cell lysis, probably by a budding mechanism similar to that of the *Escherichia coli* bacteriophage M13 (Section 10.3).

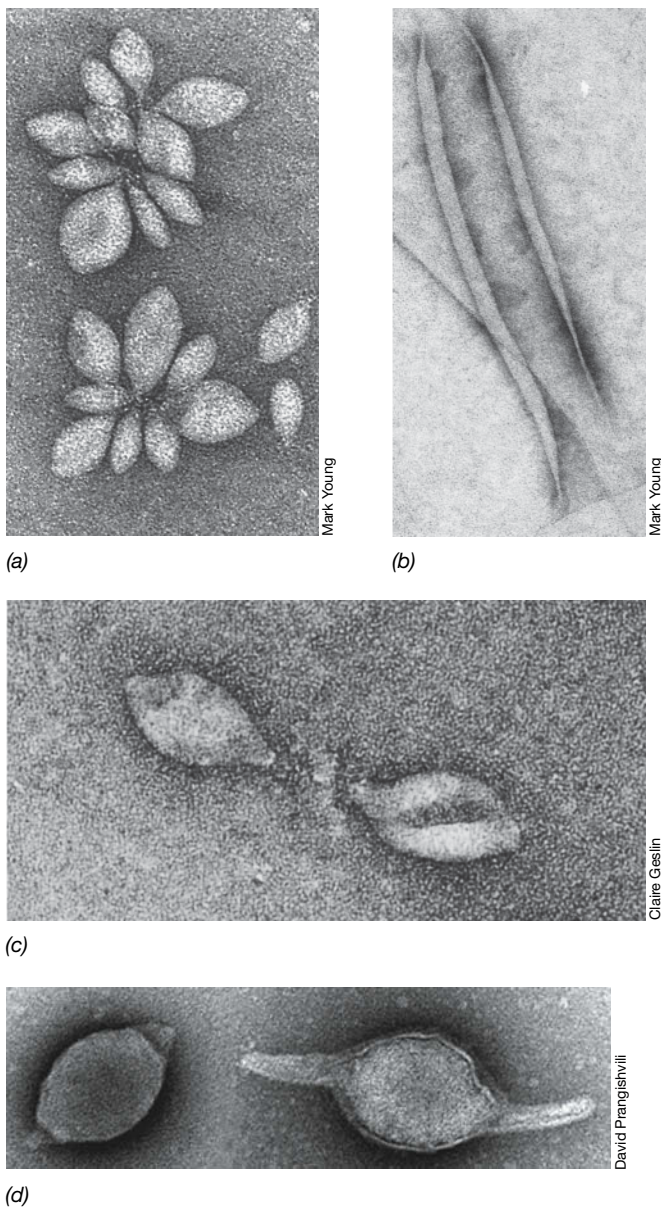


Figure 10.11 Archaeal viruses. Electron micrographs of viruses of *Crenarchaeota* (parts *a*, *b*, *d*), and a virus of a euryarchaeote (*c*). (*a*) Spindle-shaped virus SSV1 that infects *Sulfolobus solfataricus* (virions are 40×80 nm). (*b*) Filamentous virus SIFV that infects *S. solfataricus* (virions are 50×900 – 1500 nm). (*c*) Spindle-shaped virus PAV1 that infects *Pyrococcus abyssi* (virions are 80×120 nm). (*d*) ATV, the virus that infects the hyperthermophile *Acidianus convivator*. When released from the cell the virions are lemon-shaped (left), but they proceed to grow appendages on both ends (right). ATV virions are about 100 nm in diameter.

Pyrococcus has a growth temperature optimum of 100°C and thus PAV1 virions must be extremely heat-stable. Despite their similar morphologies, genomic comparisons of PAV1 and SSV-type viruses show little sequence similarity, indicating that the two types of viruses do not have common evolutionary roots.

RNA Archaeal Viruses

Thus far, RNA viruses that can replicate in the laboratory on archaeal hosts are unknown, despite the fact that a variety of RNA viruses infect *Bacteria* and eukaryotes (Figure 10.3). Although no

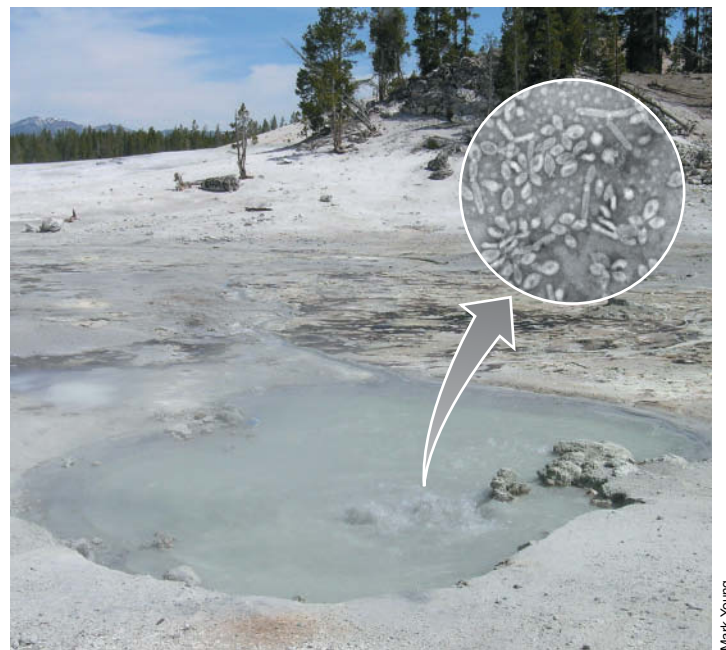


Figure 10.12 An acidic Yellowstone hot spring and its archaeal viruses. Inset: Transmission electron micrograph of a mixture of archaeal viruses from the spring. Compare with Figure 10.11.

concrete examples of archaeal RNA viruses have emerged, environmental genomics have shown that they almost certainly exist. In some acidic hot springs of Yellowstone National Park that support large communities of *Crenarchaeota*, a large number of unusually shaped and structurally tough archaeal viruses have been discovered (Figure 10.12) and grown in the laboratory. Thus far, all of these have been DNA viruses. However, using the powerful tools of metagenomics (see Section 9.8), researchers studying these hot springs have discovered viral RNAs whose RNA sequences bear no resemblance to those of any known RNA viruses that infect *Bacteria*. Because these springs are too hot for eukaryotes and cell numbers of *Bacteria* are few, the unusual RNA is almost certainly from RNA archaeal viruses that are yet to be propagated in the laboratory.

Sequence analyses of the hot spring viral RNA show that it originated from single-stranded plus-sense RNA viruses (Baltimore class IV, Figure 10.2) (Section 10.8). These viral genomes also encode an RNA replicase—a hallmark of RNA viruses—and are likely to replicate by way of polyprotein formation, a replication mechanism employed by some class IV viruses of eukaryotes, such as poliovirus (Section 10.8). Replication steps of the putative RNA archaeal viruses, including important molecular details such as the extent to which viral (rather than host) polymerases participate in the replication process, are unclear and await laboratory cultivation of the viruses. However, now that scientists know that such viruses almost certainly exist, they can be on the lookout for them in viral enrichment and isolation studies.

MINIQUIZ

- What type of genome is seen in most archaeal viruses?
- Compared with other archaeal viruses, what are two unusual features of the virus that infects *Acidianus*?

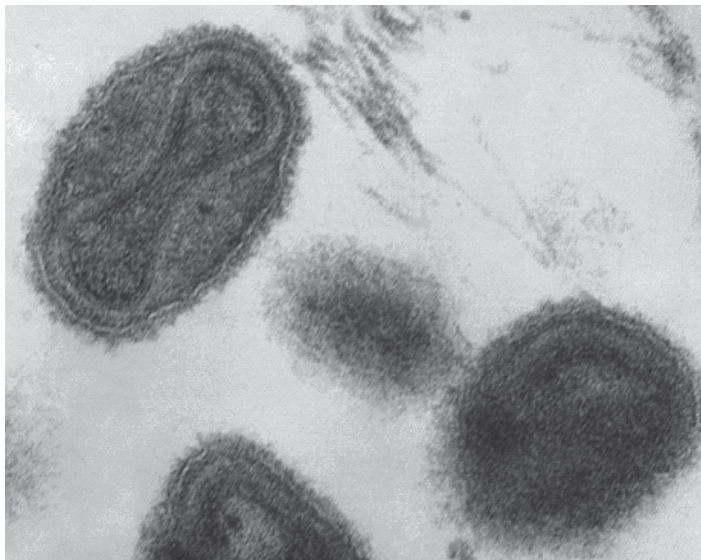
10.6 Uniquely Replicating DNA Animal Viruses

Two groups of double-stranded DNA (Baltimore class I, Figure 10.2) animal viruses show unusual replication strategies: pox viruses and adenoviruses. Pox viruses are unique because all replication events, including DNA replication, occur in the host *cytoplasm* instead of the nucleus, and adenoviruses are unique because the replication of their genome proceeds in a leading fashion on *both* DNA template strands.

Pox Viruses

Pox viruses have been important historically as well as medically. Smallpox virus was the first virus to be studied in any detail and was the first virus for which a vaccine was developed (over 200 years ago the British physician Edward Jenner was the first to protect people from infection by smallpox virus by exposing them to the similar but much less virulent cowpox virus). Pox viruses are among the largest of all viruses, the brick-shaped vaccinia virions measuring almost 400 nm in diameter (Figure 10.13). Other pox viruses of importance are cowpox and vaccinia virus. Because it closely resembles the smallpox virus but is not pathogenic, vaccinia is used as a smallpox vaccine today and as a laboratory model for smallpox virus molecular biology.

The vaccinia virus genome consists of linear double-stranded DNA about 190 kilobase pairs in length and encoding about 250 genes. Following attachment, vaccinia virions are taken up into host cells and the nucleocapsids (Figure 10.13) are liberated in the cytoplasm; all replication events take place in the cytoplasm. Uncoating of the viral genome requires the activity of a viral protein that is synthesized after infection (the gene encoding this protein is transcribed by a viral RNA polymerase contained within



CDC/PHIL, Fred Murphy and Sylvia Whitfield

Figure 10.13 Smallpox virus. Transmission electron micrograph of a negatively stained thin section of smallpox virus virions. The virions are approximately 350 nm (0.35 μm) long. The dumbbell-shaped structure inside the virion is the nucleocapsid, which contains the double-stranded DNA genome. All replication functions for pox virus occur in the host cytoplasm.

the virion). In addition to this uncoating gene, a number of other viral genes are transcribed, including genes that encode a DNA polymerase that synthesizes copies of the viral genome. These are then incorporated into virions that accumulate in the cytoplasm, and the virions are released when the infected cell lyses.

Vaccinia virus has been genetically engineered to contain certain proteins from other viruses for use in recombinant vaccines (see Section 12.8). A vaccine is a substance capable of eliciting an immune response in an animal that protects the animal from future infection with the same agent. Vaccinia virus causes no serious health effects in humans but elicits a strong immune response. Therefore, as a carrier of proteins from pathogenic viruses, vaccinia virus is a relatively safe and effective tool for stimulating an immune response against these pathogens. Success has been obtained with vaccinia virus vaccines against the viruses that cause influenza, rabies, herpes simplex type 1, and hepatitis B.

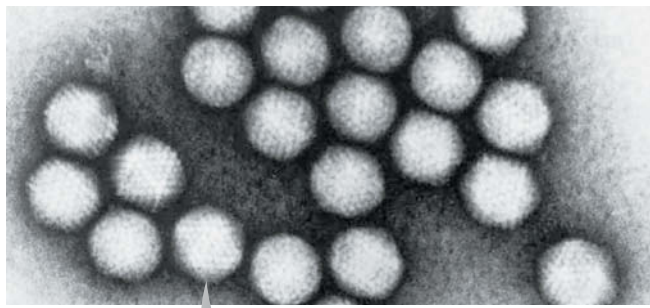
Adenoviruses

Adenoviruses are a group of small and naked icosahedral viruses (Figure 10.14a) that contain linear double-stranded DNA genomes. Adenoviruses are of minor health importance, causing mild respiratory infections in humans, but they have unique stature in virology because of the mechanism by which they replicate their genomes. Attached to the 5' end of adenoviral genomic DNA is a protein called the adenoviral *terminal protein*, and it is essential for replication of the adenoviral genome. The complementary DNA strands also have inverted terminal repeats that play a role in the replication process (Figure 10.14b).

Following infection, the adenoviral nucleocapsid is released into the host cell nucleus, and transcription of the early genes proceeds by activity of the host RNA polymerase. Most early transcripts encode important replication proteins such as the terminal protein and a viral DNA polymerase. Replication of the adenoviral genome begins at either end of the DNA genome and the terminal protein facilitates this process because it contains a covalently bound cytosine that functions as a primer for DNA polymerase (Figure 10.14b). The products of this initial replication are a completed double-stranded viral genome and a single-stranded minus-sense DNA molecule. At this point, a unique replication event occurs. The single DNA strand cyclizes by means of its inverted terminal repeats, and a complementary (plus-sense) DNA strand is synthesized beginning from its 5' end (Figure 10.14b). This mechanism is unique because double-stranded DNA is replicated *without the formation of a lagging strand*, as occurs in conventional semi-conservative DNA replication (see Section 4.3). Once sufficient copies of the adenoviral genome have formed and virion structural components accumulate in the host cell, mature adenoviral virions are assembled and released from the cell following lysis.

MINIQUIZ

- What is unusual about genome replication in pox viruses?
- What is unusual about genome replication in adenoviruses?
- Why is the adenovirus terminal protein essential for replicating its genome?



CDC/PHIL. G. William Gary, Jr.

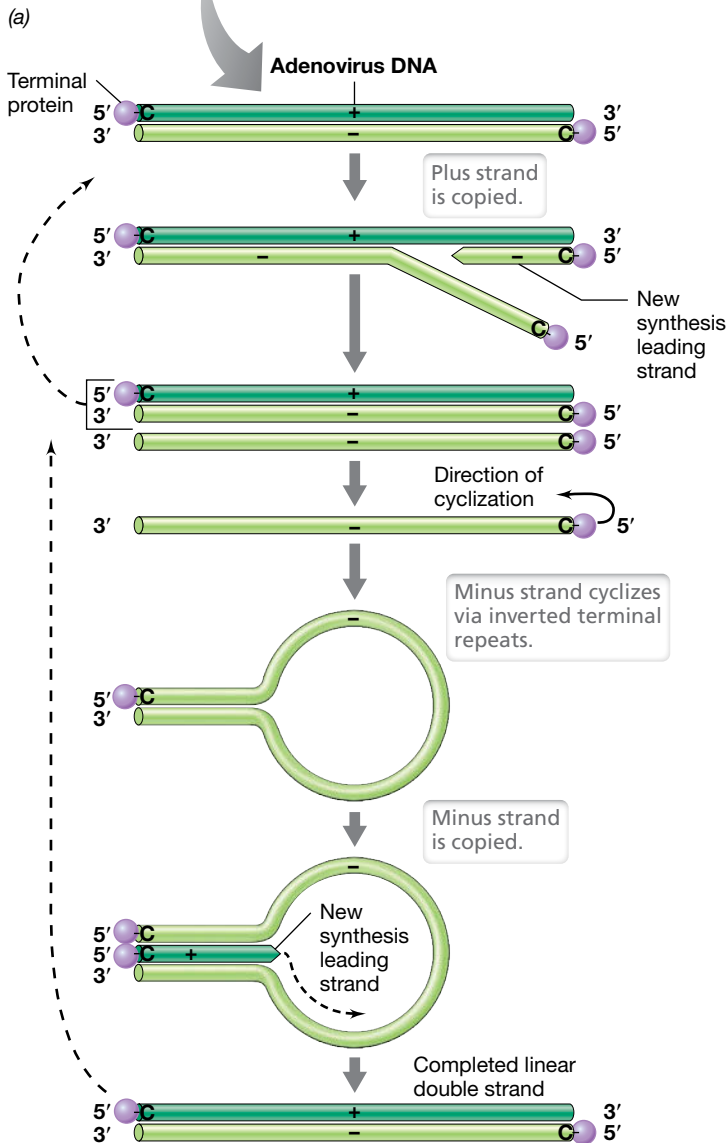


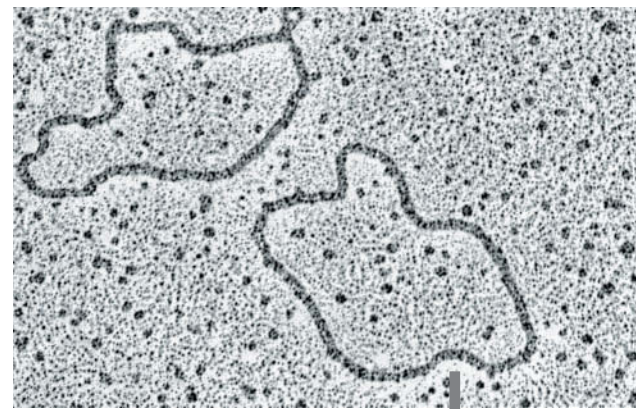
Figure 10.14 Adenoviruses and their genomic replication. (a) Transmission electron micrograph of adenoviral virions. Note the icosahedral structure. (b) Adenoviral genome replication. Because of loop formation (cyclization), there is no lagging strand; DNA synthesis is leading on both strands. A cytosine (C) is attached to the terminal protein. Adenoviruses are one of several classes of human viruses that cause upper respiratory infections such as the common cold (see Section 30.7). Rhinoviruses (single-stranded plus-sense RNA viruses, see Section 10.8) cause the vast majority of colds.

10.7 DNA Tumor Viruses

Besides catalyzing lytic events or becoming integrated into a genome in a latent state, some DNA animal viruses can induce cancers. These include viruses of the polyomavirus family and some herpesviruses, both of which contain double-stranded DNA genomes (Baltimore class II, Figure 10.2).

Polyomavirus SV40

Polyomavirus SV40 is a naked icosahedral virus that can cause tumors in small mammals, such as hamsters and rats. Its circular genome consists of double-stranded DNA (Figure 10.15a). The genome is too



Alexander EB and Jerome Vinograd

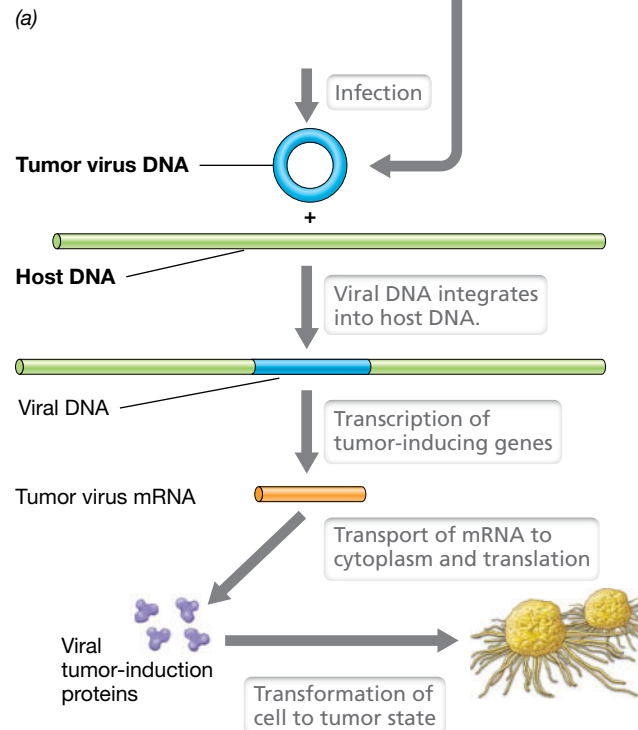


Figure 10.15 Polyomaviruses and tumor induction. (a) Transmission electron micrograph of relaxed (nonsupercoiled) circular DNA from a tumor virus. The contour length of each circle is about 1.5 μm . (b) Events in cell transformation by a polyomavirus such as SV40. Viral DNA becomes incorporated into the host genome. From there, viral genes encoding cell transformation events are transcribed and transported to the cytoplasm for translation.

small (5.2 kb) to encode its own DNA polymerase, so host DNA polymerases are used and SV40 DNA is replicated in a bidirectional fashion from a single origin of replication. Because of the small genomes of polyomaviruses, the strategy of overlapping genes, typical of many small bacteriophages (Sections 10.3 and 10.8), is also employed here. Transcription of the viral genome occurs in the nucleus and mRNAs are exported to the cytoplasm for protein synthesis. Eventually SV40 virion assembly occurs (in the nucleus) and the cell is lysed to release the new virions.

When SV40 infects a host cell, one of two outcomes can occur, depending on the host cell. In *permissive* hosts, virus infection results in the usual formation of new virions and the lysis of the host cell. In *nonpermissive* hosts, lytic events do not occur; instead, the viral DNA becomes integrated into host DNA, genetically altering the cells in the process (Figure 10.15b). Such cells can show loss of growth inhibition and become malignant, a process called *transformation* (↔ Figure 8.20). As in certain tumor-causing retroviruses (Section 10.11), expression of specific SV40 genes is required to convert the cell to the transformed state. These tumor-inducing proteins bind to and inactivate host cell proteins that control cell division, and in this way, they promote uncontrolled cell development.

Herpesviruses

Herpesviruses are a large group of double-stranded DNA viruses that cause a variety of human diseases, including fever blisters (cold sores), venereal herpes, chicken pox, shingles, and infectious mononucleosis. An important group of herpesviruses cause cancer. For example, Epstein–Barr virus causes Burkitt’s lymphoma, a tumor endemic in children of central Africa and New Guinea. A widespread herpesvirus is cytomegalovirus (CMV), present in nearly three-quarters of all adults in the United States over 40 years of age. For healthy individuals, infection with CMV comes with no apparent symptoms or long-term health consequences. However, CMV can cause pneumonia, retinitis (an eye condition), and certain gastrointestinal disorders, as well as serious disease or even death in immune-compromised individuals.

Herpesviruses can remain latent in the body for long periods of time and become active under conditions of stress or when the immune system is compromised. Herpesvirus virions are enveloped and can have many distinct structural layers over the icosahedral nucleocapsid (Figure 10.16). Following viral attachment, the host cytoplasmic membrane fuses with the virus envelope, and this releases the nucleocapsid into the cell. The nucleocapsid is transported to the nucleus, where the viral DNA is uncoated and three classes of mRNA are produced: *immediate early*, *delayed early*, and *late* (Figure 10.16). Immediate early mRNA encodes certain regulatory proteins that stimulate the synthesis of the delayed early proteins. Among the key proteins synthesized during the delayed early stage is a viral-specific DNA polymerase and a DNA-binding protein, both of which are needed for viral DNA replication. As for other viruses, late proteins are primarily viral structural proteins.

Herpesvirus DNA replication takes place in the nucleus. After infection, the herpesvirus genome circularizes and replicates by a

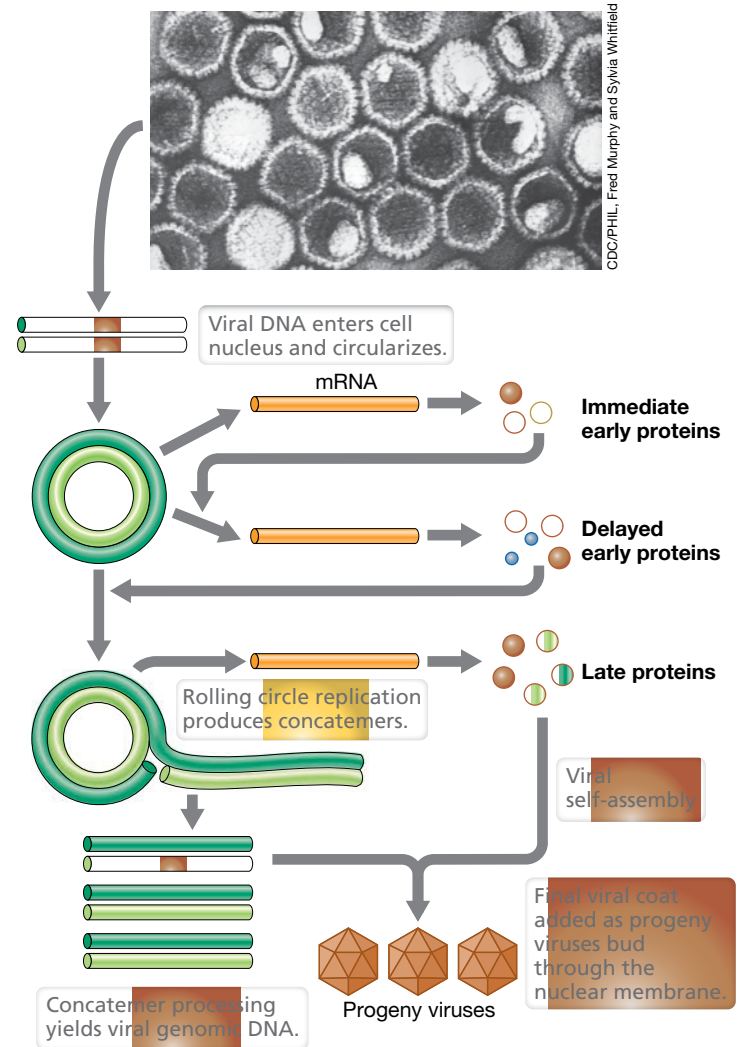


Figure 10.16 Herpesvirus. Flow of events in replication of herpes simplex virus starting from a transmission electron micrograph of herpes simplex virus (diameter of a single virion is about 150 nm). Although the viral genome is linear within the virion, it circularizes once inside the host.

rolling circle mechanism (Section 10.3). Long concatemers are formed that become processed into virus-length genomic DNA during the assembly process (Figure 10.16). Viral nucleocapsids are assembled in the nucleus, and the viral envelope is added during budding through the nuclear membrane. Mature herpesvirus virions are subsequently released through the endoplasmic reticulum to the outside of the cell. The assembly of herpesvirus virions thus differs from that of other enveloped viruses, which typically receive their envelope from the cytoplasmic membrane during exit from the cell.

MINIQUIZ

- What genomic feature does SV40 share with bacteriophage ϕ X174?
- How can the outcome of an SV40 viral infection differ in permissive versus nonpermissive hosts?
- Name two common diseases caused by herpesviruses.

III • Viruses with RNA Genomes

We have seen that RNA viruses infect a multitude of hosts and were likely the first viruses to appear on Earth (Section 10.2). As in the foregoing sections that dealt with DNA viruses, we organize our coverage of RNA viruses here by genomic characteristics. RNA viruses make up Baltimore classes III–VI (Figure 10.2).

10.8 Positive-Strand RNA Viruses

Many viruses contain single-stranded RNA genomes of the plus sense and are therefore *positive-strand RNA viruses*. In these viruses, the sequence of the genome and the mRNA are the same (Figure 10.2). A number of positive-strand animal and bacterial viruses are known, so we restrict our coverage here to just a few well-studied cases. We begin with the tiny bacteriophage MS2.

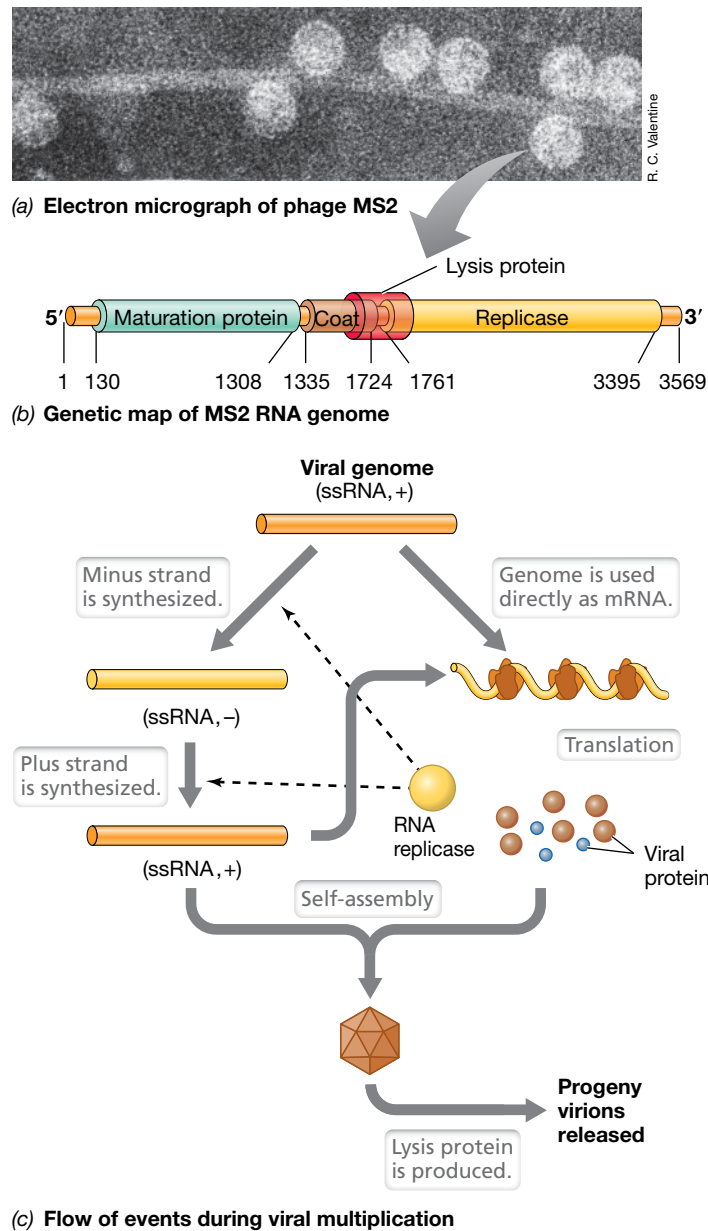
Bacteriophage MS2

Bacteriophage MS2 is about 25 nm in diameter and has an icosahedral capsid. The virus infects cells of *Escherichia coli* by attaching to the cell's pilus (Figure 10.17a), a structure that normally functions in a form of horizontal gene exchange (conjugation) in bacteria. How MS2 RNA actually gets inside the *E. coli* cell from the pilus is unknown, but once it has, MS2 replication events begin quickly; the genetic map and major activities of this virus are shown in Figure 10.17b and c.

The MS2 genome is just 3.5 kb in size and encodes only four proteins, including the maturation protein, coat protein, lysis protein, and one subunit of **RNA replicase**, the enzyme that replicates the viral RNA. Interestingly, MS2 RNA replicase is a composite protein, with three subunits encoded by the host genome and one subunit by the viral genome. The gene encoding the MS2 lysis protein overlaps that encoding the coat protein and replicase subunit (Figure 10.17b). We have seen this phenomenon of *overlapping genes* before (Section 10.3) as a strategy for making small genomes encode more proteins.

Because the genome of phage MS2 is plus-sense RNA, it is translated directly upon entry into the cell by the host RNA polymerase. When RNA replicase is made, it begins synthesis of minus-sense RNA using plus strands as templates. As minus-sense RNA copies accumulate, more plus-sense RNA is made using the minus-sense strands as templates, and some of these are translated for continued synthesis of viral structural proteins.

Phage MS2 regulates synthesis of its proteins by controlling access of host ribosomes to translational start sites on its RNA. MS2 genomic RNA is folded into a complex secondary structure. Of the four AUG translational start sites (see Section 4.9) on the MS2 RNA, the most accessible to the cell's translation machinery is that for the coat protein and replicase. Hence, translation begins at these sites very early following infection. However, as coat protein molecules accumulate, they bind to the RNA around the AUG start site for the replicase protein, effectively turning off synthesis of replicase. Although the gene for the maturation protein is at the 5' end of the RNA, the extensive folding of the RNA limits access to the maturation protein translational start site, and consequently, only a few copies are synthesized. In this way, all MS2 proteins are made in the relative amounts needed for virus assembly. Ulti-



(c) Flow of events during viral multiplication

Figure 10.17 A small RNA bacteriophage, MS2. (a) Transmission electron micrograph of the pilus of a cell of *Escherichia coli* showing virions of phage MS2 attached. (b) Genetic map of MS2. Note how the lysis protein gene overlaps with both the coat protein and replicase genes. The numbers refer to the nucleotide positions on the RNA, the entire genome being 3569 nucleotides in length. (c) Flow of events during MS2 replication.

mately, spontaneous assembly of MS2 virions begins, and the virions are released from the cell as a result of cell lysis.

Poliovirus

Several positive-strand RNA animal viruses cause disease in humans and other animals. These include poliovirus, the rhinoviruses that cause many cases of the common cold, the coronaviruses that cause respiratory syndromes, including severe acute respiratory syndrome (SARS), and the hepatitis A virus. We focus here on poliovirus and coronaviruses, both of which have linear RNA genomes.

Poliovirus is one of the smallest of all viruses with a 30-nm icosahedral structure containing the minimum 60 morphological units per virion (Figure 10.18a, b). At the 5' terminus of the viral RNA is a protein, called the VPg protein, that is attached covalently to the genomic RNA, and at the 3' terminus is a poly(A) tail (Figure 10.18c), a common feature of eukaryotic cell transcripts (Section 4.6). The poliovirus genome (about 7.4 kb) is also the mRNA, and the VPg protein facilitates binding of the RNA to host ribosomes. Translation yields a **polyprotein**, a single protein that self-cleaves into several smaller proteins including virion structural proteins. Other proteins generated from the polyprotein include the VPg protein, an RNA replicase responsible for synthesis of both minus-strand and plus-strand RNA, and a virus-encoded protease, which carries out the polyprotein cleavage (Figure 10.18c). This mechanism is called *post-translational cleavage* and is common in many animal viruses as well as animal cells.

Poliovirus replication occurs in the host cell cytoplasm. To initiate infection, the poliovirus virion attaches to a specific receptor on the surface of a sensitive cell and enters the cell. Once inside the cell, the virion is uncoated, and the genomic RNA is attached to ribosomes and translated to yield the polyprotein. Replication of viral RNA by the poliovirus RNA replicase begins shortly after infection. Both the positive and negative strands that are made pick up the VPg protein, which also functions as a primer for RNA synthesis (Figure 10.18c). Once poliovirus replication begins, host events are inhibited, and about 5 h after infection, cell lysis occurs with the release of new poliovirus virions.

Coronaviruses

Coronaviruses are single-stranded plus RNA viruses that, like poliovirus, replicate in the cytoplasm, but they differ from poliovirus in their larger size and details of replication. Coronaviruses cause respiratory infections in humans and other animals, including about 15% of common colds and SARS, an occasionally fatal infection of the lower respiratory tract in humans (Section 30.7).

Coronavirus virions are enveloped and contain club-shaped glycoprotein spikes on their surfaces (Figure 10.19a). These give the virus the appearance of having a “crown” (*corona* is Latin for crown). Coronavirus genomes are noteworthy because they are the largest of any known RNA viruses, about 30 kb. Because it is of the plus sense, the coronavirus genome can function directly in the cell as mRNA; however, most coronavirus proteins are not made by translating genomic RNA. Instead, only a portion of the genome is translated, in particular the region encoding RNA replicase (Figure 10.19b). The latter then uses the genomic RNA as a template to produce complementary negative strands from which several mRNAs are produced, and these mRNAs are translated to produce coronaviral proteins (Figure 10.19b). Full-length genomic RNA is also made off of the negative strands. New coronaviral virions are assembled within the Golgi complex, a major secretory organelle in eukaryotic cells (Section 2.16), and the fully assembled virions are released later from the cell surface.

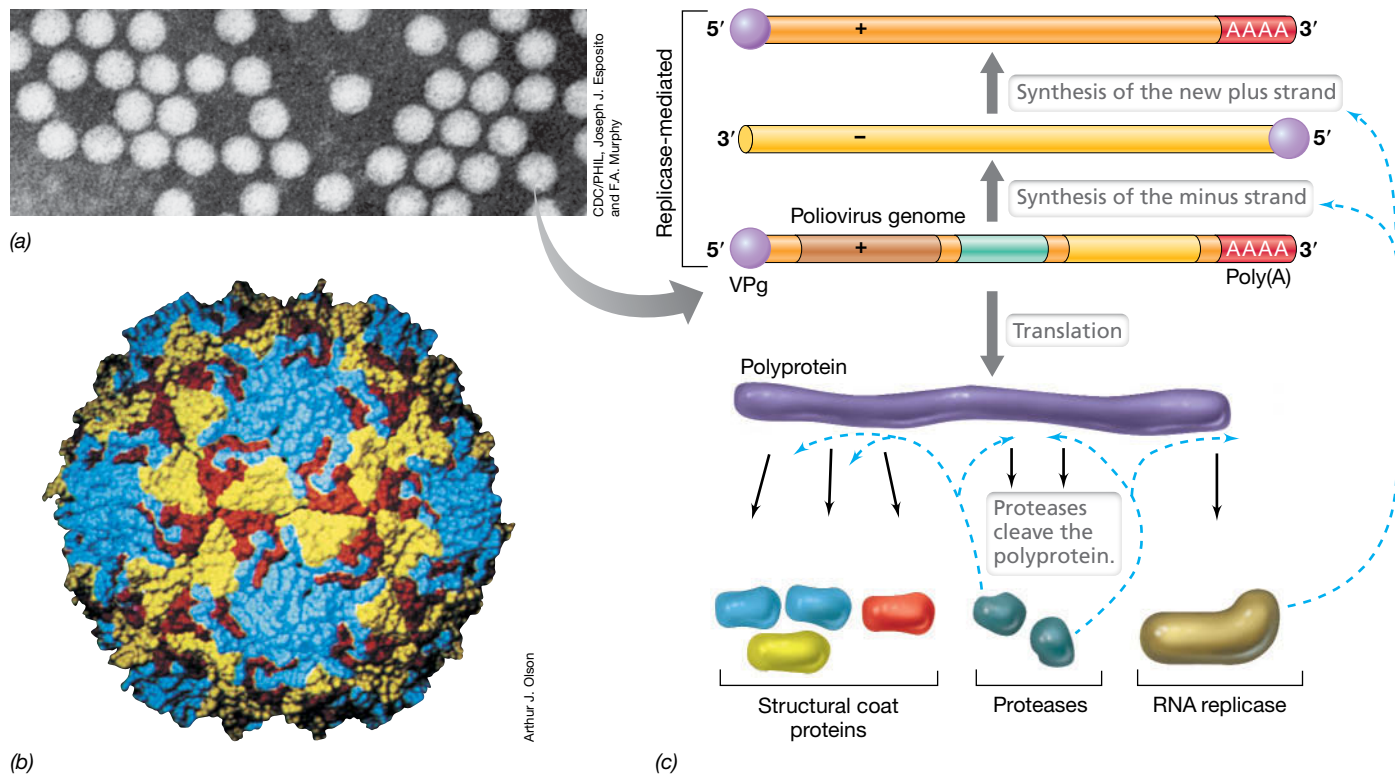


Figure 10.18 Poliovirus. (a) Transmission electron micrograph of poliovirus virions; a single virion is about 30 nm in diameter. (b) A computer model of a poliovirus virion. The various structural proteins are shown in distinct colors. (c) Genomic replication and formation of poliovirus proteins. Note the importance of the RNA replicase.

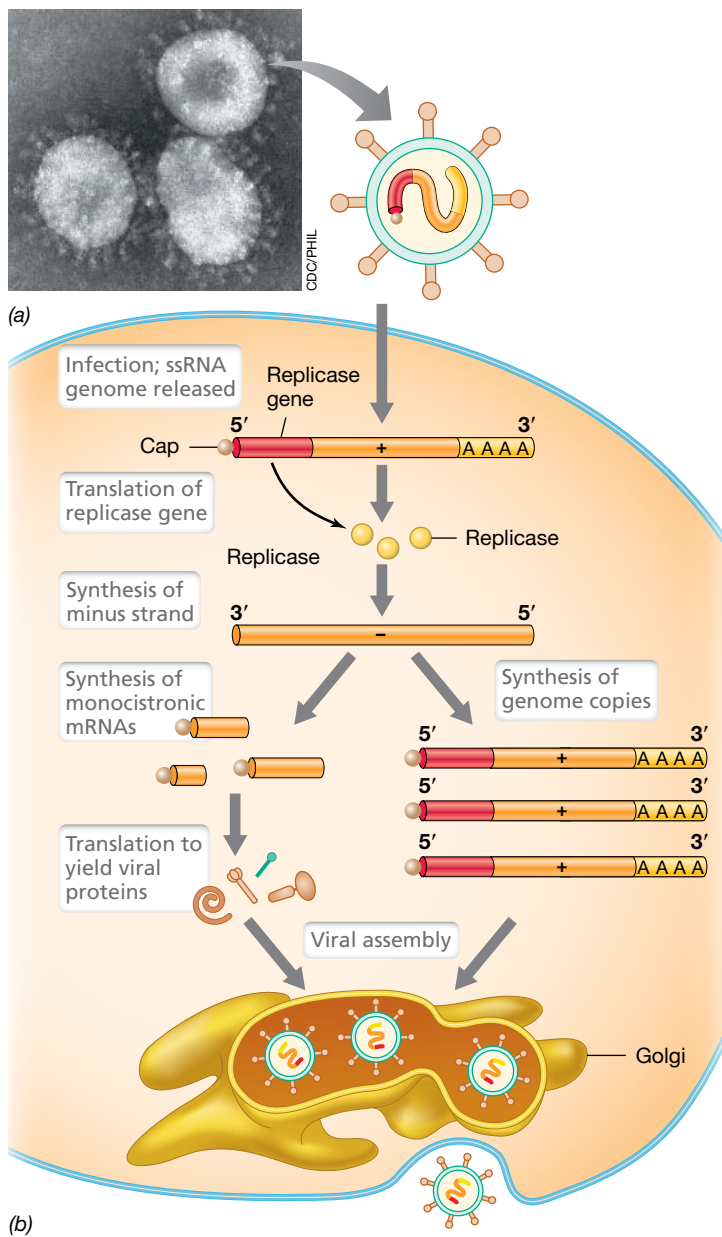


Figure 10.19 Coronavirus. (a) Electron micrograph of a coronavirus; a virion is about 150 nm in diameter. (b) Steps in coronavirus replication. The mRNA encoding viral proteins is transcribed from the negative strand made by the RNA replicase using the viral genome as a template.

Coronavirus differs from poliovirus in terms of virion and genome size, lack of the VPg protein, and absence of polyprotein formation and cleavage. Nevertheless, their single-stranded plus-sense RNA genomes dictate that many other molecular events must occur in a similar way.

MINIQUIZ

- How can poliovirus RNA be synthesized in the cytoplasm whereas host RNA must be made in the nucleus?
- What is present in the poliovirus polyprotein?
- How are protein synthesis and genomic replication similar or different in poliovirus and the SARS virus?

10.9 Negative-Strand RNA Animal Viruses

A number of animal viruses have minus-sense RNA genomes (Baltimore class V, Figures 10.2 and 10.3). In contrast to the plus-strand viruses just considered, the genomes of these negative-strand RNA viruses are *complementary* in base sequence to the mRNA that is formed. We discuss here two important examples of negative-strand RNA viruses: rabies virus and influenza virus. There are no known negative-strand RNA bacteriophages or archaeal viruses.

Rabies Virus

Rabies virus, which causes the fatal neuroinflammatory disease rabies (see Section 31.1), is a rhabdovirus, a name that refers to the characteristic shape of the virion (*rhabdos* is Greek for rod). Rhabdoviruses are commonly bullet-shaped (Figure 10.20a) and have an extensive and complex lipid envelope surrounding the helically

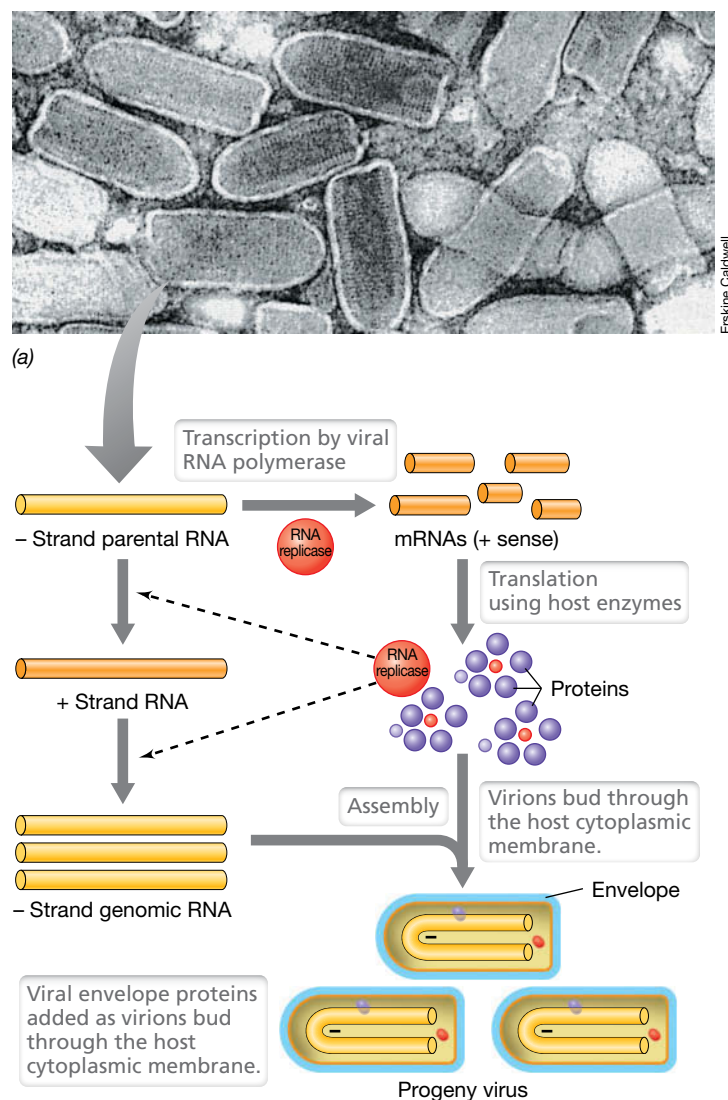


Figure 10.20 Negative-strand RNA viruses: Rhabdoviruses. (a) Transmission electron micrograph of vesicular stomatitis virus virions. A virion is about 65 nm in diameter. (b) Flow of events during replication of a negative-strand RNA virus. Note the importance of the viral-encoded RNA replicase.

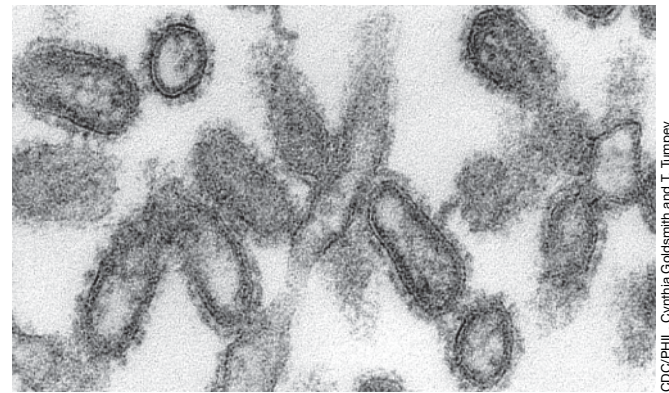
symmetrical nucleocapsid. A rhabdovirus virion contains several enzymes that are essential for the infection process, including an RNA replicase. Unlike positive-strand viruses, a rhabdovirus genome cannot be directly translated but must first be transcribed by the replicase. This occurs in the cytoplasm and generates two classes of RNAs. The first is a series of mRNAs encoding each of the viral proteins, and the second is a complementary copy of the entire viral genome; the latter functions as a template for the synthesis of genomic RNA copies (Figure 10.20*b*).

Assembly of a rhabdovirus virion is a highly orchestrated process. Two different coat proteins are made, nucleocapsid and envelope. The nucleocapsid is formed first by assembly of nucleocapsid protein molecules around the viral RNA genome. The envelope proteins are glycoproteins and they migrate to the cytoplasmic membrane where they are inserted into the membrane. Nucleocapsids then migrate to areas on the cytoplasmic membrane where these virus-specific glycoproteins are embedded and bud through them, becoming coated by the glycoprotein-enriched cytoplasmic membrane in the process. The final result is the release of new virions that can infect neighboring cells.

Influenza Virus

Another group of negative-strand RNA viruses contains the important human pathogen *influenza virus*. Influenza virus has been well studied over many years, beginning with early work during the 1918 influenza pandemic that killed millions of people worldwide (↻ Sections 29.8 and 30.8). Influenza virus is an enveloped virus in which the viral genome is present in the virion in a number of separate pieces, a condition called a *segmented genome*. In the case of influenza A virus, a common strain, the genome is segmented into eight linear single-stranded molecules ranging in size from 890 to 2341 nucleotides and totaling 13.5 kb. The nucleocapsid of the virus is of helical symmetry, about 6–9 nm in diameter and about 60 nm long, and is embedded in an envelope that has a number of virus-specific proteins as well as lipid derived from the host cytoplasmic membrane. Because of the way influenza virus buds as it leaves the cell, virions do not have a uniform shape and instead are pleomorphic (Figure 10.21*a*).

Several proteins on the outside of the influenza virion envelope interact with the host cell surface. One of these is *hemagglutinin*. Hemagglutinin is highly immunogenic (capable of stimulating the immune system) and antibodies against it prevent the virus from infecting a cell. This is the mechanism by which immunity to influenza is brought about by immunization (↻ Section 30.8). A second important influenza virus surface protein is the enzyme *neuraminidase* (Figure 10.21*b*). Neuraminidase breaks down sialic acid (a derivative of neuraminic acid) in the host cytoplasmic membrane. Neuraminidase functions primarily in virus assembly, destroying host membrane sialic acid that would otherwise block assembly or become incorporated into the virion. In addition to hemagglutinin and neuraminidase, influenza virions possess two other key enzymes. These include an RNA replicase, which converts the minus-strand genome into a plus strand, and an RNA endonuclease, which cuts the cap from host mRNAs (↻ Section 4.6) and



CCO/PHIL, Cynthia Godsmith and T. Tumpney

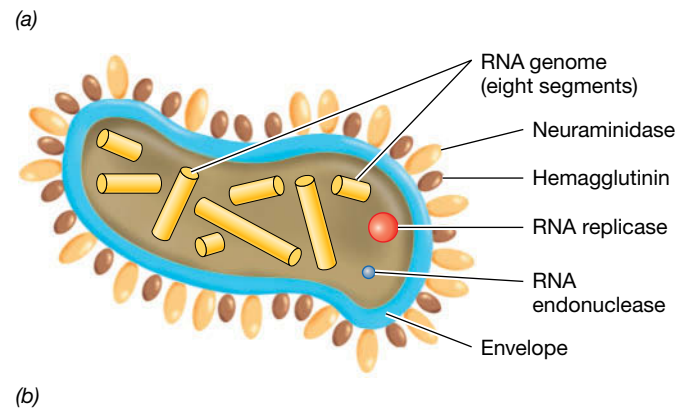


Figure 10.21 Influenza virus. (a) Transmission electron micrograph of thin sections of human influenza virus virions. (b) Some of the major components of the influenza virus, including the segmented genome.

uses them to cap viral mRNAs so they can be translated by the host translational machinery.

After the influenza virion enters the cell, the nucleocapsid separates from the envelope and migrates to the nucleus. Uncoating activates the virus RNA replicase and transcription begins. Ten proteins are encoded by the eight segments of the influenza virus genome; the mRNAs transcribed from six segments each encode a single protein, while the other two segments encode two proteins each. Some of the viral proteins are needed for influenza virus RNA replication, whereas others are structural proteins of the virion. The overall pattern of genomic RNA synthesis resembles that of the rhabdoviruses (Figure 10.20*b*), with full-length positive-strand RNA used as a template for making negative-strand genomic RNA. The complete enveloped virion forms by budding, as for the rhabdoviruses.

The segmented genome of the influenza virus has important practical consequences. Influenza virus exhibits a phenomenon called *antigenic shift* in which segments of the RNA genome from two different strains of the virus infecting the same cell are reassorted. This generates hybrid influenza virions that express unique surface proteins unrecognized by the immune system. Antigenic shift is thought to trigger major outbreaks of influenza because immunity to the new forms of the virus is essentially absent from the population. We discuss antigenic shift, and a related phenomenon called *antigenic drift*, in Section 30.8.

MINIQUIZ

- Why is it essential that negative-strand viruses carry an enzyme in their virions?
- What is a segmented genome?
- In influenza virus, what is antigenic shift and how does it occur?

10.10 Double-Stranded RNA Viruses

Viruses with double-stranded RNA genomes (Baltimore class III, Figure 10.2) infect animals, plants, fungi, and a few bacteria. *Reoviruses* are an important family of animal viruses, and we focus on them here.

Rotavirus is a typical reovirus and is the most common cause of diarrhea in infants 6 to 24 months of age. Other reoviruses cause respiratory infections and some infect plants. Reovirus virions consist of a nucleocapsid 60–80 nm in diameter, surrounded by a double shell of icosahedral symmetry (Figure 10.22a, b). As we have seen with single-stranded RNA viruses, the virions of double-stranded RNA viruses must carry their own enzyme to synthesize their mRNA and replicate their RNA genomes. Like the influenza virus genome, the reoviral genome is segmented, in this case into 10–12 molecules of linear double-stranded RNA totaling 18 kb.

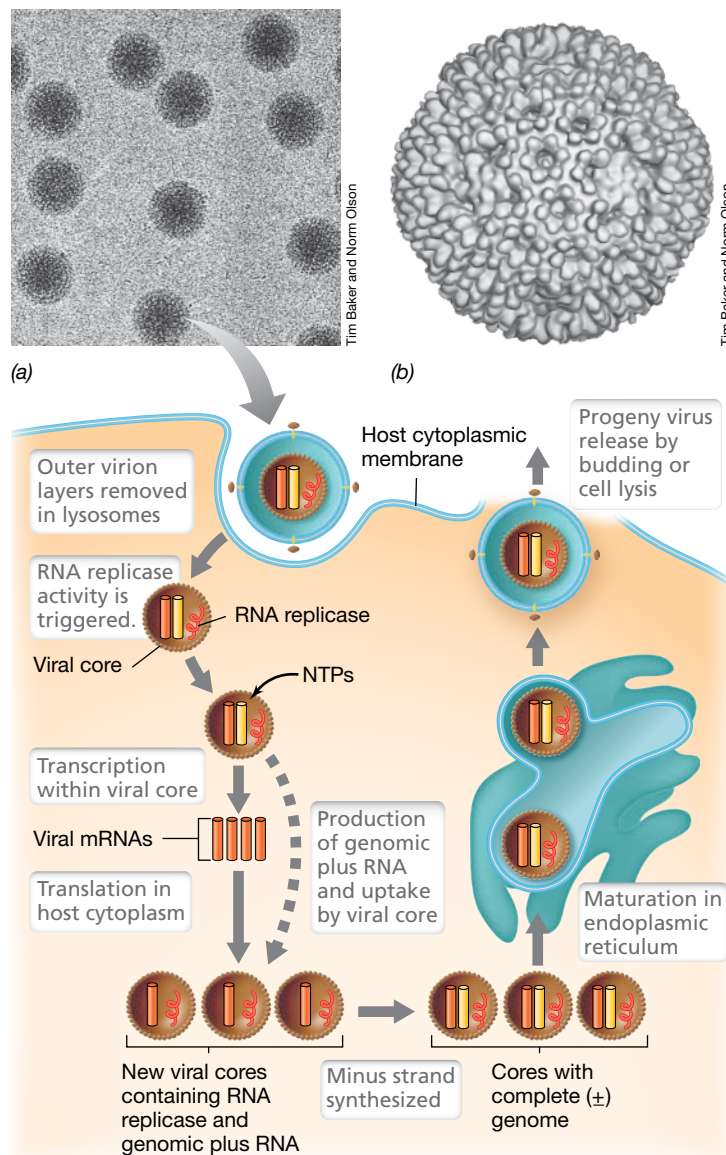
To initiate infection, a reovirus virion binds to a cellular receptor protein. The attached virus then enters the cell and is transported into lysosomes, where normally it would be destroyed (see Section 2.16). However, only the outer coats of the virion are removed by the lysosome, revealing the nucleocapsid; the latter is released into the cytoplasm. This uncoating process activates the viral RNA replicase and initiates virus replication (Figure 10.22c).

Reovirus Replication

Reovirus replication occurs exclusively in the host cytoplasm but *within* the nucleocapsid itself (Figure 10.22c) because the host has enzymes that recognize double-stranded RNA as foreign and would destroy it. The plus strand of the reoviral genome is inactive as mRNA, and thus the first step in replication is the synthesis of plus-sense mRNA by the viral-encoded RNA replicase, using minus-strand RNA as a template; the nucleotide triphosphates necessary for RNA synthesis are supplied by the host (Figure 10.22c). The mRNAs are then capped with a methylated nucleotide (as is typical of eukaryotic mRNAs, see Section 4.6) by viral enzymes and exported from the nucleocapsid into the cytoplasm and translated by host ribosomes.

Most RNAs in the reovirus genome encode a single protein, although in a few cases the protein formed is cleaved to yield the final products. However, one of the reovirus mRNAs encodes two proteins but the RNA does not have to be processed in order to translate both of these. Instead, a ribosome occasionally “misses” the start codon for the first gene in this mRNA and travels on to the start codon of the second gene to begin translation. When this occurs, the second protein, needed in small amounts, is made but the first protein is not. This “molecular mistake” can be viewed as a primitive form of translational control that ensures that viral proteins are made in their proper amounts.

As viral proteins are formed in the host cytoplasm, they aggregate to form new nucleocapsids, trapping copies of RNA replicase inside



(c)

Figure 10.22 Double-stranded RNA viruses: The reoviruses. (a) Transmission electron micrograph showing reovirus virions (diameter, about 70 nm). (b) Three-dimensional computer reconstruction of a reovirus virion calculated from electron micrographs of frozen-hydrated virions. (c) The reovirus life cycle. All replication and transcription steps occur inside the nucleocapsids. NTPs, nucleotide triphosphates.

as they form (Figure 10.22c). Newly formed nucleocapsids then take up the correct complement of genomic (plus-strand) RNA fragments—probably by recognition of specific sequences on each fragment—and as each single-stranded RNA enters a newly formed nucleocapsid, a double-stranded form is produced from it by RNA replicase. Once genomic synthesis is complete, viral coat proteins are added in the host’s endoplasmic reticulum, and the mature reoviral virions are released by budding or cell lysis (Figure 10.22c).

Reoviruses and RNA Replication

In addition to the unusual genome structure of reoviruses that forces them to employ special mechanisms to protect their double-stranded RNA genomes from cleavage by host ribonucleases

(Figure 10.22c), genomic replication in these viruses is also unique and differs in a fundamental way from that of cells and all other viruses.

Because the reovirus RNA genome is double-stranded, one might predict that reovirus replication would parallel that of organisms with double-stranded DNA genomes, but this is not the case. RNA replication in reoviruses is actually a *conservative* process rather than the well-known *semiconservative* process typical of cellular DNA replication and replication in viruses that contain double-stranded DNA genomes (↻ Section 4.3 and Figure 10.2). This is because synthesis of reovirus mRNA occurs *only* off of the minus strand as a template in the infecting nucleocapsids, whereas synthesis of double-stranded genomic RNA from assimilated plus-strand RNA copies in progeny virions occurs *only* off of the plus strand as a template (Figure 10.22c). Hence, in addition to their unique double-stranded RNA genomes, reoviruses also display their unusual molecular biology by employing a unique genomic replication mechanism that is neither semiconservative nor rolling circle (Figure 10.7) in nature.

MINIQUIZ

- What does the reovirus genome consist of?
- How does reovirus genome replication resemble that of influenza virus, and how does it differ?
- Why must reoviral replication events occur within the nucleocapsid?

10.11 Viruses That Use Reverse Transcriptase

Two different classes of viruses use *reverse transcriptase*, and they differ in the type of nucleic acid in their genomes; retroviruses have RNA genomes while hepadnaviruses have DNA genomes (Baltimore classes VI and VII, respectively, Figure 10.2). Besides their unique molecular properties, both classes of viruses include important human pathogens, including HIV (a retrovirus) and hepatitis B (a hepadnavirus).

Retroviruses: Integration of Viral Genes into the Host Genome

Retroviruses have enveloped virions that contain two identical copies of the RNA genome (↻ Figure 8.21a). The virion also contains several enzymes, including reverse transcriptase, and also a specific viral tRNA. Enzymes for retrovirus replication must be carried in the virion because although the retroviral genome is of the plus sense, it is not used directly as mRNA. Instead, the genome is converted to DNA by reverse transcriptase and integrated into the host genome. The DNA formed is a linear double-stranded molecule and is synthesized within the virion and then released to the cytoplasm. The major steps in reverse transcription are presented in **Figure 10.23**.

Reverse transcriptase has three enzymatic activities: (1) *reverse transcription* (to synthesize DNA from an RNA template), (2) *ribonuclease activity* (to degrade the RNA strand of an RNA:DNA hybrid), and (3) *DNA polymerase* (to make double-stranded

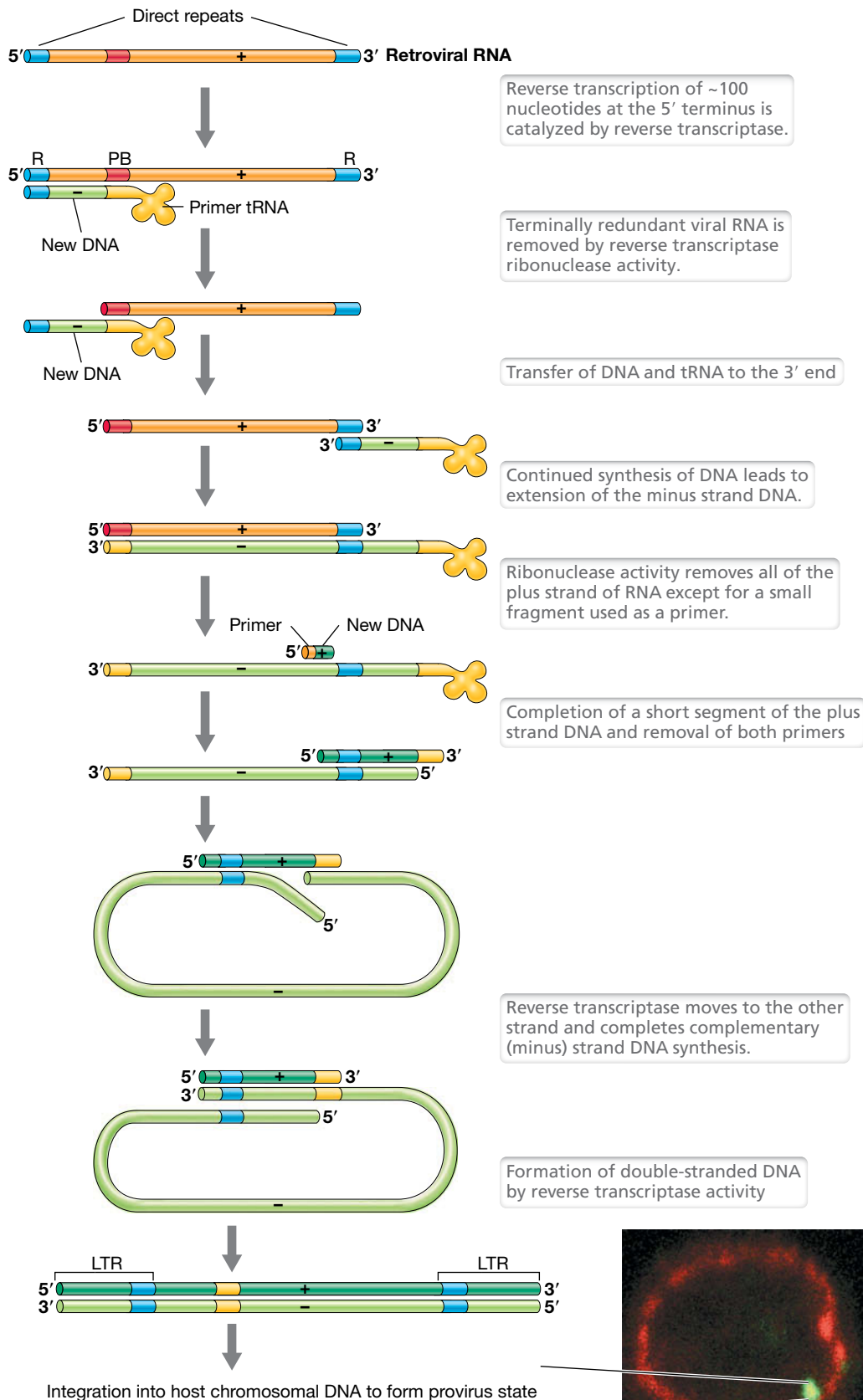
DNA from single-stranded DNA). Reverse transcriptase needs a primer for DNA synthesis and this is the function of the viral tRNA. Using this primer, nucleotides near the 5' terminus of the RNA are reverse-transcribed into DNA. Once reverse transcription reaches the 5' end of the RNA, the process stops. To copy the remaining RNA, a different mechanism comes into play. First, terminally redundant RNA sequences at the 5' end of the molecule are removed by reverse transcriptase. This leads to the formation of a small, single-stranded DNA that is complementary to the RNA segment at the *other end* of the viral RNA. This short, single-stranded piece of DNA then hybridizes with the other end of the viral RNA molecule, where synthesis of DNA begins once again.

Continued reverse transcription leads to the formation of a double-stranded DNA molecule with long terminal repeats, and these assist in integration of the retroviral DNA into the host chromosome (Figure 10.23). For HIV, the chromosomal integration site is not random. Through the use of a special form of fluorescence microscopy, scientists have shown that HIV incorporates into chromosomal loci near the outer shell of the nucleus (photo inset, Figure 10.23). This location is likely favored due to the short life of the viral integrase. Recall from Section 8.8 that reverse transcription of a retrovirus occurs in the nucleocapsid. Thus the retroviral DNA must be integrated quickly into the host genome upon entry into the nucleus.

Retroviruses: Induction to Form New Retrovirus Virions

Once integrated, retroviral DNA becomes a permanent part of the host chromosome; the genes may be expressed or they may remain in a latent state indefinitely. However, if induced, retroviral DNA is transcribed by a cellular RNA polymerase to form RNA transcripts that can be either packaged into virions as genomic RNA or translated to yield retroviral proteins. Translation and processing of retroviral mRNAs is shown in **Figure 10.24**. All retroviruses have the genes *gag*, *pol*, and *env*, arranged in that order in their genomes (↻ Figure 8.21). The *gag* gene at the 5' end of the mRNA actually encodes several viral structural proteins. These are first synthesized as a single protein (polyprotein) that is subsequently processed by a protease which itself is part of the polyprotein. The structural proteins make up the capsid, and the protease is packaged in the virion.

Next, the *pol* gene is translated into a large polyprotein that also contains the *gag* proteins (Figure 10.24a). Compared to *gag* proteins, *pol* proteins are required in only small amounts. This regulation is achieved because *pol* protein synthesis requires the ribosome to either read through a stop codon at the end of the *gag* gene or switch to a different reading frame in this region. Both of these are rare events and can be considered a form of translational regulation. Once produced, the *pol* gene product is processed to yield *gag* proteins, reverse transcriptase, and integrase; the latter is the protein required for viral DNA integration into the host chromosome. For the *env* gene to be translated, the full-length mRNA is first processed to remove the *gag* and *pol* regions, and then the *env* product is made and immediately processed into two distinct envelope proteins by the viral-encoded protease (Figure 10.24b). Retroviral assembly occurs on the inner side of the host cytoplasmic membrane and virions are released across the membrane by budding (↻ Figure 8.22).



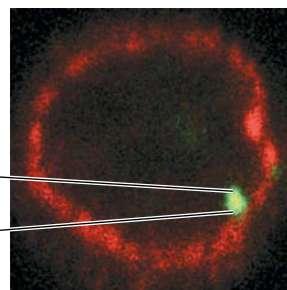
With retroviruses we thus see a replication scheme that is perhaps the most complex of all known viruses. Despite the complexity of retroviruses, molecular studies of them suggest an ancient origin and possible central importance in the transition from self-replicating RNA life forms to the DNA world of cellular organisms. It is the signature enzyme of retroviruses—reverse transcriptase, the only enzyme known that can make DNA from RNA—that brings retroviruses into this evolutionary limelight, and so it is possible that all cells owe their very existence to this class of viruses (Section 10.2 and Figure 10.4).

Hepadnaviruses

In addition to the retroviruses a second class of unusual viruses employ the enzyme reverse transcriptase; these are the **hepadnaviruses**, such as human hepatitis B virus (Figure 10.25a). The tiny DNA genomes of hepadnaviruses are unusual because they are neither single-stranded nor double-stranded but *partially* double-stranded. Despite their small size (3–4 kilobase pairs), the hepadnavirus genomes encode several proteins by employing overlapping genes, a strategy we have seen before in very small viruses (Sections 10.3 and 10.8).

Besides the usual activities of reverse transcriptase we have just considered (Figure 10.23), hepadnaviral reverse transcriptase also functions as a protein primer for synthesis of one of its own DNA strands. In terms of its role in replication events, however, reverse transcriptase plays different roles in

Figure 10.23 Formation of double-stranded DNA from retrovirus single-stranded genomic RNA. The sequences labeled R on the RNA are direct repeats found at either end. The sequence labeled PB is where the primer (tRNA) binds. Note that DNA synthesis yields longer direct repeats on the DNA than were originally on the RNA. These are called long terminal repeats (LTRs). Inset: A special form of fluorescence microscopy allows visualization of HIV genome integration (green) into a chromosome region near the nuclear membrane (red) of a CD4 lymphocyte. Image courtesy of Marina Lusic, University Hospital, Heidelberg, Germany.



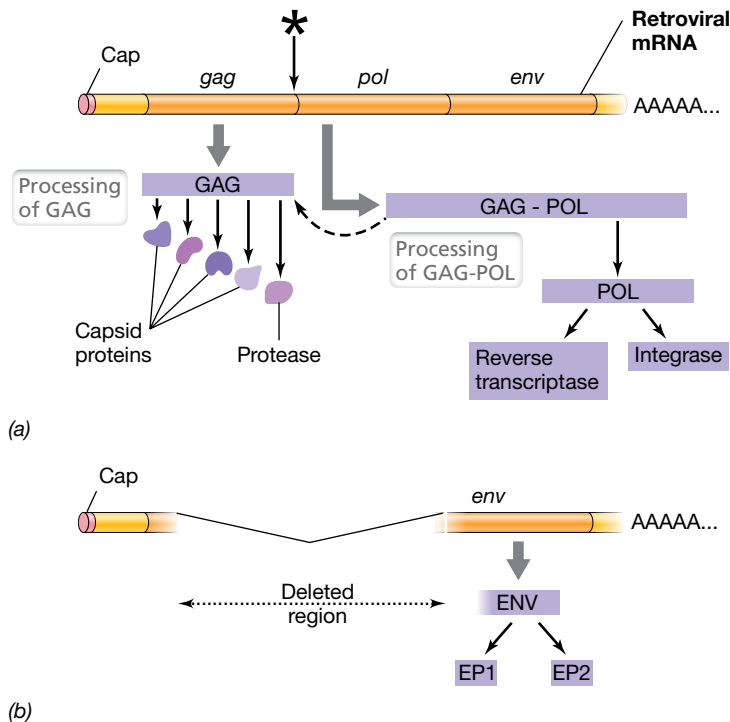
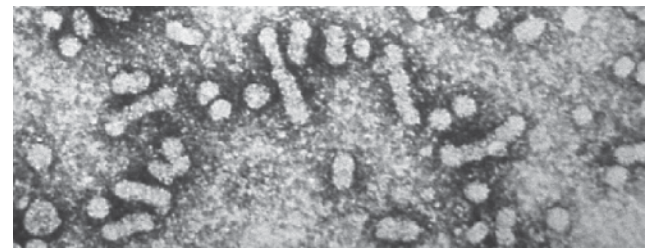


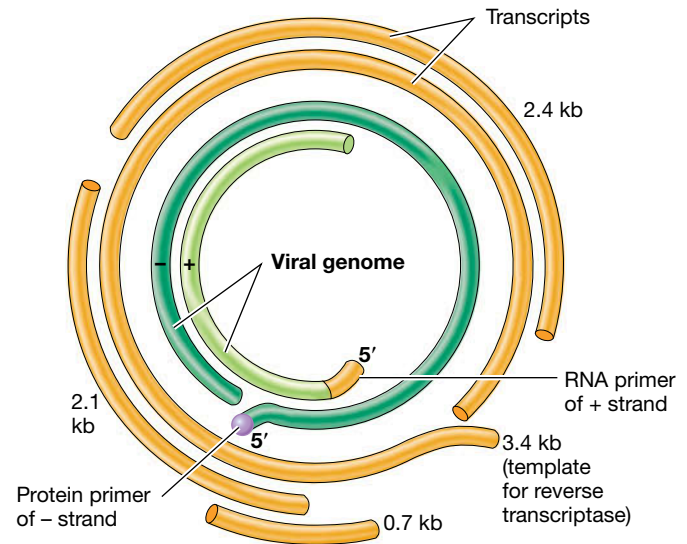
Figure 10.24 Translation of retrovirus mRNA and processing of the proteins. (a) Full-length retroviral mRNA encodes *gag*, *pol*, and *env*. The asterisk shows the site where a ribosome must read through a stop codon or do a precise shift of reading frame to synthesize the GAG-POL polyprotein. The thick gray arrows indicate translation, and the black arrows indicate protein-processing events. One of the *gag* gene products is a protease. The POL product is processed by this protease to yield reverse transcriptase and integrase, two key enzymes that catalyze retrovirus replication events (Figure 10.23). (b) The mRNA has been processed to remove most of the *gag-pol* region. This shortened message is translated to give the ENV polyprotein, which is cleaved into two envelope proteins (EP), EP1 and EP2.

retroviral and hepadnaviral genome replication. In hepadnaviruses, the DNA genome is replicated through an RNA intermediate, whereas in retroviruses, the RNA genome is replicated through a DNA intermediate (Figures 10.23 and 10.25).

Upon infection, the hepadnavirus nucleocapsid enters the host nucleus where the partial genomic DNA strand is completed by the viral polymerase to form an entire double-stranded molecule. Transcription by host RNA polymerase yields four size classes of viral mRNAs (Figure 10.25b), which are subsequently translated to yield the hepadnaviral proteins. The largest of these transcripts is slightly larger than the viral genome and, together with reverse transcriptase, associates with viral proteins in the host cytoplasm to form genomes for new virions. Reverse transcriptase forms single-stranded DNA off of this large transcript inside the virion to form the minus-sense strand of the DNA genome and then uses this as a template to form a portion of the plus-sense strand, yielding the partially double-stranded genome characteristic of hepadnaviruses (Figure 10.25b). Once mature virions are produced, these associate with membranes in the endoplasmic reticulum and Golgi complex, from which they are exported across the cytoplasmic membrane by budding.



(a)



(b)

Figure 10.25 Hepadnaviruses. (a) Electron micrograph of hepatitis B virions. (b) Hepatitis B genome. The partially double-stranded genome is shown in the standard green colors. The sizes of the transcripts (orange) are also shown; all of the genes in the hepatitis B virus overlap. Reverse transcriptase produces the DNA genome from a single genome-length mRNA made by host RNA polymerase.

MINIQUIZ

- Why are protease inhibitors an effective treatment for human AIDS?
- Contrast the genomes of HIV and hepatitis B virus.
- How does the role of reverse transcriptase in the replication cycles of retroviruses and hepadnaviruses differ?

IV • Viral Ecology

Viruses can be found everywhere on and in Earth where cellular life is present (including on and in plants and animals) and are present in some environments in enormous numbers. The number of bacterial and archaeal cells on Earth is far greater than the total number of eukaryotic cells; estimates of total prokaryotic cell numbers are on the order of 10^{30} . However, the number of viruses is even greater than this, an estimated 10^{31} (10 nonillion)! Thus, one might expect that, despite their small size, viruses would play a major ecological role in nature. We consider some aspects of viral ecology here including a mechanism that protects *Bacteria* and *Archaea* from viral destruction (and countermeasures

that have evolved in viruses) and explore the viral world that exists in and on the human body, the *human virome*.

10.12 The Bacterial and Archaeal Virosphere

The best estimates of the total number of cells of *Bacteria* and *Archaea* and their respective viruses have come from quantitative studies of seawater, although surveys of soil, freshwater, Earth's deep subsurface, and microbial mats, among many other habitats, are also teeming with microbes and the viruses that feed on them. We focus on seawater to give a feel for the numbers involved and how viruses interact with their prokaryotic hosts.

Bacteriophages and Archaeal Viruses in Seawater

There are about 10^6 prokaryotic cells/ml of seawater and approximately ten times as many viruses. It has been estimated that at least 5% and as many as 50% of the *Bacteria* in seawater are killed by bacteriophages each day, and most of the others are eaten by protozoa. For example, Syn5 is a bacteriophage that attacks and lyses *Synechococcus* species, who, along with their relative *Prochlorococcus*, are the major primary producers of the ocean and account for over 30% of the CO_2 fixed globally (↻ Section 20.10) (Figure 10.26). The cytoplasm released as a result of viral attack on cells of these species (Figure 10.26c) provides a significant amount of organic matter for other microbes in the ocean. Although viruses account for most of the total microbes present in seawater in terms of numbers, because of their very small sizes they constitute only about 5% of the total microbial biomass (Figure 10.27).

The most common bacteriophages in the oceans are head-and-tail bacteriophages containing double-stranded DNA genomes (Baltimore class I, Figure 10.2); by contrast, RNA-containing phages are relatively rare. As we have seen, lysogenic bacteriophages can integrate into the genomes of their bacterial hosts (↻ Section 8.7), and when they do, they can confer new properties on the cell. Moreover, some lytic phages facilitate the transfer of bacterial genes from one cell to another through the process of transduction, a major means of horizontal gene transfer in

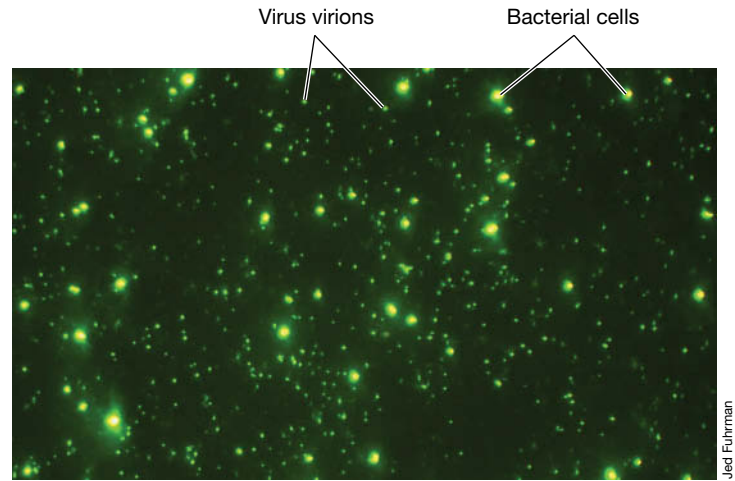


Figure 10.27 Viruses and bacteria in seawater. A fluorescence photomicrograph of seawater stained with the dye SYBR Green to reveal prokaryotic cells and viruses. Although viruses are too small to be seen with the light microscope, fluorescence from a stained virus is visible.

which a virus ferries host genes between cells, for example by picking up host DNA and becoming a nonlytic *transducing particle* (↻ Section 11.7). As agents of transduction, bacteriophages are thought to have a major influence on bacterial evolution. For example, transferred genes may confer new metabolic or other beneficial properties on the recipient cells and allow them to successfully colonize new habitats.

A good example of phage gene transfer are the cyanophages that have been shown to transfer certain photosynthesis genes among strains of *Synechococcus* and *Prochlorococcus*. When these phages are released from their lysed host cells (Figure 10.26c), some of them incorporate host genes that encode photosystem (PS) II, one of the key components of oxygenic photosynthesis (↻ Section 14.4). When such a phage infects a new host cell, it provides the cell with genes encoding a modified PSII. It is hypothesized that a more diverse complement of PSII proteins improves both cell and phage fitness by allowing the host cell to better adapt to changing environmental

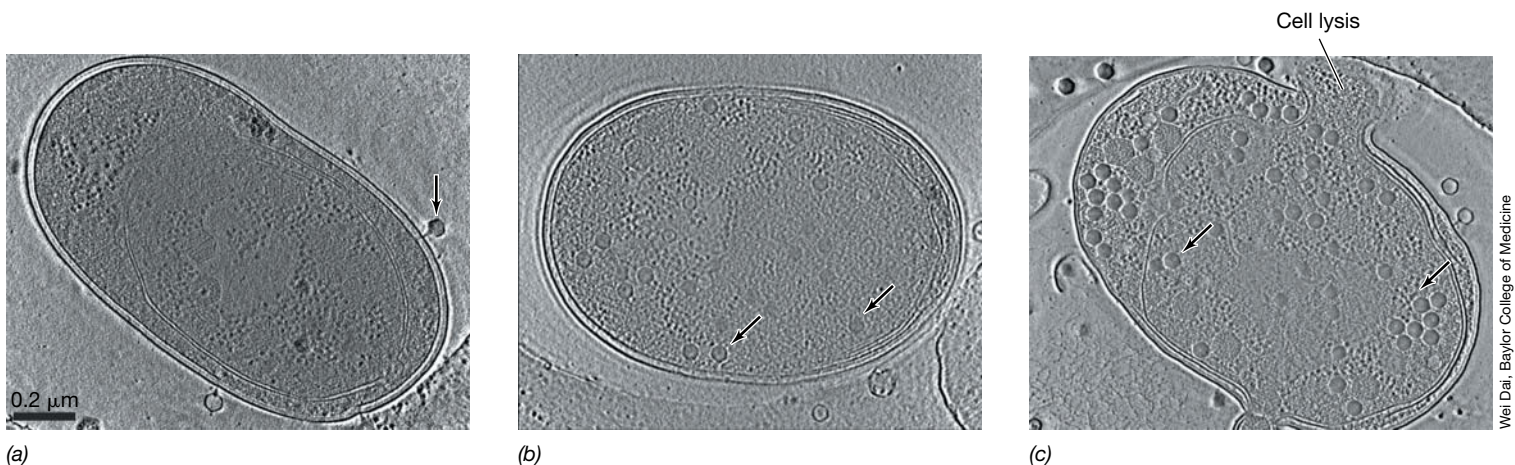


Figure 10.26 Cyanophage Syn5 infection of *Synechococcus*. Phase contrast–electron cryotomography sections of cells in various stages of infection. (a) Early. (b) Intermediate. (c) Late. Arrows point to phage virions.

conditions—for example, changes in light intensity or light quality—and in the process, producing more cyanophage.

Many *Archaea* are present in the oceans, and a major group of marine *Archaea* of ecological importance is the *Thaumarchaeota* (↻ Section 17.5). These ammonia-oxidizing chemolithotrophs are capable of consuming the vanishingly low levels of ammonia present in planktonic (open ocean) waters. Although lytic archaeal viruses have yet to be isolated for this group, several species of the thaumarchaeotan genus *Nitrosopumilus* have been shown to harbor viral genomes within their chromosome (that is, the cells contain a provirus) whose genes suggest that the infecting virus was either of the head-and-tailed dsDNA bacteriophage type (Section 10.4) or was icosahedral-shaped, similar to the herpesviruses (Section 10.7) (Figure 10.3b). It is thus likely that at least some, and perhaps even many, of the viruses in seawater (Figure 10.27) infect marine *Archaea* instead of marine *Bacteria*. This is bolstered by the observation that virtually all known archaeal viruses contain double-stranded DNA genomes (Figure 10.3a), and these are the most commonly observed viral genomes in the oceans.

Survival Strategies and Metagenomics of Viruses in Nature

When hosts are plentiful in nature, it is thought that bacteriophages adopt the lytic lifestyle and thus large numbers of host cells are killed. By contrast, when host numbers are low, it may be difficult for viruses to find a new host cell, and under such circumstances, lysogeny would be favored if the virus is lysogenic (↻ Section 8.7). Under these conditions, the virus would survive as a prophage until host numbers rebounded and a lytic lifestyle could once again be supported. This hypothesis is consistent with the observation that in the depths of the ocean where bacterial numbers are lower than in surface waters, around half the bacteria examined have been found to contain one or more lysogenic viruses. As far as is known, no single-stranded DNA viruses and no RNA viruses can enter a lysogenic state, and so how these viruses might survive periods of low host numbers is unknown.

Most of the genetic diversity on Earth resides in viruses, mostly bacteriophages. The *viral metagenome* is the sum total of all the virus genes in a particular environment. Several viral metagenomic studies have been undertaken, and they invariably show that immense viral diversity exists on Earth. For example, approximately 75% of the gene sequences found in viral metagenomic studies show no similarity to any other genes cataloged in viral or cellular gene databases. By comparison, surveys of bacterial metagenomes typically reveal approximately 10% unknown genes. Thus, most viruses still await discovery and most viral genes have unknown functions. This makes the study of the virosphere and viral diversity one of the most exciting areas of microbiology today.

MINIQUIZ

- What type of bacteriophages are most common in the oceans?
- How can bacteriophages affect bacterial evolution?
- What does the viral metagenome suggest about our understanding of viral diversity?

10.13 Viral Defense Mechanisms of Bacteria and Archaea

The viruses of *Bacteria* and *Archaea* can be viewed in two different ways: as deadly predators that cull cell populations, or as agents of diversity, enriching their hosts through gene transfers. Despite the importance of the latter, with viruses outnumbering their prokaryotic hosts by a factor of about 10, *Bacteria* and *Archaea* have evolved an arsenal of strategies to defend against their viral predators and we consider these here.

The Microbial Arms Race

The relationship between bacteriophages and their hosts is not static but instead is extremely dynamic. Although bacteria possess several weapons to battle phage attack, phages counter these weapons with weapons of their own, triggering a microbiological “arms race” for cell survival and viral propagation.

We previously discussed how *Bacteria* produce restriction endonucleases—proteins that target and destroy foreign DNA—and how in the case of bacteriophage T4 the virus avoids restriction enzyme surveillance by its *Escherichia coli* host by substituting the base 5-hydroxymethylcytosine for cytosine in its genome (↻ Section 8.5). In response, some *E. coli* strains have evolved altered restriction enzyme systems that recognize this viral DNA modification and still degrade incoming T4 DNA, preventing infection. This host adaptation has then selected for T4 bacteriophages that modify their DNA in a different manner—by glycosylating (adding sugars) to specific DNA bases. Then, as one might expect in an expanding arms race, some strains of *E. coli* have evolved restriction systems that recognize glycosylated viral DNA and destroy it. As a countermeasure to this, some T4 strains have evolved a protein that inhibits these modified restriction enzymes. In a constant attempt to stay ahead, other *E. coli* strains have evolved endonucleases that are not inhibited by these bacteriophage proteins, and so it goes, back and forth as both predator and prey fight to survive and propagate in the face of “the enemy.”

Alterations in viral receptor sites are also a common mechanism to avoid viral infection; for a virus to attach to a host cell, the virus must first recognize and attach to a cell receptor (↻ Section 8.5 and Figure 8.11). To prevent viral infection, hosts can modify the structure of the cell receptor or protect the receptor by producing a shield, such as an outer cell surface capsule (↻ Section 2.7). However, viruses can counter these shielding mechanisms either by a mutated viral receptor able to bind to its modified complement on the cell surface or by carrying enzymes within the capsid that can degrade the capsule. Some temperate bacteriophages (↻ Section 8.7) ensure their self-preservation by hijacking the host-encoded toxin–antitoxin system (↻ Section 7.11). They do this by way of a prophage-encoded protein that inactivates the host’s antitoxin protein and replaces it with a phage antitoxin protein. Thus for the cell to avoid toxicity from its chromosomally encoded toxin (a protein that slows growth during stressful conditions in order to preserve resources but which could make cells uncompetitive in times of plenty if not controlled by its antitoxin), the cell is forced to retain the prophage and gain protection from its encoded antitoxin.

The Antiviral System of *Bacteria* and *Archaea*: CRISPR

A major antiviral defense of both *Bacteria* and *Archaea* is CRISPR, the clustered regularly interspaced short palindromic repeats found in the chromosomes of many species that help protect them from bacteriophage infection (see Section 11.12). CRISPR regions contain short repeats of *constant* DNA sequence alternating with short *variable* DNA sequences called *spacers* (Figure 10.28 and see Figure 11.33). These spacers correspond to pieces of viral or other foreign DNA and function as a “memory bank” of past encounters with a virus in a manner analogous to how animals produce antibodies and long-lasting memory cells against an infecting virus (adaptive immunity, Chapter 27).

Besides the spacer regions, another essential component of the CRISPR system are the *Cas* (CRISPR-associated) proteins. These proteins possess endonuclease activity and both mediate the defense against foreign DNA and incorporate new spacer regions into the CRISPR region. When a virus attaches to a host cell and injects its DNA, the *Cas* proteins of a CRISPR region may recognize specific DNA sequences known as *protospacer adjacent motifs* (PAMs) (Figure 10.28a). The *Cas* protein then cleaves the viral DNA at a locus near this PAM (termed the *protospacer*) and inserts the short DNA region into the CRISPR region of the chromosome, where it becomes a *spacer* (Figure 10.28a). The insertion of a spacer into the CRISPR region confers “genetic memory”

(referred to as *immunization*) on the cell and sets the stage for later encounters with the same virus.

Immune Memory and Other Aspects of CRISPR

When an immunized cell encounters the same virus at a later date, the *Cas* proteins quickly destroy the incoming DNA in an RNA-dependent process. While the genomic CRISPR region does not contain open reading frames, it does have a promoter (see Section 4.5). The resulting transcript is considered the *pre-CRISPR RNA* (pre-crRNA) and contains an array of RNA sequences complementary to both the repeat and spacer regions. The *Cas* proteins then process the transcript into individual spacer RNAs by targeting the repeat regions (Figure 10.28b). These crRNAs then associate with *Cas* proteins within the cleavage complex and begin surveillance for complementary incoming viral DNAs. Any viral DNA:crRNA duplexes formed are cleaved by the endonuclease activity of *Cas* proteins and the invading DNA is degraded in a process called *interference* (Figure 10.28b). With part of its genome destroyed in this way, an invading virus cannot proceed to replicate and the infection (and threat to the cell) is thwarted.

One of the major conundrums of the CRISPR system is how a host can survive the *initial* viral invasion long enough to become immunized. For the CRISPR system to successfully prevent viral infection, a spacer corresponding to a region of the viral genome

must already be present in the CRISPR locus. Immunization probably occurs when an incoming virus has been inactivated by environmental factors (such as ultraviolet radiation) or when the host’s restriction enzyme system cleaves the invading DNA before a successful infection can be initiated.

Because of the dynamic nature of the phage-host interaction, viruses have evolved mechanisms to avoid CRISPR surveillance and destruction. These include mutation of the PAM regions recognized by the memorizing complex of *Cas* proteins and the production of proteins that inhibit activity of the cleavage complex of *Cas* proteins (Figure 10.28b). In addition, a *phage-encoded* (in contrast to *cell-encoded*) CRISPR system has been identified in the genome of a bacteriophage that infects *Vibrio cholerae*, the bacterium that causes cholera. The phage-encoded crRNAs target genes in *V. cholerae* that encode a defense system that prevents bacteriophage propagation; when these genes are inactivated, the cell’s defense system is defunct, a rather elaborate example of the “arms race” between bacteria and their viruses.

While CRISPRs are essential for viral surveillance in *Bacteria* and *Archaea*, we discuss in later chapters how their mode of action is also beneficial for maintaining cellular genome integrity (see Section 11.12), and how the CRISPR/*Cas* system has been developed as a powerful tool for synthetic biology (see Section 12.12).

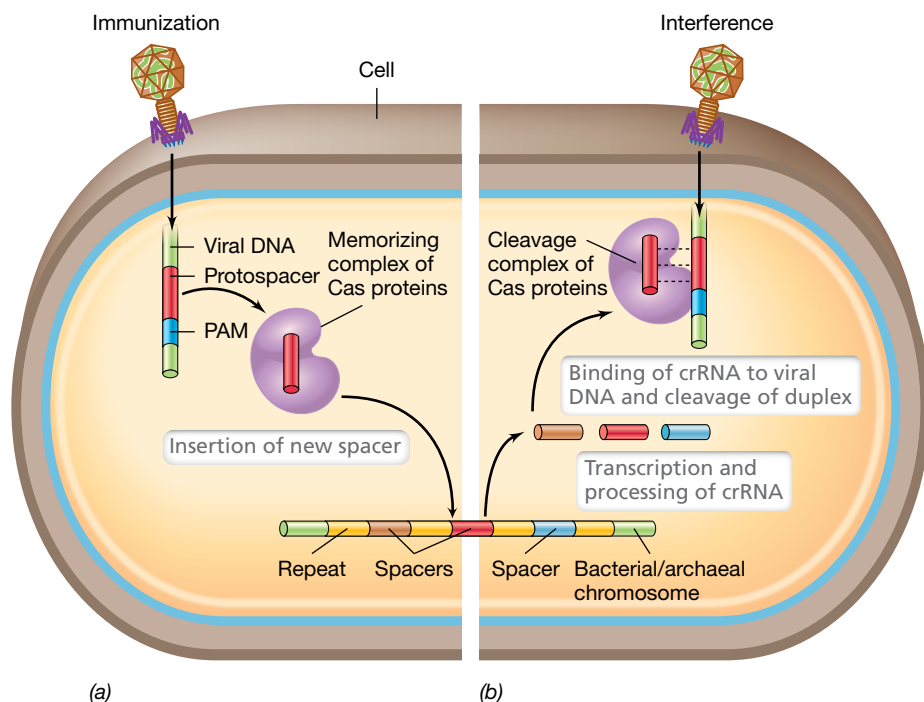


Figure 10.28 CRISPR defense against viruses. (a) Immunization. Incoming viral DNA is targeted by the memorizing complex of *Cas* proteins. This complex selects a protospacer region based on protospacer adjacent motif (PAM) sequences located on the virus genome. Once the protospacer is excised from the viral genome, the memorizing complex inserts the protospacer into the CRISPR region of the chromosome, resulting in a unique spacer region. (b) Interference. The chromosomal CRISPR region is transcribed and processed into crRNAs that correspond to the individual spacer regions. *Cas* proteins bind to these crRNAs and search for complementary DNA. If a crRNA binds to the DNA of an invading virus, forming a crRNA:DNA duplex, the endonuclease activity of the cleavage complex is triggered and results in the degradation of incoming viral DNA.

MINIQUIZ

- Describe two ways that prokaryotic cells can avoid viral infection and how viruses may overcome these defenses.
- How does a prokaryotic cell become immunized against a specific virus?
- How do the crRNAs and Cas proteins protect the cell from invading viral DNA?

10.14 The Human Virome

In Chapter 8 we highlighted the infection process of animal viruses, and in Figures 10.2 and 10.3 we described the general morphology and genomic structure of common animal viruses. Animal viruses differ from the bacteriophages and archaeal viruses in that an animal cell takes up the *entire virion* instead of just the viral genome. There are also more than two lifestyles for animal viruses, including virulent infection, latent infection, persistent infection, and cellular transformation (↔ Figure 8.20). A healthy human is teeming with viruses—not just animal viruses but also bacteriophages and possibly even some plant viruses—and we now explore this remarkable suite of viruses living within us and on us.

The Human Body and the Virome

While profiling the viruses of the human body has been hampered by limitations in cell culturing, the power of metagenomics has not only allowed for the human *microbiome* to be characterized (Chapter 24), but also the human **virome**. The human virome encompasses the entire population of viruses present in and on the

human body (Figure 10.29). And, as for the human microbiome, the human virome is both unique to an individual and relatively stable over long periods.

Depending on the individual, animal viruses of the human virome include those that cause severe diseases such as hepatitis (Section 10.11) and severe acute respiratory syndrome (SARS, a coronavirus, Section 10.8) as well as those that cause milder acute infections such as influenza (Section 10.9) and viruses of the common cold (rhinoviruses and coronaviruses, Section 10.8, and adenoviruses, Section 10.6). Other common animal viruses of the human virome cause latent infections, such as the herpesviruses (Section 10.7), in particular human cytomegalovirus, present in most human adults.

The viromes of healthy humans are dominated by viruses that contain DNA genomes. Areas of the human body in which the virome has been assessed are the nose, skin, mouth, and gastrointestinal tract (from fecal samples). Animal viruses commonly detected in these areas include the single-stranded DNA anelloviruses and circoviruses and the double-stranded DNA adenoviruses, polyomaviruses, and papillomaviruses (Figure 10.29a). Anelloviruses are nonenveloped viruses that establish persistent infections in the body early in life with no known connection to disease. Circoviruses possess some of the smallest of all viral genomes (Figure 10.1) and are commonly found in poultry and pigs, suggesting that their presence in the human gastrointestinal tract may originate from these food sources.

Adenoviruses are common respiratory viruses that have been detected in children with fevers, but are also found in the nose and upper respiratory tract of healthy humans. Similarly, polyomaviruses

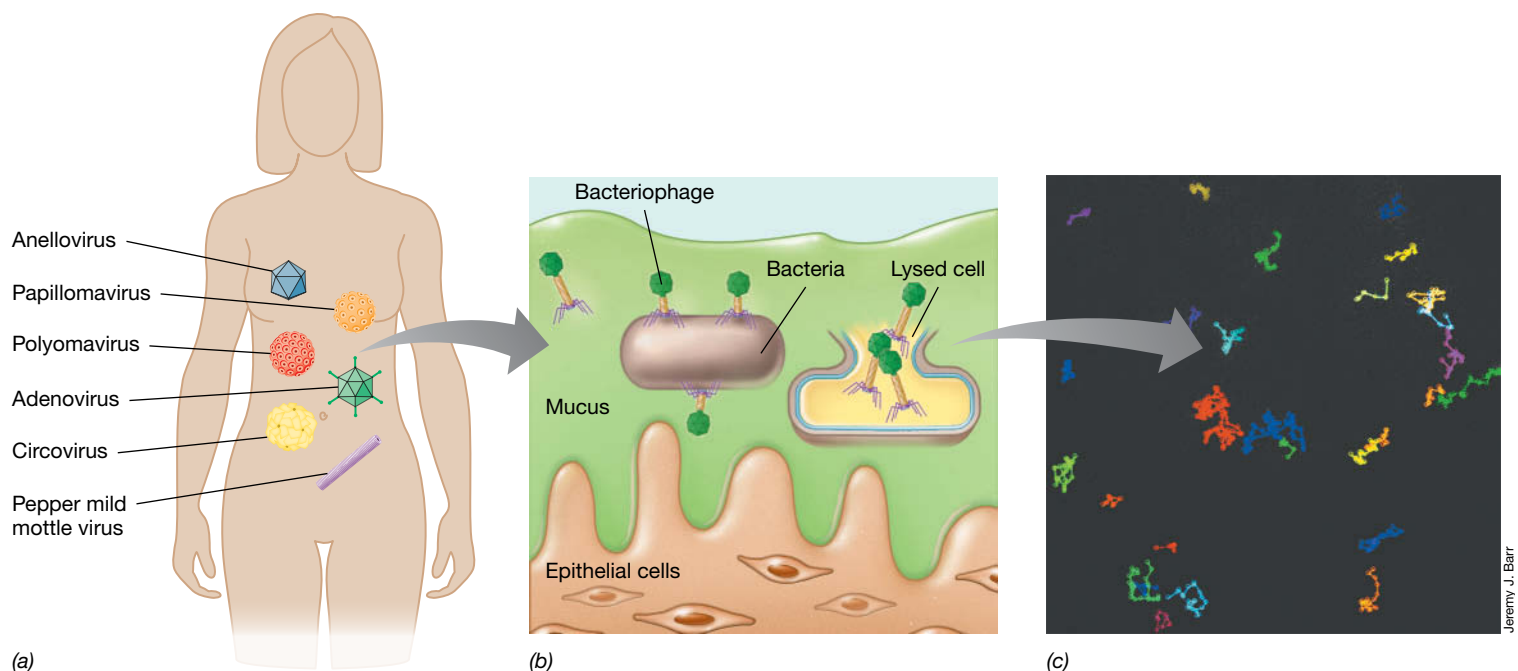


Figure 10.29 The human virome. (a) Common viruses of eukaryotes found in the virome of a healthy human. (b) Bacteriophages within the mucosa of the human respiratory and gastrointestinal tracts protect epithelial cells from the invasion of bacterial pathogens. (c) Movement of T4 bacteriophages in mucus. Tracks represent movement of fluorescently labeled individual bacteriophage virions within a microfluidic chip containing 1% mucin to simulate a mucosal matrix.

are commonly found in healthy individuals, but can also cause a type of brain disorder called leukoencephalopathy as well as urinary tract infections in immunocompromised individuals. Several different papillomaviruses are also common in the human virome, especially in the skin and saliva. *Papillomaviruses* are nonenveloped double-stranded DNA viruses (Baltimore class I, Figure 10.2) that replicate on skin and in mucosal epithelia. Most papillomavirus infections are asymptomatic; however, the human papillomavirus (HPV) causes persistent infections that can develop into skin warts, and certain strains of HPV can cause premalignant lesions in the female reproductive tract that can lead to cervical cancer later in life.

Besides these common animal viruses, most human viromes also contain viruses that infect plants, such as the pepper mild mottle virus (Figure 10.29a). These viruses are undoubtedly transmitted to humans from foods and pass through the intestines. While the host specificity of these viruses is for plants, it is thought that their presence in humans may trigger inflammation, which may lead to some of the symptoms susceptible individuals have to spicy foods or certain plant products. Another possible interaction between the human virome and the immune system is the prevalence of *human endogenous retrovirus* (HERV) elements in human chromosomes; HERVs are remnants of retroviral genes that have been incorporated into and constitute 5–8% of the human genome. Although most HERVs are thought to be harmless, connections have been proposed between certain HERVs and the autoimmune disorders rheumatoid arthritis and systemic lupus erythematosus (↪ Section 27.9), and with other afflictions such as inflammatory bowel (Crohn's) disease and multiple sclerosis.

Bacteriophages and the Human Virome

While every human body contains a unique mixture of different animal viruses, the most abundant viruses in all body sites are not animal viruses but instead are bacteriophages. The large intestine is the hotbed for such viruses, as this organ contains a roughly equal abundance of prokaryotic cells and viruses (about 10^9 of each per gram of feces). As with the animal viruses of the human virome, DNA bacteriophages dominate, and the majority of these viruses are thought to benefit bacterial cells by transferring genes encoding antibiotic resistance or enzymes for specialized metabolisms through the processes of transduction (↪ Section 11.7) and lysogeny (↪ Section 8.7). Genetic transfers from gut bacteriophages to their hosts likely help the gut microbiota adapt to changing nutrient conditions and stabilize it from the stresses of antibiotic treatment.

Bacteriophages of the human virome can also be a first line of defense against certain pathogens, especially within mucosal surfaces where bacteriophages accumulate. It has been estimated that 20 times more bacteriophages than bacteria exist in the mucosa of our lungs and intestines. The bacteriophages present in mucous linings are anchored to sugar residues produced by mucosal cells (Figure 10.29b). Here bacteriophages scavenge invading pathogens and kill them before they can cross the mucous barrier. Thus, phages within the mucous layer can be considered to have a symbiotic relationship with the human host and provide a form of host-independent immunity. Using fluorescently labeled T4 phages, movement of the phages through a model mucus layer can actually be tracked microscopically in the laboratory (Figure 10.29c).

Besides killing or conferring antibiotic resistance to bacteria within the microbiome, the bacteriophage component of the virome can also *enhance* the pathogenicity of certain bacteria. An example of this is the temperate (lysogenic, ↪ Section 8.7) bacteriophage CTX ϕ and its host, the bacterium *Vibrio cholerae* (cholera). It is the phage genome rather than the host genome that actually encodes the cholera toxin that induces disease symptoms (↪ Section 32.3), and cells of *V. cholerae* that are not CTX ϕ lysogens are nonpathogenic. This phage–bacterium link is also seen in the *toxin-coregulated pilus*, a structure essential for *V. cholerae* cells to attach to an intestinal cell. Genes encoding the pilus are part of a viral genome that is integrated into the host's genome (a prophage, ↪ Section 8.7) and is only present in pathogenic strains of *V. cholerae*.

While the human microbiome has been extensively profiled (Chapter 24), these surveys have relied on targeting genes encoding ribosomal RNAs, highly conserved phylogenetic barometers present in all cells. However, because viruses do not possess a universal gene marker, characterization of the human virome has understandably lagged behind that of the cellular microbiome. Nevertheless, recognition of the potentially huge impact on human health by the virome coupled with the continually improving field of metagenomics should bring the human virome into clearer focus in the near future.

MINIQUIZ

- How does the virome differ from the microbiome?
- How do bacteriophages benefit the microbiome?
- How do bacteriophages benefit and harm human hosts?

V • Subviral Agents

We conclude our genomic tour of the viral world by considering two *subviral* agents: the viroids and the prions. These are infectious agents that resemble viruses but lack either nucleic acid (prions) or protein (viroids) and are thus not viruses.

10.15 Viroids

Viroids are infectious RNA molecules that lack a protein component. Viroids are small, circular, single-stranded RNA molecules that are the smallest known pathogens. They range in size from 246 to 399 nucleotides and show a considerable degree of sequence homology to each other, suggesting that they have common evolutionary roots. Viroids cause a number of important plant diseases and can have a severe agricultural impact (Figure 10.30). A few well-studied viroids include coconut cadang-cadang viroid (246 nucleotides) and potato spindle tuber viroid (359 nucleotides). No viroids are known that infect animals or microorganisms.

Viroid Structure and Function

The extracellular form of a viroid is naked RNA; there is no protein capsid. Although the viroid RNA is a single-stranded, covalently closed circle, its extensive secondary structure forms a hairpin-shaped double-stranded molecule with closed ends (Figure 10.31).



Yijun Qi and Biao Ding

Figure 10.30 Viroids and plant diseases. Photograph of healthy tomato plant (left) and one infected with potato spindle tuber viroid (PSTV) (right). The host range of most viroids is quite restricted. However, PSTV infects tomatoes as well as potatoes, causing growth stunting, a flat top, and premature plant death.

This apparently makes the viroid sufficiently stable to exist outside the host cell. Because it lacks a capsid, a viroid does not use a receptor to enter the host cell. Instead, the viroid enters a plant cell through a wound, as from insect or other mechanical damage. Once inside, viroids move from cell to cell via plasmodesmata, which are the thin strands of cytoplasm that link plant cells (Figure 10.32).

Viroid RNAs do not encode proteins and thus the viroid is totally dependent on its host for replication. Plants have several RNA polymerases, one of which has RNA replicase activity, and this is the enzyme that replicates the viroid. The replication mechanism itself resembles the rolling circle mechanism used for genome synthesis by some small viruses (Sections 10.3 and 10.7). The result is a large RNA molecule containing many viroid units joined end to end. The viroid has ribozyme (catalytic RNA) activity and uses it to self-cleave the large RNA molecule, releasing individual viroids.

Viroid Disease

Viroid-infected plants can be symptomless or show symptoms that range from mild to lethal, depending on the viroid (Figure 10.30). Most disease symptoms are growth related, and it is believed that viroids mimic or in some way interfere with plant small regulatory RNAs. In fact, viroids could themselves be derived from regulatory RNAs that have evolved away from carrying out beneficial roles in the cell to inducing destructive events. Viroids are known to yield small interfering RNAs (siRNAs) as a by-product of replication. It has been proposed that these siRNAs may function by way of the RNA-interference silencing pathway to suppress



Figure 10.31 Viroid structure. Viroids consist of single-stranded circular RNA that forms a seemingly double-stranded structure by intra-strand base pairing.

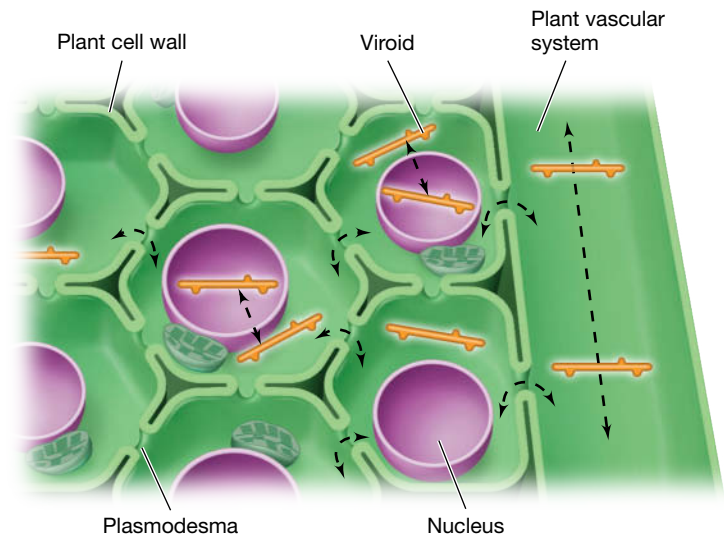


Figure 10.32 Viroid movement inside plants. After entry into a plant cell, viroids (orange) replicate either in the nucleus or the chloroplast. Viroids can move between plant cells via the plasmodesmata (thin threads of cytoplasm that penetrate the cell walls and connect plant cells). On a larger scale, viroids can also move around the plant via the plant vascular system.

the expression of plant genes that show some homology to viroid RNA, and in this way induce disease symptoms. This mechanism of regulation is similar to how some bacterial and archaeal regulatory small RNAs target mRNAs for degradation (Figure 6.27b).

MINIQUIZ

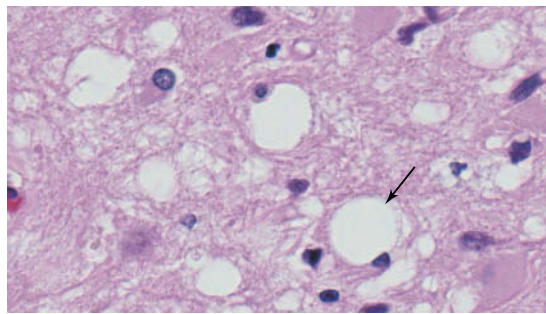
- If viroids are circular molecules, why are they depicted as hairpins?
- How might viroids cause disease in plants?

10.16 Prions

Prions represent the opposite extreme from that of viroids. Prions are infectious agents whose extracellular form consists entirely of protein. That is, a prion lacks both DNA and RNA. Prions cause several neurological diseases such as scrapie in sheep, bovine spongiform encephalopathy (BSE or “mad cow disease”) in cattle, chronic wasting disease in deer and elk, and kuru and variant Creutzfeldt–Jakob disease in humans by catalyzing protein conformational changes that lead to protein clumping and accumulation. No prion diseases of plants are known, although prions have been detected in yeast. Collectively, animal prion diseases are known as *transmissible spongiform encephalopathies*.

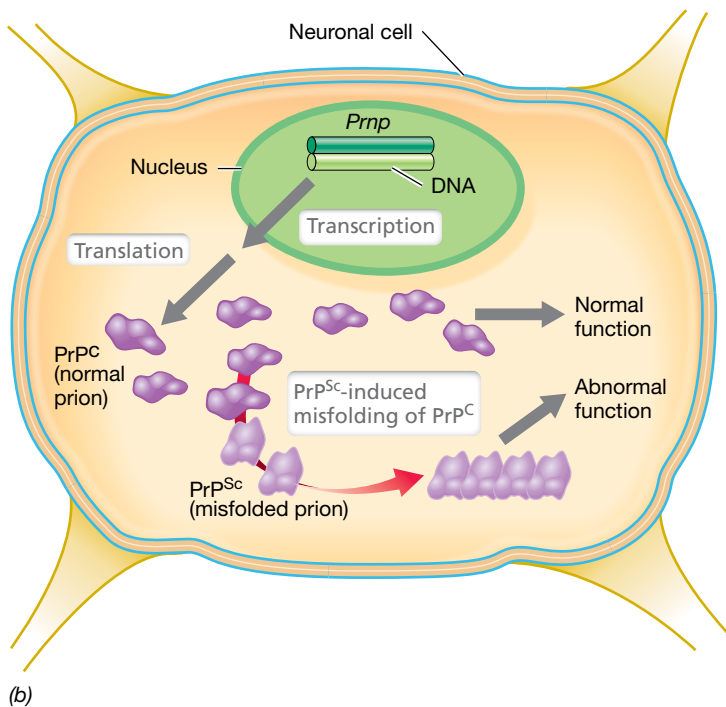
Prion Proteins and the Prion Infectious Cycle

If prions lack nucleic acid, how is prion protein encoded? The answer to this conundrum is that the host cell itself encodes the prion. The host contains a gene, *Prnp* (Prion protein), which encodes the native form of the prion, known as *PrP^C* (Prion Protein Cellular). This is primarily found in the neurons of healthy animals, especially in the brain (Figure 10.33a). The pathogenic form of the prion protein is designated *PrP^{Sc}* (prion protein



(a)

CDC/PHIL, Teresa Hammett



(b)

Figure 10.33 Prions. (a) Section through brain tissue of a human with variant Creutzfeldt-Jakob disease. Note the spongy nature of the tissue (clearings, arrow) where neural tissue has been lost. (b) Mechanism of prion misfolding. Neuronal cells produce the native form of the prion protein. The pathogenic form catalyzes the refolding of native prions into the pathogenic form. The pathogenic form is protease resistant, insoluble, and forms aggregates in neural cells. This eventually leads to destruction of neural tissues (see part a) and neurological symptoms.

Scrapie) because the first prion disease to be discovered was that of scrapie in sheep. PrP^{Sc} is identical in amino acid sequence to PrP^C from the same animal species, but it has a different conformation. For example, native prion proteins are largely α -helical, whereas the pathogenic forms contain less α -helix and more β -sheet secondary structure. Prion proteins from different species of mammals are similar but not identical in amino acid sequence, and host range is linked in some way to protein sequence. For example, PrP^{Sc} from BSE-diseased cattle can infect humans, whereas PrP^{Sc} from scrapie-infected sheep apparently cannot.

When the PrP^{Sc} form enters a host cell that is expressing PrP^C, it promotes the conversion of PrP^C into the pathogenic form (Figure 10.33b). That is, the pathogenic prion “replicates” by converting

preexisting native prions into the pathogenic form. As the pathogenic prions accumulate and aggregate, they form insoluble crystalline fibers referred to as amyloids in neural cells; this leads to disease symptoms including the destruction of brain and other nervous tissue (Figure 10.33a). PrP^C functions in the cell as a cytoplasmic membrane glycoprotein, and it has been shown that membrane attachment of pathogenic prions is necessary before disease symptoms commence. Mutant versions of PrP^{Sc} that can no longer attach to nerve cell cytoplasmic membranes still aggregate but no longer cause disease.

Besides the transmissible spongiform encephalopathies, amyloids are also associated with debilitating human diseases such as Alzheimer’s, Huntington’s, Parkinson’s, and type 2 diabetes. Whether all of these are truly prion diseases is unclear. However, protein aggregation has been linked to these diseases, so the possibility remains that they are manifestations of prion infection.

Nonpathogenic Prions

Certain fungi have proteins that fit the prion definition of an inherited self-perpetuating change in protein structure, although these proteins do not cause noticeable disease. Instead they adapt the fungal cells to survive dynamic environmental conditions by conferring traits such as altered nutrient utilization, antibiotic resistance, and biofilm formation. In yeast, for example, the [URE3] prion is a protein that regulates the transcription of genes encoding certain nitrogen metabolism functions. The normal (soluble) form of this protein functions to repress genes encoding proteins that metabolize certain nitrogen sources. However, when the [URE3] prion accumulates, it forms insoluble aggregates and when this occurs, the normally repressed genes are derepressed and the nitrogen sources are metabolized.

Humans also have beneficial prions. MAVS is a human protein that is part of our innate immune system and has been shown to convert to a self-perpetuating prion-like form in cells that become infected with a virus. The aggregation of MAVS protein triggers the production of immune modulators called *interferons* (see Section 26.10). These proteins trigger the recruitment of phagocytic cells such as macrophages (cells that ingest and destroy pathogens, foreign particles, and cell debris) and other immune factors to the viral-infected cell, resulting in its destruction. While the conversion of MAVS to a prion-like form ultimately kills the cell, it prevents the virus from hijacking the cell’s molecular machinery, which would allow the virus to replicate and lyse the host cell to infect adjacent cells.

MINIQUIZ

- On what basis can prions be differentiated from all other infectious agents?
- What is the difference between the native and pathogenic forms of the prion protein?
- How does a prion differ from a viroid? How does a prion differ from a virus?

Chapter Review

I • Viral Genomes and Evolution

10.1 Viral genomes can be single-stranded or double-stranded DNA or RNA and vary from a few to hundreds of kilobases in size. Viral mRNA is always of the plus configuration by definition, but single-stranded genomes can be of the plus or minus configuration. Viruses with RNA genomes must either carry an RNA replicase in their virions or encode this enzyme in their genomes in order to synthesize RNA from an RNA template.

Q Describe the classes of viruses based on their genomic characteristics. For each class, describe how viral mRNA is made and how the viral genome is replicated.

10.2 Viruses may have evolved as agents of gene transfer in cells, or they may be the remnants of “virocells” that contained RNA genomes and eventually streamlined their biology until they became dependent on a host cell for replication. RNA viruses are likely more ancient than DNA viruses, and the latter may have triggered the transition from RNA genomes to DNA genomes in cells.

Q How would virocells have differed from the RNA viruses we know today, and how would they have been similar?

II • DNA Viruses

10.3 Single-stranded DNA viruses contain DNA of the plus configuration, and a double-stranded replicative form is necessary for transcription and genome replication. The genome of the virus ϕ X174 is so small that some of its genes overlap, and the genome replicates by a rolling circle mechanism. Some related viruses, such as M13, have filamentous virions that are released from the host cell without lysis.

Q Describe how the genome of bacteriophage ϕ X174 is transcribed and translated.

10.4 The head-and-tail bacteriophage T7 contains a double-stranded DNA genome that encodes both early genes, transcribed by the host RNA polymerase, and late genes, transcribed by a virus-encoded RNA polymerase. Replication of the T7 genome employs T7 DNA polymerase and involves terminal repeats and concatemers. Bacteriophage Mu is a temperate virus that is also a transposable element. Mu replicates by transposition in the host chromosome.

Q Why can it be said that transcription of the bacteriophage T7 genome requires two enzymes?

Why is bacteriophage Mu mutagenic, and what features are necessary for Mu to insert into DNA?

10.5 Several double-stranded DNA viruses infect cells of *Archaea*, most of which inhabit extreme environments. Many of these genomes are circular, in contrast to the linear double-stranded DNA genomes of bacteriophages. Although many head-and-tail-type viruses are known, other archaeal viruses have an unusual spindle-shaped morphology.

Q How is a spindle-shaped double-stranded DNA archaeal virus infecting *Acidianus* capable of surviving in an extremely hot and acidic environment?

10.6 Pox viruses are large double-stranded DNA viruses that replicate entirely in the cytoplasm and are responsible for several human diseases, including smallpox. Adenoviruses are double-stranded DNA viruses whose genome replication employs protein primers and a mechanism that occurs without lagging-strand synthesis.

Q Of all the double-stranded DNA animal viruses, pox viruses stand out concerning one unique aspect of their DNA replication process. What is this unique aspect, and how can this be accomplished without special DNA replication enzymes being packaged in the virion?

10.7 Some double-stranded DNA viruses cause cancer in humans. SV40 is such a tumor virus and has a tiny genome containing overlapping genes. The virus can trigger cell transformation (tumor induction) from the activity of certain genes. Some herpesviruses also cause cancer, but most cause various human infectious diseases. Herpesviruses can maintain themselves in a latent state in the host indefinitely, initiating viral replication periodically.

Q How does the way herpesviruses obtain their envelope differ from other enveloped viruses?

III • Viruses with RNA Genomes

10.8 In single-stranded plus-sense RNA viruses, the genome is also the mRNA, and a negative strand is synthesized to produce more mRNA and genome copies. The tiny bacteriophage MS2 contains only four genes, one of which encodes a subunit of its RNA replicase. In poliovirus, the viral RNA is translated directly, producing a polyprotein that is cleaved into several small viral proteins. Coronaviruses are large RNA viruses that resemble poliovirus in some but not all of their replication features.

Q What is the function of the VPg protein of poliovirus, and how can coronaviruses replicate without a VPg protein?

10.9 In negative-strand viruses, the virus RNA is not mRNA but must first be copied to form mRNA by RNA replicase present in the virion. The positive strand is the template for production of genome copies. Important pathogenic negative-strand viruses include rabies virus and influenza virus.

Q Rabies virus and poliovirus both have single-stranded RNA genomes, but only in poliovirus can the genome be translated directly. Explain.

10.10 Reoviruses contain segmented linear double-stranded RNA genomes. Like negative-strand RNA viruses, reoviruses contain an RNA-dependent RNA polymerase within the virion. All replication events occur within newly forming virions.

Q Compare the reovirus genome to those of influenza virus and bacteriophage MS2. Why do reovirus replication events have to happen in the nucleocapsid?

10.11 Some viruses employ reverse transcriptase, including retroviruses (HIV) and hepadnaviruses (hepatitis B). Retroviruses have single-stranded RNA genomes and use reverse transcriptase to make a DNA copy. Hepadnaviruses contain partially double-stranded DNA genomes and use reverse transcriptase to make a single strand of genomic DNA from a full-length complementary strand of RNA.

Q Why do both hepadnaviruses and retroviruses require reverse transcriptase when their genomes are double-stranded DNA and single-stranded RNA, respectively?

IV • Viral Ecology

10.12 The number of viruses on Earth is greater than the number of cells by 10-fold. Most of the genetic diversity on Earth resides in virus genomes, most of which are still to be investigated. Viruses affect their host cells by either culling the host population or by carrying out horizontal gene transfer from one bacterial cell to another. In the oceans, both *Bacteria* and *Archaea* are likely to be infected with viruses.

Q How do viral numbers compare to those of bacteria in seawater?

10.13 Prokaryotic cells employ various mechanisms to defend against viral infection. However, the fast mutation rate and genome plasticity of viruses allows them to adapt to these strategies. The CRISPR system is a type of immune system for *Bacteria* and *Archaea* that resists viral infection by recognizing and degrading incoming foreign DNA.

Q Explain CRISPR and its mechanism.

10.14 The human virome is the entire population of viruses associated with the human body. Each individual possesses a unique and relatively stable virome. Animal viruses that cause asymptomatic infections are common in a healthy individual's virome. However, bacteriophages are the major constituent of the human virome. These phages can protect against pathogen invasion and also increase the fitness of the microbiome through the genetic transfer of beneficial genes.

Q What type of viruses are most abundant in the human virome? How are they beneficial to humans?

V • Subviral Agents

10.15 Viroids are circular single-stranded RNA molecules that do not encode proteins and are dependent on host-encoded enzymes for replication. Unlike viruses, viroid RNA is not enclosed within a capsid, and all known viroids are plant pathogens.

Q What are the similarities and differences between viruses and viroids?

10.16 Prions consist of protein but have no nucleic acid of any kind. Prions exist in two conformations, the native cellular form and the pathogenic form, which takes on a different protein structure. The pathogenic form “replicates” itself by converting native prion proteins, encoded by the host cell, into the pathogenic conformation.

Q What are the similarities and differences between prions and viruses?

Application Questions

- Not all proteins are made from the RNA genome of bacteriophage MS2 in the same amounts. Can you explain why? One of the proteins functions very much like a repressor, but it functions at the translational level. Which protein is it and how does it function?
- Replication of both strands of DNA in adenoviruses occurs in a continuous (leading) fashion. How can this happen without violating the rule that DNA synthesis always occurs in a 5' → 3' direction?
- Imagine that you are a researcher at a pharmaceutical company charged with developing new drugs against human RNA viral pathogens. Describe at least two types of drugs you might pursue, what class of virus they would affect, and why you feel that the drugs would not harm the patient.
- Reoviruses contain genomes that are unique in all of biology. Why? Why can't reovirus replication occur in the host cytoplasm? Contrast reovirus genomic replication events with those of a cell. Why can it be said that reovirus genome replication is not semiconservative even though the reovirus genome consists of complementary strands?

Chapter Glossary

Hepadnavirus a virus whose DNA genome replicates by way of an RNA intermediate

Negative strand a nucleic acid strand that has the opposite sense to (is complementary to) the mRNA

Overlapping genes two or more genes in which part or all of one gene is embedded in another gene

Polyprotein a large protein expressed from a single gene and subsequently cleaved to form several individual proteins

Positive strand a nucleic acid strand that has the same sense as the mRNA

Prion an infectious protein whose extracellular form lacks nucleic acid

Replicative form a double-stranded molecule that is an intermediate in the replication of viruses with single-stranded genomes

Retrovirus a virus whose RNA genome has a DNA intermediate as part of its replication cycle

Reverse transcription the process of copying genetic information found in RNA into DNA

RNA replicase an enzyme that can produce RNA from an RNA template

Rolling circle replication a mechanism, used by some plasmids and viruses, of replicating circular DNA, which starts by

nicking and unrolling one strand. For a single-stranded genome, this is preceded by using the still-circular strand as a template for DNA synthesis; for a double-stranded genome, the unrolled strand is used as a template for DNA synthesis

Transposase an enzyme that catalyzes the insertion of DNA segments into other DNA molecules

Viroid an infectious RNA whose extracellular form lacks protein

Virome the entire population of viruses associated with the human body

11

Genetics of *Bacteria* and *Archaea*

microbiologynow

Killing and Stealing: DNA Uptake by the Predator *Vibrio cholerae*

Vibrio cholerae, the causative agent of cholera, is a marine bacterium that can also flourish in the nutrient-rich human intestine. The pathogen reaches new human hosts by releasing a toxin that triggers diarrheal purges, resulting in transmission of the pathogen into new environments. Besides this virulent lifestyle, *V. cholerae* also competes for nutrients in marine environments by employing an arsenal of proteins to kill neighboring microbial cells. Using an elegant contractile structure known as the type VI secretion system (T6SS), predatory *V. cholerae* cells inject toxic molecules called effectors into prey cells. Cells lacking immunity to these effectors ultimately lyse and release their cellular contents.

While predation reduces competition and releases nutrients, microbiologists have recently observed another fascinating aspect of the “assassin” lifestyle of *V. cholerae*. Once prey cells have been killed using T6SS, *V. cholerae* predator cells can scavenge the DNA released from their

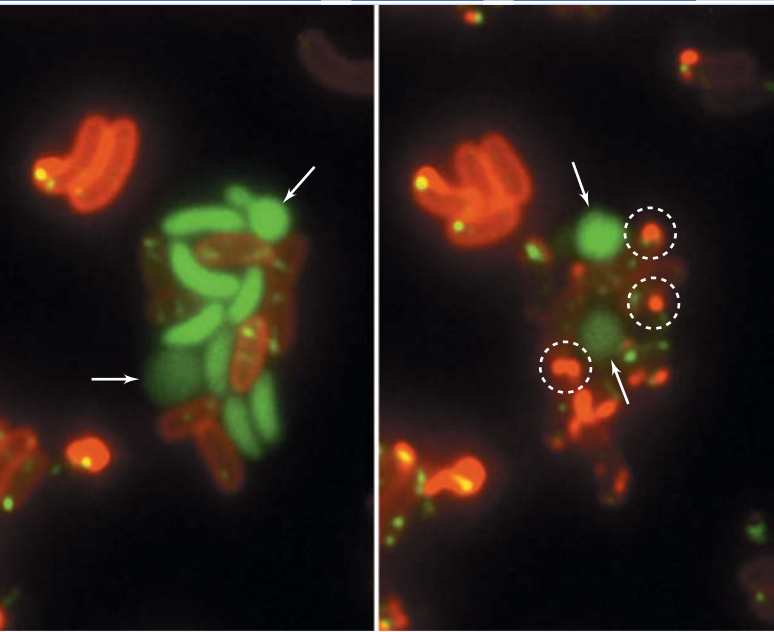
victims using a special DNA-uptake system (natural competence). This system facilitates horizontal gene transfer, as it can incorporate DNA fragments containing up to 40 genes. If prey cell genes recombine with the *V. cholerae* chromosome, the *V. cholerae* population stands to benefit from the transformation event.

These images show predator cells killing and stealing DNA from defenseless prey cells. DNA-uptake proteins of the competence system are labeled with a red fluorescent tag and proteins of the T6SS with a green fluorescent tag so that predatory *V. cholerae* cells appear red with green “crossbows” inside. In contrast, prey lacking T6SS and immunity proteins appear as solid green cells. The panel on the left shows red predator cells with their green crossbows attacking prey cells. The white arrows point to dead prey cells, which circularize upon death. The panel on the right shows the same population of cells 30 minutes later. As the predator cells take up DNA from killed prey, the corresponding DNA-binding proteins localize for effective uptake (circled).

The antagonistic nature of *V. cholerae* provides this predatory bacterium with a mechanism for stealing valuable traits such as antibiotic resistance and virulence factors from its prey by exploiting horizontal gene transfer. Thus, this form of microbial warfare not only reduces competition for nutrients but also functions to increase the predator’s genetic fitness.



Source: Borgeaud, S., L.C. Metzger, T. Scignari, and M. Blokesch. 2015. The type VI secretion system of *Vibrio cholerae* fosters horizontal gene transfer. *Science* 347: 63–67.



- I Mutation 343
- II Gene Transfer in *Bacteria* 349
- III Gene Transfer in *Archaea* and Other Genetic Events 360

In 1946 the microbiologist Joshua Lederberg made a groundbreaking discovery—like plants and animals, bacteria can also exchange genes! Lederberg’s work showcasing genetic recombination in bacteria not only earned him a Nobel Prize but also helped launch the field of molecular biology and the use of bacteria to study how genes work in higher organisms such as animals.

Understanding the varied mechanisms by which *Bacteria* and *Archaea* exchange genes has helped tackle the conundrum of how these microbes can exhibit so much diversity while reproducing asexually. Gene exchange, along with genetic innovations that arise from random changes in a cell’s genetic blueprint, confer selectable advantages that ultimately drive genetic diversity.

In this chapter we discuss mechanisms of genetic exchange in *Bacteria* and *Archaea*. We first describe how changes arise in the genome, and then we consider how *horizontal gene transfer* can move genes from one cell to another. While changes to the genome underlie microbial diversity and habitat adaptation, microorganisms also possess mechanisms to maintain genomic stability, and we end this chapter by considering these. Taken together, both genomic change and genomic stability are important to the evolution of an organism and its competitive success in nature.

I • Mutation

All organisms contain a specific sequence of nucleotide bases in their genome, their genetic blueprint. A **mutation** is a *heritable* change in the base sequence of that genome, that is, a change that is passed from the mother cell to progeny cells. Mutations can lead to changes in the properties of an organism; some mutations are beneficial, some are detrimental, but most are neutral and have no effect. Although the rate of spontaneous mutation is low (Section 11.3), the speed at which many prokaryotic cells divide and their characteristic exponential growth ensure that mutations accumulate in a population surprisingly fast. Moreover, whereas a single mutation typically brings about only a small change in a cell, genetic exchange often generates much larger change. Taken together, mutation and genetic exchange fuel the evolutionary process.

We begin by considering the molecular mechanism of mutation and the properties of mutant microorganisms.

11.1 Mutations and Mutants

The genomes of cells consist of double-stranded DNA. In viruses, by contrast, the genome may consist of double- or single-stranded DNA (or RNA) (Chapters 8 and 10). By convention, a strain of an organism or a virus isolated from nature is referred to as the **wild-type strain** and therefore contains the wild-type genome. A cell or virus derived from the wild type that carries a change in nucleotide sequence is called a **mutant**. A mutant by definition differs from the wild-type strain in its **genotype**, the nucleotide sequence of its genome. In addition, the observable properties of the mutant—its **phenotype**—may also be altered relative to its parent (Figure 11.1). This altered phenotype is called a *mutant phenotype*.

The term “wild-type” may be used to refer to an entire organism or just to the status of a particular gene that is under investigation.

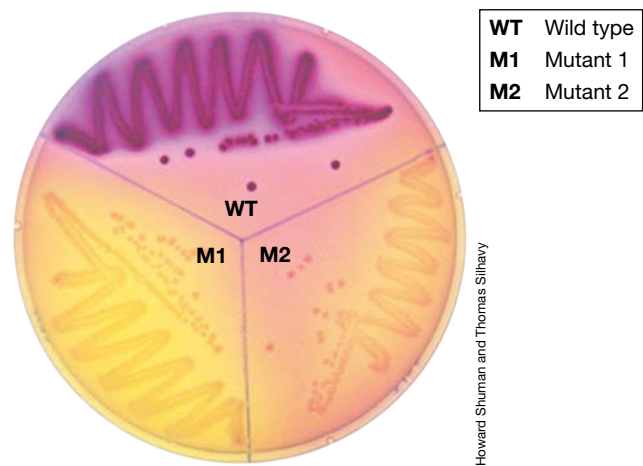


Figure 11.1 Wild-type versus mutant phenotype. Growth of wild-type *Escherichia coli* and maltose utilization mutants on a plate of MacConkey agar, a differential medium. The medium contains maltose as the carbon source and a pH indicator that turns red if maltose is fermented. Mutants 1 and 2 are unable to ferment maltose due to a deletion of the *malB* gene and a point mutation in the *malQ* gene, respectively.

Mutant derivatives can be obtained either directly from a wild-type strain or from another strain—referred to as a *parental strain*—previously derived from the wild type; for example, another mutant. Figure 11.1 shows a plate of MacConkey agar (a culture medium that contains a pH indicator that turns red if sugar is fermented) that shows the phenotypic difference between wild-type *Escherichia coli* and mutant derivatives in the sugar utilization pathway.

Depending on the mutation, a mutant strain may or may not differ in phenotype from its parent. By convention in bacterial genetics, the *genotype* of an organism is designated by three lowercase letters followed by a capital letter (all in italics) indicating a particular gene. For example, the *hisC* gene of *E. coli* encodes a protein called HisC that functions in biosynthesis of the amino acid histidine. Mutations in the *hisC* gene would be designated as *hisC1*, *hisC2*, and so on, the numbers referring to the order of isolation of the mutant strains. Each *hisC* mutation would be different, and each *hisC* mutation might affect the HisC protein in different ways.

The *phenotype* of an organism is designated by a capital letter followed by two lowercase letters, with either a plus or minus superscript to indicate the presence or absence of that property. For example, a His^+ strain of *E. coli* is one that is capable of making its own histidine, whereas a His^- strain is not. The His^- strain would therefore require a histidine supplement for growth. A mutation in the *hisC* gene will lead to a His^- phenotype if it eliminates the function of the HisC protein.

Isolation of Mutants: Screening versus Selection

Virtually any characteristic of an organism can be changed by mutation. Some mutations are *selectable*, conferring some type of advantage on organisms possessing them, whereas others are nonselectable, even though they may lead to a very clear change in the phenotype of an organism. A selectable mutation confers a

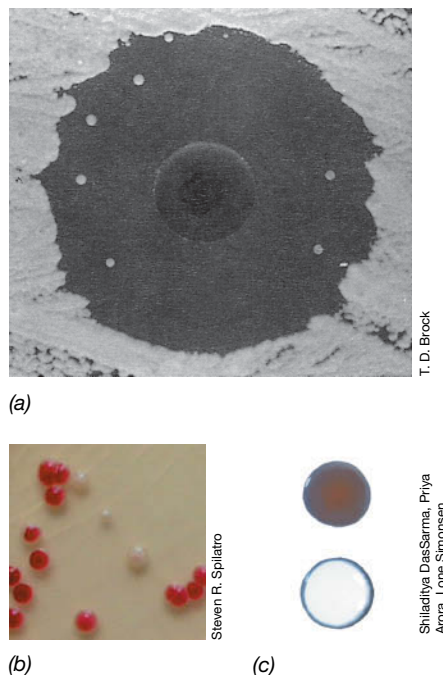


Figure 11.2 Selectable and nonselectable mutations. (a) Development of antibiotic-resistant mutants, a type of easily selectable mutation, within the inhibition zone of an antibiotic assay disc. (b) Nonselectable mutations. UV-radiation-induced nonpigmented mutants of *Serratia marcescens*. The wild type has a dark red pigment. The white or colorless mutants make no pigment. (c) Colonies of mutants of a species of *Halobacterium*, a member of the *Archaea*. The wild-type colonies are white. The orangish-brown colonies are mutants that lack gas vesicles (↔ Section 2.9). The gas vesicles scatter light and mask the color of the colony.

clear advantage on the mutant strain under certain environmental conditions, so the progeny of the mutant cell are able to outgrow and replace the parent. A good example of a selectable mutation is drug resistance: An antibiotic-resistant mutant can grow in the presence of an antibiotic that inhibits or kills the parent (Figure 11.2a) and is thus selected under these conditions. It is relatively easy to detect and isolate selectable mutants by choosing the appropriate environmental conditions. **Selection** is therefore an extremely powerful genetic tool, allowing the isolation of a single mutant from a population containing millions or even billions of parental cells.

An example of a nonselectable mutation is color loss in a pigmented organism (Figure 11.2b, c). Nonpigmented cells usually have neither an advantage nor a disadvantage over the pigmented parent cells when grown in the laboratory, although pigmented organisms may have a selective advantage in nature. We can detect nonselectable mutations only by examining large numbers of colonies and looking for the “different” ones, a process called **screening**. In microbial genetics, screening is typically a much more laborious and time-consuming process than is selection. Thus in a genetic experiment if selection is possible, it is almost always the preferred strategy.

Isolation of Nutritional Auxotrophs

Although screening is more tedious than selection, useful methods have been developed for screening large numbers of colonies for certain types of mutations. For instance, nutritionally defective mutants

can be detected by the technique of *replica plating* (Figure 11.3). A colony from a master plate can be transferred onto an agar plate lacking the nutrient by using a sterile loop, toothpick, or even a robotic arm. Parental colonies will grow normally, whereas those of the mutant will not. Thus, the inability of a colony to grow on medium lacking the nutrient signals that it is a mutant. The colony on the master plate corresponding to the vacant spot on the replica plate can then be picked, purified, and characterized.

A mutant strain with an additional nutritional requirement for growth is called an **auxotroph**, and the parental strain from which it was derived is called a *prototroph*. For instance, mutants of *E. coli* with a His^- phenotype are histidine auxotrophs, while the parental His^+ strain from which the auxotroph was derived is the prototroph of such strains. As described earlier, many different mutations can lead to a strain showing a His^- phenotype, and thus an initial step in characterizing the genetics of a metabolic pathway (such as histidine biosynthesis) would be the isolation of several His^- strains followed by their comparative genetic analyses (Section 11.5).

Examples of common classes of mutants and the means by which they are detected are listed in Table 11.1.

MINIQUIZ

- Distinguish between a mutation and a mutant.
- Distinguish between screening and selection.
- How does an auxotroph differ from a prototroph?

11.2 Molecular Basis of Mutation

Mutations can be either spontaneous or induced. **Spontaneous mutations** are those that occur without external intervention, and most result from occasional errors in the pairing of bases by DNA polymerase during DNA replication. **Induced mutations**, by contrast, are those caused by agents in the environment and include mutations made deliberately by humans. Induced mutations can result from exposure to natural radiation (cosmic rays and so on) that alters the structure of bases in the DNA, or from a variety of chemicals that chemically modify DNA (Section 11.4).

Mutations that change only one base pair are called **point mutations** and occur when a single base-pair substitution takes place in the DNA. Many point mutations do not actually cause any phenotypic change, as discussed below. However, as for all mutations, any phenotypic change that results from a point mutation depends on exactly where in the genome the mutation occurs and the nature of the nucleotide change.

Base-Pair Substitutions: Missense, Nonsense, and Silent Mutations

If a point mutation is within the region of a gene that encodes a polypeptide, any change in the phenotype of the cell is most likely the result of a change in the amino acid sequence of that polypeptide. The error in the DNA is transcribed into mRNA, and the erroneous mRNA in turn is translated to yield a polypeptide. Figure 11.4 shows the consequences of some base-pair substitutions.

In interpreting the results of a mutation, we must first recall that the genetic code is degenerate (↔ Section 4.9 and Table 4.4).

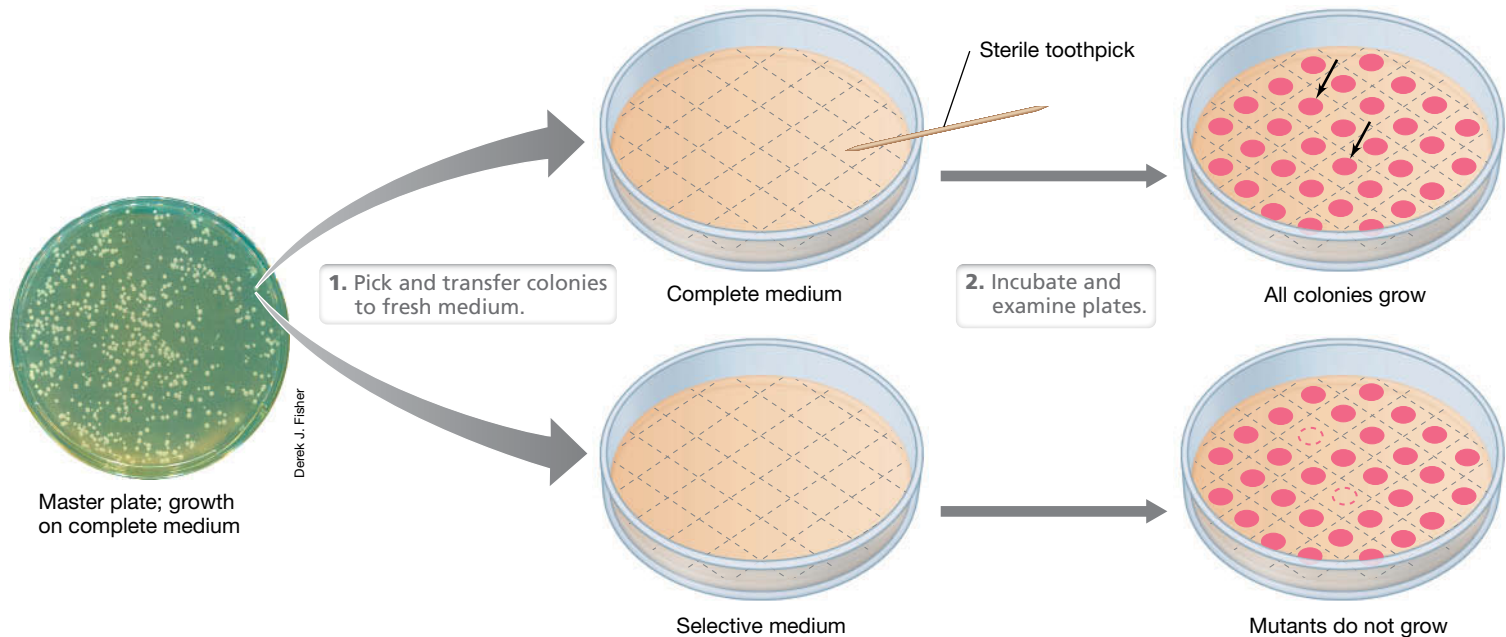


Figure 11.3 Screening for nutritional auxotrophs. The replica-plating method can be used for the detection of nutritional mutants. Colonies from the master plate are transferred using a sterile toothpick to a gridded plate containing different media for selection. The colonies not appearing on the selective medium are indicated with arrows. The selective medium lacked one nutrient (leucine) present in the master plate. Therefore, the colonies indicated with arrows on the master plate are leucine auxotrophs.

Consequently, not all mutations in the base sequence encoding a polypeptide will change the polypeptide. This is illustrated in Figure 11.4, which shows several possible results when the DNA that encodes a single tyrosine codon in a polypeptide is mutated. First, a change in the RNA from UAC to UAU would have no apparent effect because UAU is also a tyrosine codon. Although they do not affect the sequence of the encoded polypeptide, such changes in the DNA are considered one type of **silent mutation**, that is, a mutation that does not affect the phenotype of the cell. Note that

silent mutations in coding regions are almost always in the third base of the codon (arginine and leucine can also have silent mutations in the first position).

Changes in the first or second base of the codon more often lead to significant changes in the amino acid sequence of the polypeptide. For instance, a single base change from UAC to AAC (Figure 11.4) results in an amino acid change within the polypeptide from tyrosine to asparagine at a specific site. This is called a **missense mutation** because the informational “sense” (precise sequence

TABLE 11.1 Some examples of mutants

Phenotype	Nature of change	Detection of mutant
Auxotroph	Loss of enzyme in biosynthetic pathway	Inability to grow on medium lacking the nutrient
Temperature-sensitive	Alteration of an essential protein so it is more heat-sensitive	Inability to grow at a high temperature that normally supports growth
Cold-sensitive	Alteration of an essential protein so it is inactivated at low temperature	Inability to grow at a low temperature that normally supports growth
Drug-resistant	Detoxification of drug or alteration of drug target or permeability to drug	Growth on medium containing a normally inhibitory concentration of the drug
Rough colony	Loss or change in lipopolysaccharide layer	Granular, irregular colonies instead of smooth, glistening colonies
Nonencapsulated	Loss or modification of surface capsule	Small, rough colonies instead of larger, smooth colonies
Nonmotile	Loss of flagella or nonfunctional flagella	Compact instead of flat, spreading colonies; lack of motility by microscopy
Pigmentless	Loss of enzyme in biosynthetic pathway leading to loss of one or more pigments	Presence of different color or lack of color
Sugar fermentation	Loss of enzyme in degradative pathway	Lack of color change on agar containing sugar and a pH indicator
Virus-resistant	Loss of virus receptor	Growth in presence of large amounts of virus

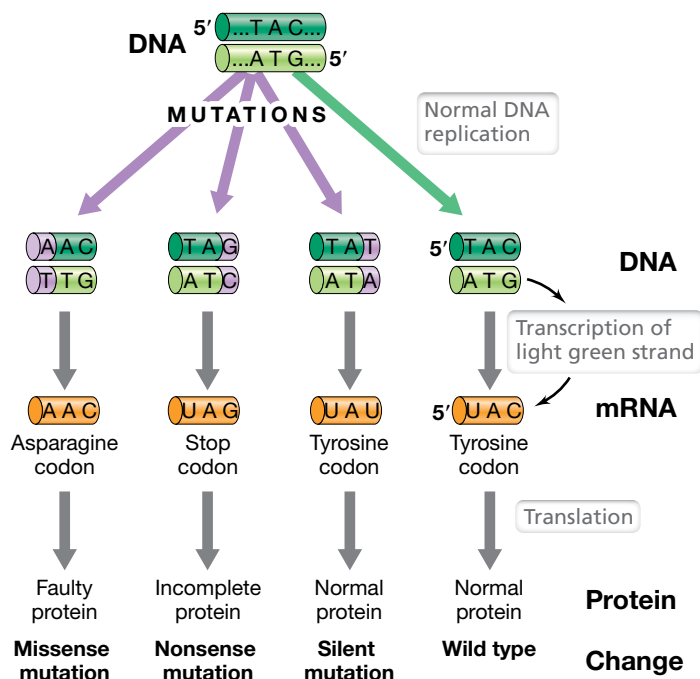


Figure 11.4 Possible effects of base-pair substitution in a gene encoding a protein. Three different protein products are possible from changes in the DNA for a single codon.

of amino acids) in the polypeptide has changed. If the change is at a critical location in the polypeptide chain, the protein could be inactive or have reduced activity. However, not all missense mutations necessarily lead to nonfunctional proteins. The outcome depends on where the substitution lies in the polypeptide chain and on how it affects protein folding and activity. For example, mutations in the active site of an enzyme are more likely to destroy activity than mutations in other regions of the protein.

Another possible outcome of a base-pair substitution is the formation of a nonsense (stop) codon. This results in premature termination of translation, leading to an incomplete polypeptide (Figure 11.4). Mutations of this type are called **nonsense mutations** because the change is from a sense (coding) codon to a nonsense codon (↔ Table 4.4). Unless the nonsense mutation is very near the end of the gene, the product is considered *truncated* (incomplete). Truncated proteins are completely inactive or, at the very least, lack normal activity.

Other terms are occasionally used in microbial genetics to describe the precise type of base substitution in a point mutation. **Transitions** are mutations in which one purine base (A or G) is substituted for another purine, or one pyrimidine base (C or T) is substituted for another pyrimidine. **Transversions** are point mutations in which a purine base is substituted for a pyrimidine base, or vice versa.

Frameshifts and Other Insertions or Deletions

Because the genetic code is read from one end of the nucleic acid in consecutive blocks of three bases (codons), any deletion or insertion of a single base pair results in a shift in the reading frame. These **frameshift mutations** often have serious consequences.

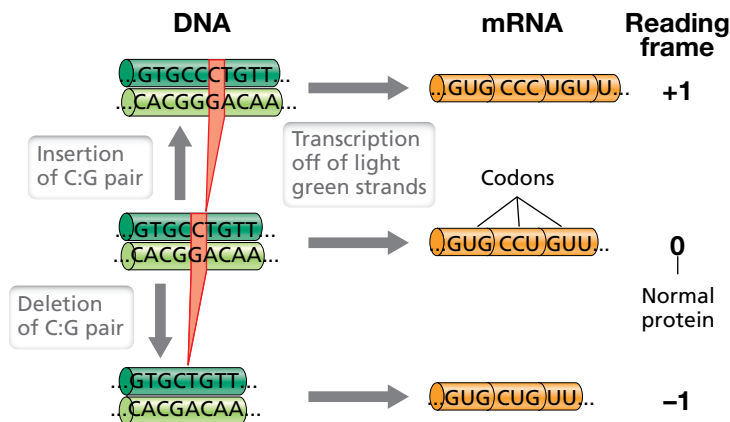


Figure 11.5 Shifts in the reading frame of mRNA caused by insertions or deletions. The reading frame in mRNA is established by the ribosome, which begins at the 5' end (toward the left in the figure) and proceeds by units of three bases (codons). The normal reading frame is referred to as the 0 frame, that missing a base the -1 frame, and that with an extra base the +1 frame.

Single base insertions or deletions change the primary sequence of the encoded polypeptide, typically in a major way (Figure 11.5). Such microinsertions or microdeletions can result from replication errors. Insertion or deletion of two base pairs also causes a frameshift. However, insertion or deletion of three base pairs does not cause a frameshift but does add or remove a codon; this results in the addition or deletion of a single amino acid in the polypeptide sequence. Although an amino acid addition or deletion may well be deleterious to protein function, it is usually not as bad as a frameshift, which scrambles the entire polypeptide sequence downstream of the mutation.

Insertions or deletions can also result in the gain or loss of hundreds or even thousands of base pairs. Such changes inevitably result in complete loss of gene function. Some deletions are so large that they may include several genes. If any of the deleted genes are essential, the mutation will be lethal. Such deletions cannot be restored through further mutations, but only through genetic recombination. Larger insertions and deletions may arise as a result of errors during genetic recombination (Section 11.5). In addition, many large insertion mutations are due to the insertion of specific identifiable DNA sequences called *transposable elements* (Section 11.11). The effect of transposable elements on the evolution of bacterial genomes was discussed in Section 9.6.

MINIQUIZ

- Do missense mutations occur in genes encoding tRNA? Why or why not?
- Why do frameshift mutations generally have more serious consequences than missense mutations?

11.3 Reversions and Mutation Rates

The rates at which different kinds of mutations occur vary widely. Some types of mutations occur so rarely that they are almost impossible to detect, whereas others occur so frequently that they

present difficulties for an experimenter trying to maintain a genetically stable stock culture. Sometimes a second mutation can reverse the effect of an initial mutation. Furthermore, all organisms possess a variety of systems for DNA repair. Consequently, the observed mutation rate depends not only on the frequency of DNA changes but also on the efficiency of DNA repair.

Reversions (Back Mutations) and Suppressors

Point mutations are typically reversible, a process known as **reversion**. A revertant is a strain in which the original phenotype that was changed in the mutant is restored by a second mutation. Revertants can be of two types, same site or second site. In *same-site* revertants, the mutation that restores activity is at the same site as the original mutation. If the back mutation is not only at the same site but also restores the original sequence, it is called a *true* revertant.

In *second-site* revertants, the mutation is at a different site in the DNA. Second-site mutations can restore a wild-type phenotype if they function as *suppressor mutations*—mutations that compensate for the effect of the original mutation. Several classes of suppressor mutations are known. These include (1) a mutation somewhere else in the same gene that restores enzyme function, such as a second frameshift mutation near the first that restores the original reading frame; (2) a mutation in another gene that restores the function of the original mutated gene; and (3) a mutation in another gene that results in the production of an enzyme that can replace the nonfunctional one.

Suppressors can be best illustrated by mutations in tRNAs. Nonsense mutations can be suppressed by changing the anticodon sequence of a tRNA molecule so that it now recognizes a stop codon (Figure 11.6). Such an altered tRNA is called a *suppressor tRNA* and will insert the amino acid it carries at the stop codon that it now reads. Suppressor tRNA mutations would be lethal unless a cell has more than one tRNA for a particular codon. One tRNA may then be mutated into a suppressor, while the other performs the original function. Most cells have multiple tRNAs and so suppressor mutations are reasonably common, at least in microorganisms. Sometimes the amino acid inserted by the suppressor tRNA is identical to the original amino acid and the protein is fully active. In other cases, however, a different amino acid is inserted and only a partially active protein may be produced.

Mutation Rates

For most microorganisms, errors in DNA replication occur at a frequency of 10^{-6} to 10^{-7} per thousand bases during a single round of replication. A typical gene has about 1000 base pairs. Therefore, the frequency of a mutation *in a given gene* is also in the range of 10^{-6} to 10^{-7} per round of replication. For instance, in a bacterial culture having 10^8 cells/ml, there are likely to be a number of different mutants for any given gene in each milliliter of culture. Eukaryotes with very large genomes tend to have replication error rates about 10-fold lower than typical bacteria, whereas DNA viruses, especially those with very small genomes, may have error rates 100-fold to 1000-fold higher than those of cellular organisms. RNA viruses have even higher error rates due to less proofreading (Section 4.4) and the lack of RNA repair mechanisms.

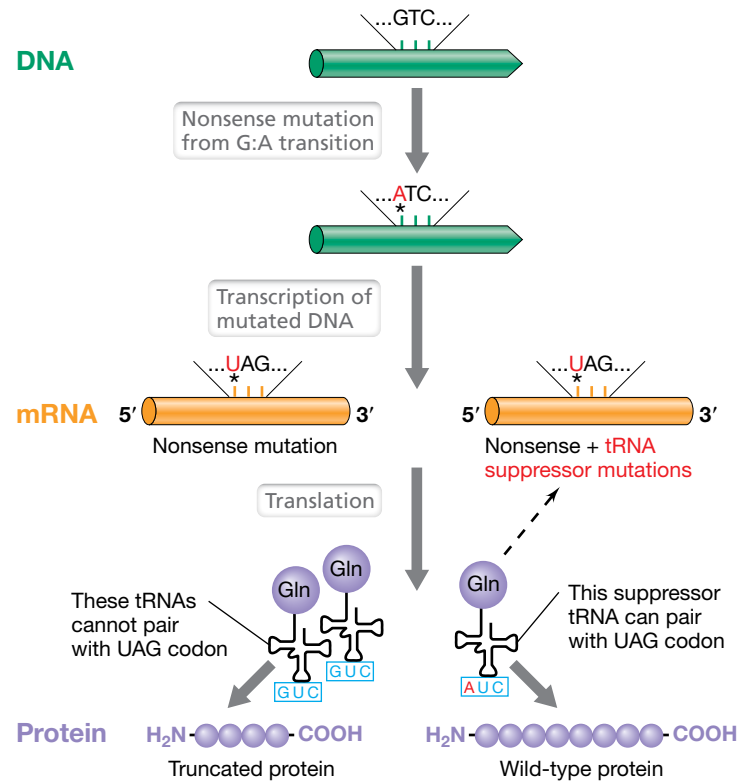


Figure 11.6 Suppression of nonsense mutations. Introduction of a nonsense mutation in a gene encoding a protein results in the incorporation of a stop codon (indicated by the *) in the corresponding mRNA. This single mutation leads to the production of a truncated polypeptide. The mutation is suppressed if a second mutation occurs in the anticodon of a tRNA, a tRNA charged with glutamine in this example, which allows the mutated tRNA or suppressor tRNA to bind to the nonsense codon.

Single base errors during DNA replication are more likely to lead to missense mutations than to nonsense mutations because most single base substitutions yield codons that encode other amino acids (Table 4.4). The next most frequent type of codon change caused by a single base change leads to a silent mutation. This is because for the most part alternate codons for a given amino acid differ from each other by a single base change in the “silent” third position. A given codon can be changed to any of 27 other codons by a single base substitution, and on average, about two of these will be silent mutations, one a nonsense mutation, and the rest missense mutations.

Unless a mutation can be selected for, its experimental detection is difficult, and much of the skill of the microbial geneticist requires increasing the efficiency of mutation detection. This can be done most effectively by increasing the pool of mutations. As we see in the next section, it is possible to greatly increase the mutation rate by treatment with mutagenic agents. In addition, the mutation rate may change under certain circumstances, such as when cells are placed under high-stress conditions.

MINIQUIZ

- Why are suppressor tRNA mutations not lethal?
- Which class of mutation, missense or nonsense, is more common, and why?

11.4 Mutagenesis

The spontaneous rate of mutation is very low, but a variety of chemical, physical, and biological agents can increase the mutation rate and are therefore said to induce mutations. These agents are called **mutagens**, and we discuss some of the major categories of mutagens and their activities here.

Chemical Mutagens and Radiation

An overview of some of the major chemical mutagens and their modes of action is given in **Table 11.2**. Several classes of chemical mutagens exist. The *nucleotide base analogs* are molecules that resemble the purine and pyrimidine bases of DNA in structure yet display faulty base-pairing properties (**Figure 11.7**). If a base analog is incorporated into DNA in place of the natural base, the DNA may replicate normally most of the time. However, DNA replication errors occur at higher frequencies at these sites due to incorrect base pairing. The result is the incorporation of a mismatched base into the new strand of DNA and thus introduction of a mutation. During subsequent segregation of this strand in cell division, the mutation is revealed.

Other chemical mutagens induce *chemical modifications* in one base or another, resulting in faulty base pairing or related changes (Table 11.2). For example, alkylating agents (chemicals that react

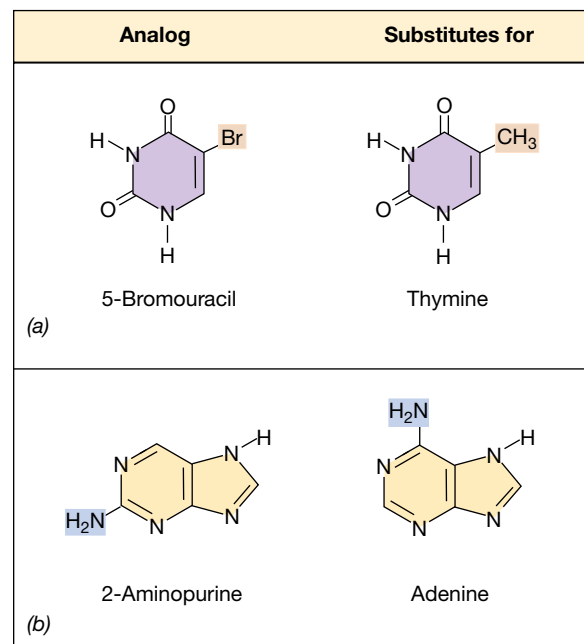


Figure 11.7 Nucleotide base analogs. Structure of two common nucleotide base analogs used to induce mutations and the normal nucleic acid bases for which they substitute. (a) 5-Bromouracil can base-pair with guanine, causing AT to GC substitutions. (b) 2-Aminopurine can base-pair with cytosine, causing AT to GC substitutions.

TABLE 11.2 Chemical and physical mutagens and their modes of action

Agent	Action	Result
Base analogs		
5-Bromouracil	Incorporated like T; occasional faulty pairing with G	AT → GC and occasionally GC → AT
2-Aminopurine	Incorporated like A; faulty pairing with C	AT → GC and occasionally GC → AT
Chemicals reacting with DNA		
Nitrous acid (HNO ₂)	Deaminates A and C	AT → GC and GC → AT
Hydroxylamine (NH ₂ OH)	Reacts with C	GC → AT
Alkylating agents		
Monofunctional (for example, ethyl methanesulfonate)	Puts methyl on G; faulty pairing with T	GC → AT
Bifunctional (for example, mitomycin, nitrogen mustards, nitrosoguanidine)	Cross-links DNA strands; faulty region excised by DNase	Both point mutations and deletions
Intercalating agents		
Acridines, ethidium bromide	Inserts between two base pairs	Microinsertions and microdeletions
Radiation		
Ultraviolet (UV)	Pyrimidine dimer formation	Repair may lead to error or deletion
Ionizing radiation (for example, X-rays)	Free-radical attack on DNA, breaking chain	Repair may lead to error or deletion

with amino, carboxyl, and hydroxyl groups by substituting them with alkyl groups) such as nitrosoguanidine are powerful mutagens and generally induce mutations at higher frequency than base analogs. Unlike base analogs, which have an effect only when incorporated during DNA replication, alkylating agents can introduce changes even in nonreplicating DNA. Both base analogs and alkylating agents tend to induce base-pair substitutions (Section 11.2).

Another group of chemical mutagens, the acridines, are planar molecules that function as *intercalating agents*. These mutagens become inserted between two DNA base pairs and push them apart. Then, during replication, this abnormal conformation can trigger single base insertions or deletions. Thus, acridines typically induce frameshift rather than point mutations (Section 11.2). Ethidium bromide, which is commonly used to detect DNA in gel electrophoresis, is also an intercalating agent and therefore a mutagen.

Nonionizing and *ionizing* radiation are two forms of electromagnetic radiation that are highly mutagenic (**Figure 11.8**). Ultraviolet (UV) radiation is widely used to generate mutations as the purine and pyrimidine bases of nucleic acids absorb UV radiation strongly (the absorption maximum for DNA and RNA is at 260 nm). The primary mutagenic effect is the production of *pyrimidine dimers*, in which two adjacent pyrimidine bases (cytosine or thymine) on the same strand of DNA become covalently bonded to one another. This either greatly impedes DNA polymerase activity or greatly increases the probability of DNA polymerase misreading the sequence at this point. Thus the killing of cells by UV radiation is due primarily to its effect on DNA. Conversely, ionizing radiation is more powerful than UV radiation and includes short-wavelength radiation such as X-rays, cosmic rays, and gamma rays

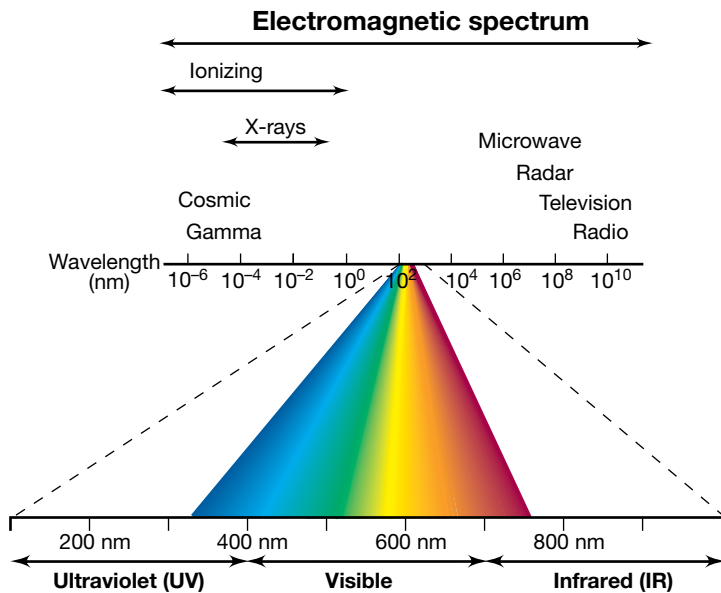


Figure 11.8 Wavelengths of radiation. Ultraviolet radiation consists of wavelengths just shorter than visible light. For any electromagnetic radiation, the shorter the wavelength, the higher the energy. DNA absorbs strongly at 260 nm.

(Figure 11.8). These rays cause water and other substances to ionize, resulting in the formation of free radicals such as the hydroxyl radical ($\text{OH}\cdot$, [↔](#) Section 5.14) that can damage macromolecules in the cell, including DNA. This causes double-stranded and single-stranded breaks that may lead to rearrangements or large deletions.

DNA Repair and the SOS System

By definition, a mutation is a *heritable* change in the genetic material. Therefore, if damaged DNA can be corrected before the cell divides, no mutation will occur. While cells have a variety of different DNA repair processes to correct mistakes ([↔](#) Section 4.4) or repair damage, some are error-prone and the repair process itself introduces the mutation. Some types of DNA damage, especially large-scale damage from highly mutagenic chemicals or large doses of radiation, may cause lesions that interfere with replication. If such lesions are not removed before replication occurs, DNA replication will stall and lethal breaks in the chromosome will result.

In *Bacteria*, stalled replication or major DNA damage activates the **SOS repair system**. The SOS system initiates a number of DNA repair processes, some of which are error-free. However, the SOS system also allows DNA repair to occur without a template, that is, with random incorporation of dNTPs. As might be expected, this results in many errors and hence many mutations. However, mutations may be less detrimental to cell survival than breaks in the chromosome, as mutations can often be corrected but chromosome breaks usually cannot.

In *Escherichia coli* the SOS repair system controls the transcription of approximately 40 genes located throughout the chromosome that participate in DNA damage tolerance and DNA repair (the SOS system thus forms a regulon, [↔](#) Section 6.3). In DNA damage tolerance, DNA lesions remain in the DNA, but are

bypassed by specialized DNA polymerases that can move past DNA damage—a process known as *translesion synthesis*. Even if no template is available to allow insertion of the correct bases, it is less dangerous to cell survival in the long run to fill the gap than to let it remain. Consequently, translesion synthesis generates many errors. In *E. coli*, in which the process of mutagenesis has been studied in great detail, the two error-prone repair polymerases are DNA polymerase V, an enzyme encoded by the *umuCD* genes, and DNA polymerase IV, encoded by *dinB* (Figure 11.9). Both are induced as part of the SOS repair system.

The master regulators of the SOS system are the proteins LexA and RecA. LexA is a repressor that normally prevents expression of the SOS system. The RecA protein, which normally functions in genetic recombination (Section 11.5), is activated by the presence of DNA damage, in particular by the single-stranded DNA that results when replication stalls. The activated form of RecA then stimulates LexA to inactivate itself by self-cleavage. This leads to derepression of the SOS system and the coordinate expression of proteins that participate in DNA repair. Because some of the DNA repair mechanisms of the SOS system—such as DNA polymerases IV and V—are inherently error-prone, many mutations arise. However, once the original DNA damage has been repaired, the SOS regulon is repressed and further mutagenesis ceases.

MINIQUIZ

- How do mutagens cause mutations?
- What is meant by “error-prone” DNA repair?

II • Gene Transfer in *Bacteria*

Comparative genomic analyses of closely related microbes that exhibit different phenotypes have revealed distinct genome differences. Often these idiosyncratic differences result from *horizontal gene transfer*, the movement of genes between cells that are not direct descendants of one another ([↔](#) Section 9.6). Horizontal gene transfer allows cells to quickly acquire new characteristics and fuels metabolic diversity.

Three mechanisms of genetic exchange are known in bacteria: (1) *transformation*, in which free DNA released from one cell is taken up by another (Section 11.6); (2) *transduction*, in which DNA transfer is mediated by a virus (Section 11.7); and (3) *conjugation*, in which DNA transfer requires cell-to-cell contact and a conjugative plasmid in the donor cell (Sections 11.8 and 11.9). These processes are contrasted in Figure 11.10, and it should be noted that DNA transfer typically occurs in only one direction, *from donor to recipient*.

Before discussing the mechanisms of transfer, we consider the fate of transferred DNA. Regardless of how it was transferred, DNA that enters the cell by horizontal gene transfer faces three possible fates: (1) It may be degraded by the recipient cell’s restriction enzymes or other DNA destruction systems (Section 11.12); (2) it may replicate by itself (but only if it possesses its own origin of replication, such as a plasmid or phage genome); or (3) it may recombine with the recipient cell’s chromosome.

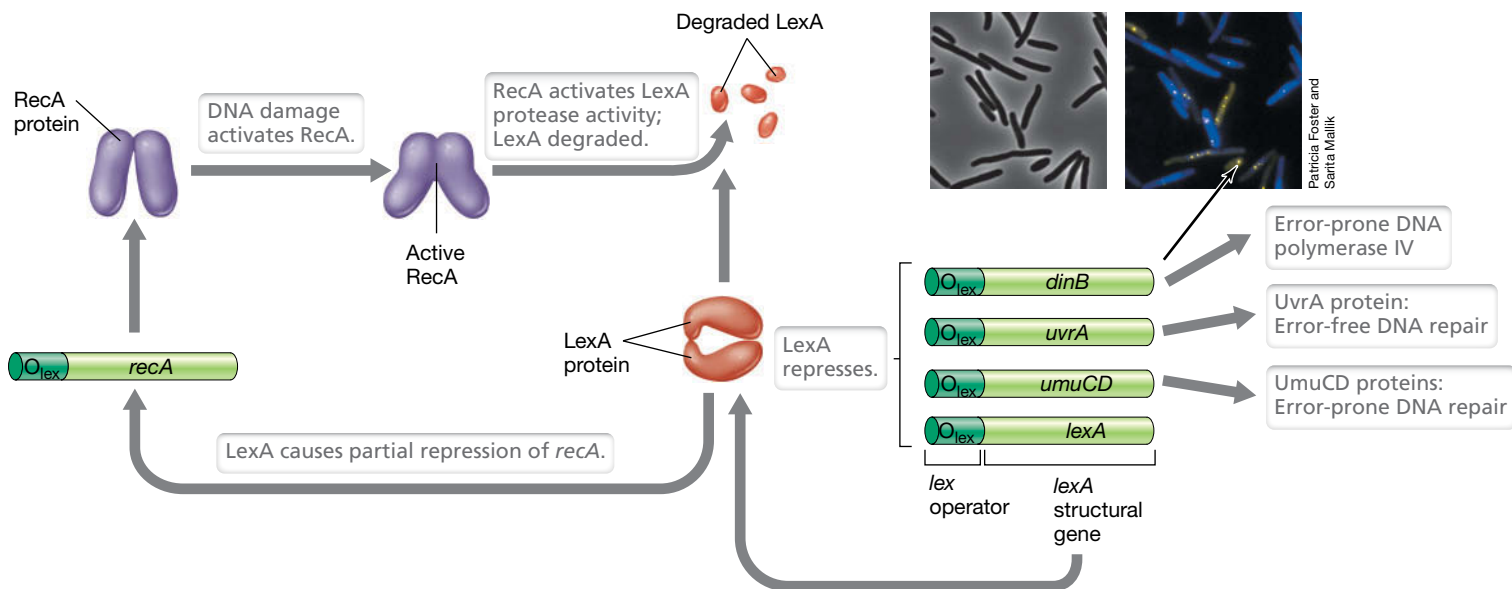


Figure 11.9 SOS response to DNA damage. DNA damage activates RecA protein, which in turn activates the protease activity of LexA, resulting in self-cleavage. LexA normally represses the activities of *recA*, the DNA repair genes *uvrA*, *umuCD* (the UmuCD proteins are part of DNA polymerase V), and *dinB*, which encodes DNA polymerase IV. However, repression is not complete. Some

RecA protein is produced even in the presence of LexA protein. When LexA is inactivated, DNA repair genes become highly active. Inset photos: Both photos show DNA polymerase IV localization to the nucleoid during the SOS response in *Escherichia coli*. Cells containing a fluorescently tagged DNA polymerase IV (DinB) were treated with an antibiotic to induce DNA damage.

Left: phase-contrast micrograph. Right: fluorescence micrograph showing cells stained with DAPI (blue) and DNA polymerase IV (yellow, in the nucleoid region). Expression of *dinB* requires not only the loss of LexA repression but also the protein RpoS, an RNA polymerase sigma factor whose synthesis is triggered by various stress responses.

11.5 Genetic Recombination

Recombination is the physical exchange of DNA between *genetic elements* (structures that carry genetic information). Here we focus on *homologous recombination*, a process that results in genetic exchange between homologous DNA sequences from two different sources. Homologous DNA sequences are those that have nearly the same sequence; therefore, bases can pair over an extended length of the two DNA molecules to facilitate exchange. This type of recombination drives the process of “crossing over” in classical genetics.

Molecular Events in Homologous Recombination

The RecA protein, previously mentioned in regard to the SOS repair system (Section 11.4 and Figure 11.9), is the key to homologous recombination. RecA is essential in nearly every homologous recombination pathway. RecA-like proteins have been identified in all *Bacteria* examined, as well as in the *Archaea* and most *Eukarya*.

A molecular mechanism for homologous recombination between two DNA molecules is shown in **Figure 11.11**. An enzyme that cuts DNA in the middle of a strand, called an *endonuclease*, begins the process by nicking one strand of the donor DNA molecule. This nicked strand is separated from the other strand by proteins with helicase activity; the resulting single-stranded segment binds single-strand binding protein (SSB; Section 4.3) and then RecA. This results in a complex that promotes base pairing with the complementary sequence in the recipient DNA molecule. Base pairing, in turn, displaces the other strand of the recipient DNA molecule (Figure 11.11) and is appropriately called *strand invasion*.

The base pairing of one strand from each of the two DNA molecules over long stretches generates recombination intermediates containing long **heteroduplex** regions, where each strand has originated from a different chromosome. The linked molecules are then resolved (separated) by enzymes that cut and rejoin the previously unbroken strands of both original DNA molecules. Depending on the orientation of the junction during resolution, two types of products, referred to as “patches” or “splices,” are formed that differ in the conformation of the heteroduplex regions remaining after resolution (Figure 11.11).

Effect of Homologous Recombination on Genotype

For homologous recombination to generate new genotypes, the two homologous sequences must be related but genetically distinct. This is obviously the case in a diploid eukaryotic cell, which has two sets of chromosomes, one from each parent. In bacteria, genetically distinct but homologous DNA molecules are brought together in different ways. Genetic recombination in bacteria occurs after fragments of homologous DNA from a donor chromosome are transferred to a recipient cell by transformation, transduction, or conjugation. It is only after the transfer event, when the DNA fragment from the donor is in the recipient cell, that homologous recombination occurs.

For physical exchange of DNA segments to be detected, the cells resulting from recombination must be phenotypically different from both parents (**Figure 11.12**). Genetic crosses in bacteria usually depend on using recipient strains that lack some selectable character that the recombinants will gain. The recipient may be unable to grow on a particular medium or may exhibit a specific phenotype, while the genetic recombinants can grow on a particular

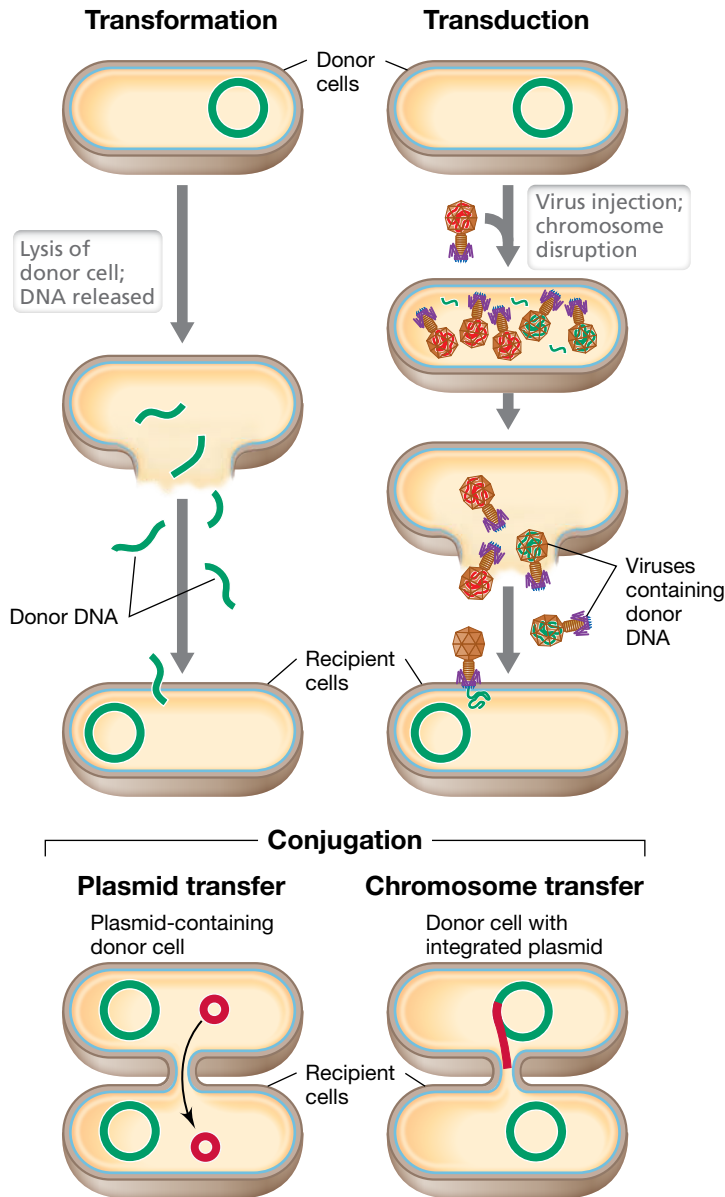


Figure 11.10 Processes by which DNA is transferred from donor to recipient bacterial cell. Just the initial steps in transfer are shown.

medium or exhibit a phenotype different from the recipient (Figures 11.1 and 11.12). Various kinds of selectable markers, such as drug resistance and nutritional requirements, were discussed in Section 11.1. The exceedingly great sensitivity of the selection process allows even a few recombinant cells to be detected in a large population of nonrecombinant cells, and thus selection is an important tool for the microbial geneticist.

Complementation

In all three methods of bacterial gene transfer, only a portion of the donor chromosome enters the recipient cell and thus transfer is just the first step; unless recombination takes place with the recipient chromosome, the donor DNA will be lost because it cannot replicate independently in the recipient. Nonetheless, it is possible to stably maintain a state of partial diploidy for use in

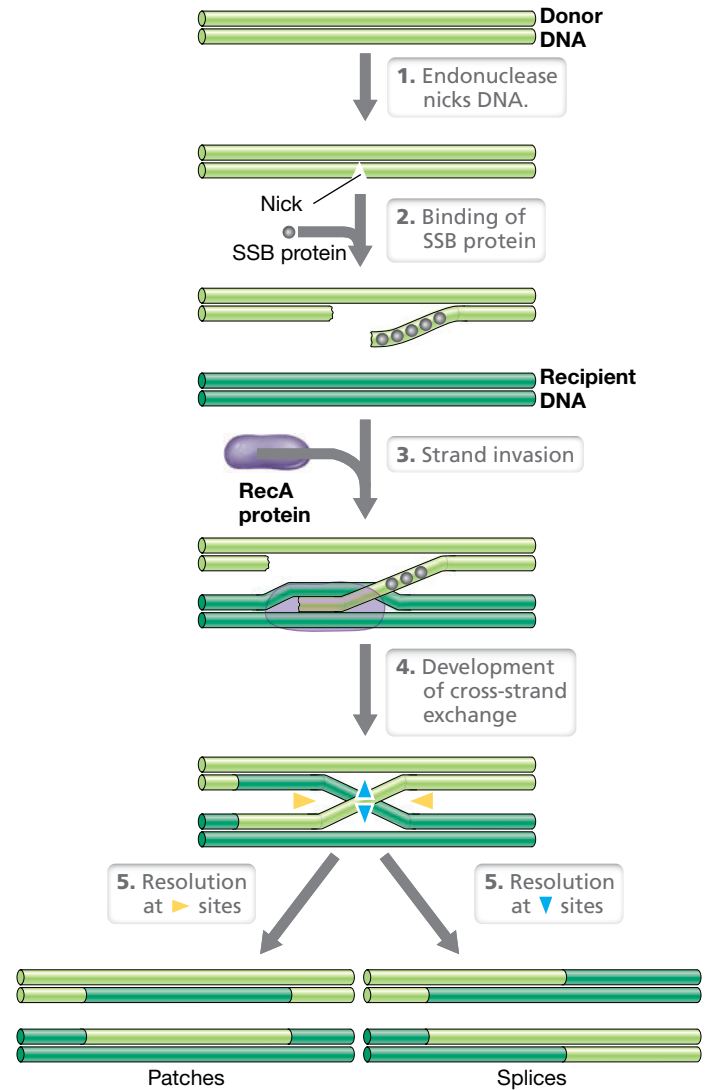


Figure 11.11 A simplified version of homologous recombination. Homologous DNA molecules pair and exchange DNA segments. The mechanism requires breakage and reunion of paired segments. Two of the participating proteins, single-strand binding (SSB) protein and the RecA protein, are shown. The other participating proteins are not shown. The diagram is not to scale: Pairing may occur over hundreds or thousands of bases. Resolution occurs by cutting and rejoining the cross-linked DNA molecules. Note that there are two possible outcomes, patches or splices, depending on where strands are cut during the resolution process.

bacterial genetic analysis. A bacterial strain that carries *two copies* of any particular chromosomal segment is known as a partial diploid, or *merodiploid*. In general, one copy is present on the chromosome itself and the second copy on another genetic element, such as a plasmid or a bacteriophage.

Consequently, if the chromosomal copy of a gene is defective due to a mutation, it is possible to supply a functional (wild-type) copy of the gene on a plasmid or bacteriophage. For example, if one of the genes for biosynthesis of the amino acid tryptophan has a mutation resulting in a nonfunctional enzyme, this will yield a Trp^- phenotype. That is, the mutant strain will be a tryptophan auxotroph and must be supplied with tryptophan for growth. However, if a copy of the wild-type gene is introduced

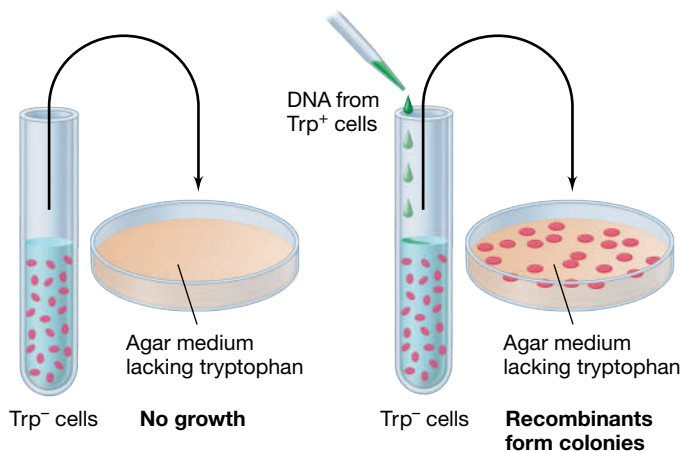


Figure 11.12 Using a selective medium to detect rare genetic recombinants. On the selective medium only the rare recombinants form colonies even though a very large population of bacteria was plated. Procedures such as this, which offer high resolution for genetic analyses, can ordinarily be used only with microorganisms. The type of genetic exchange being illustrated is transformation, but a similar outcome could result from any of the other forms of horizontal gene transfer.

into the same cell on a plasmid or viral genome, this gene will encode the necessary protein and, assuming the gene is transcribed and translated, will restore the wild-type phenotype. This process is called *complementation* because the wild-type gene is said to *complement* the mutation, in this case converting the Trp^- cell into Trp^+ (Figure 11.12).

MINIQUIZ

- Which protein, found in virtually all cells, facilitates the pairing required for homologous recombination?
- Explain the fate of transferred chromosomal DNA if recombination does not occur in the recipient cell.
- What is a merodiploid, and what is genetic complementation?

11.6 Transformation

Transformation is a genetic transfer process by which *free DNA* is incorporated into a recipient cell and brings about genetic change. Several organisms are naturally transformable, including certain species of both gram-negative and gram-positive *Bacteria* and also some species of *Archaea* (Section 11.10). Because the DNA in prokaryotic cells is present as a large single molecule, when a cell is gently lysed, its DNA pours out.

Bacterial chromosomes break easily because of their extreme length (if linearized, the *Bacillus subtilis* chromosome would be 1700 μm long). Even after gentle extraction, the *B. subtilis* chromosome of 4.2 megabase pairs is converted to fragments of about 10 kilobase pairs each. Because an average gene contains about 1000 nucleotides, each of the fragments of *B. subtilis* DNA contains about ten genes. This is a typical transformable size. A single cell typically incorporates only one or at most a few DNA fragments, so only a small proportion of the genes of one cell can be transferred to another in a single transformation event.

Competence in Transformation

A cell that is able to take up DNA and be transformed is said to be *competent*, and this capacity is genetically determined. Competence in most naturally transformable bacteria is regulated, and special proteins play a role in the uptake and processing of DNA. These competence-specific proteins include a membrane-associated DNA-binding protein, a cell wall autolysin, and various nucleases. One pathway of natural competence in *B. subtilis*—an easily transformed species—is regulated by quorum sensing, a regulatory system that responds to cell density (see Section 6.8). Cells produce and excrete a small peptide during growth, and the accumulation of this peptide to high concentrations induces the cells to become competent. But not all cells in a population become competent and stay that way for several hours. By contrast, in *Streptococcus*, 100% of the cells can become competent, but only for a brief period during the growth cycle.

Multiple layers of regulation control natural competence in other bacteria. *Vibrio cholerae* (the causative agent of cholera) is naturally found in marine and freshwater environments associated with crustacean exoskeletons, which are composed of chitin. Competence in *V. cholerae* is controlled not only by quorum sensing but also by chitin sensing and catabolite repression (see Section 6.4; see also page 342). *V. cholerae* can catabolize chitin (a polymer of *N*-acetylglucosamine and an abundant nutrient in the marine environment), and as cells aggregate on a chitin surface, they are in close proximity to one another and more likely to successfully exchange DNA (Figure 11.13).

High-efficiency, natural transformation is rare among *Bacteria*. For example, *Acinetobacter*, *Bacillus*, *Streptococcus*, *Haemophilus*, *Neisseria*, and *Thermus* are naturally competent and easy to transform. This natural competence provides a nutritional advantage, as free DNA is rich in carbon, nitrogen, and phosphorus. By contrast, many *Bacteria* are poorly transformed, if at all, under natural conditions. For example, *Escherichia coli* and many other gram-negative bacteria fall into this category. However, if cells of *E. coli* are treated with high concentrations of Ca^{2+} and then chilled, they become adequately competent. Cells treated in this manner take up double-stranded DNA, and therefore transformation of *E. coli* by plasmid DNA can be relatively efficient. This is important because getting DNA into *E. coli*—the workhorse of genetic engineering—is critical for biotechnology, as we will see in Chapter 12.

Electroporation is a physical technique that is used to get DNA into organisms that are difficult or impossible to transform, especially cells that contain thick cell walls. In electroporation, cells are mixed with DNA and then exposed to brief, high-voltage electrical pulses. This makes the cell envelope permeable and allows entry of the DNA. Electroporation works for getting free DNA into most types of cells, including *E. coli*, most other *Bacteria*, some species of *Archaea*, and even yeast and certain plant cells.

Uptake and Integration of DNA in Transformation

During natural transformation (Figure 11.14), competent bacteria reversibly bind DNA. Soon, however, the binding becomes irreversible. Competent cells bind much more DNA than do noncompetent cells—as much as 1000 times more. As noted earlier, the

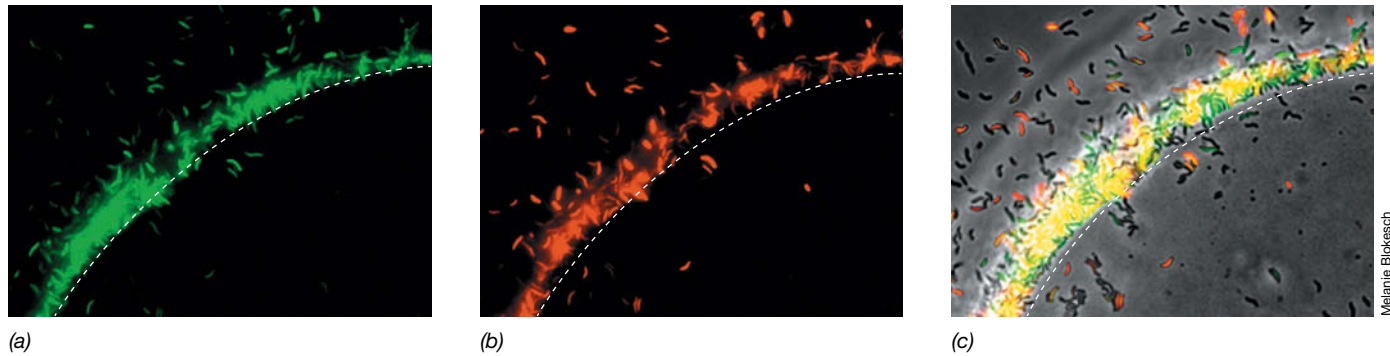


Figure 11.13 Regulation of natural competence in *Vibrio cholerae*. *V. cholerae* cells with fluorescent reporter genes linked to the promoters of competence genes were grown in the presence of chitin beads. White dashed line indicates edge of bead surface. (a) Cells with

the *pilA* (pilus protein) promoter linked to a green fluorescent protein (GFP). (b) Cells with the *comEA* (DNA uptake proteins) promoter linked to a red fluorescent protein. (c) Merged image of parts a and b illustrating expression of competence genes in cells associated with

the chitin bead. Cells of *V. cholerae* are about 0.5 μm wide and 1.5 μm long. Adapted from Lo Scrudato, M., and M. Blokesch. 2012. *PLoS Genetics* 8(6): e1002778. See page 342 for a different view of transformation in *V. cholerae*.

sizes of the transforming fragments are much smaller than that of the whole genome, and the fragments are further degraded during the uptake process. In *Streptococcus pneumoniae* (the cause of bacterial pneumonia) each cell can bind only about ten molecules of double-stranded DNA of 10–15 kbp each. However, as these fragments are taken up, they are converted into single-stranded pieces of about 8 kb, with the complementary strand being degraded. The DNA fragments in the mixture compete with each other for uptake and thus the probability of a transformant taking up DNA that confers an advantage or a selectable marker decreases.

During transformation, DNA is bound at the cell surface by a DNA-binding protein (Figure 11.14). In many species, this DNA-binding protein resembles a pilus (↔ Section 2.7) that is able to pull the DNA into the periplasm of a gram-negative bacterium or through the thick cell wall of a gram-positive bacterium. Next, either the entire double-stranded fragment is taken up, or a nuclease degrades one strand and the remaining strand is taken up, depending on the organism. After uptake, a competence-specific protein binds the donor DNA. This protects the DNA from nuclease attack until it reaches the recipient's chromosome, where the

RecA protein takes over. The DNA is integrated into the genome of the recipient by recombination (Figures 11.11 and 11.14). The preceding applies only to small pieces of *linear* DNA. Many naturally transformable *Bacteria* are transformed only poorly by plasmid DNA because the plasmid must remain double-stranded and circular in order to replicate.

MINIQUIZ

- During transformation a cell usually incorporates only one or a few fragments of DNA. Explain.
- In genetic transformation, what is meant by the word competence?

11.7 Transduction

In **transduction**, a bacterial virus (bacteriophage) transfers DNA from one cell to another. Viruses can transfer host genes in two ways. In the first, called *generalized transduction*, DNA

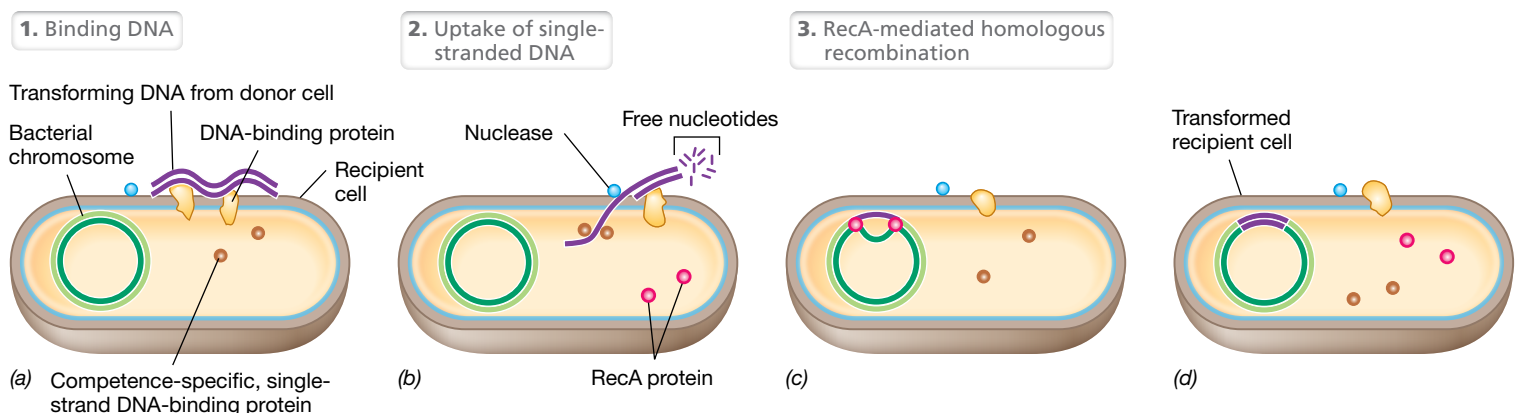


Figure 11.14 Mechanism of transformation in a gram-positive bacterium. (a) Binding of double-stranded DNA by a membrane-bound DNA-binding protein. (b) Passage of one of the two strands into the cell while nuclease activity degrades the other strand. (c) The single strand of DNA in the cell is bound by specific proteins, and recombination with homologous regions of the bacterial chromosome is mediated by RecA protein. (d) Transformed cell.

derived from virtually any portion of the host genome is packaged inside the mature virion in place of the virus genome. In the second, called *specialized transduction*, DNA from a specific region of the host chromosome is integrated directly into the virus genome—usually replacing some of the virus genes. This occurs only with certain temperate viruses such as phage lambda (↔ Section 8.7).

In generalized transduction, the bacterial donor genes cannot replicate independently and are not part of a viral genome. Thus, unless the donor genes recombine with the recipient bacterial chromosome, they will be lost. In specialized transduction, homologous recombination may also occur. However, since the donor bacterial DNA is actually a part of a temperate phage genome, it may be integrated into the host chromosome during lysogeny (↔ Section 8.7).

Transduction occurs in a variety of *Bacteria*, including the genera *Desulfovibrio*, *Escherichia*, *Pseudomonas*, *Rhodococcus*, *Rhodobacter*, *Salmonella*, *Staphylococcus*, and *Xanthobacter*, as well as *Methanothermobacter thermautotrophicus*, a species of *Archaea*. Not all phages can transduce and not all bacteria are transducible, but with bacteriophages estimated to outnumber prokaryotic cells in nature by 10-fold, transduction likely plays an important role in gene transfer in the environment. Some examples of genes transferred by transducing bacteriophages include multiple-antibiotic-resistance genes among strains of *Salmonella enterica* (*typhimurium*), Shiga-like toxin genes in *Escherichia coli*, virulence factors in *Vibrio cholerae*, and genes encoding photosynthetic proteins in cyanobacteria (↔ Section 10.12).

Generalized Transduction

In generalized transduction, virtually any gene on the donor chromosome can be transferred to the recipient, forming a *transducant*. Generalized transduction was first discovered and extensively studied in the bacterium *S. enterica* with phage P22 and has also

been studied with phage P1 in *E. coli*. The mechanism of transduction is shown in **Figure 11.15**. When a bacterial cell is infected with a transducing phage, the lytic cycle may occur. However, during lytic infection, the enzymes responsible for packaging viral DNA into the bacteriophage sometimes package host DNA accidentally. The result is called a *transducing particle*. These cannot lead to a viral lytic infection because they contain no viral DNA, and are therefore said to be *defective*.

Upon lysis of the cell, transducing particles are released along with normal virions that contain the virus genome. When this lysate is used to infect a population of recipient cells, most of the cells are infected with a normal (lytic) virus. However, a small proportion of the population receives transducing particles that inject the DNA they packaged from the previous host bacterium. Although this DNA cannot replicate, it can recombine with the DNA (Section 11.5) of the new host (**Figure 11.16**). Because only a small proportion of the particles in the lysate are defective, and each of these contains only a small fragment of donor DNA, the probability of a given transducing particle containing a particular gene is quite low. Typically, only about 1 cell in 10^6 to 10^8 cells is transduced for any given gene.

Lysogeny and Specialized Transduction

Generalized transduction allows the transfer of any gene from one bacterium to another, but at a low frequency. In contrast, specialized transduction allows extremely efficient transfer but is selective and transfers only a small region of the bacterial chromosome. In the first case of specialized transduction to be discovered, genes for galactose catabolism were transduced by the temperate phage lambda of *E. coli*.

When lambda lysogenizes a host cell, the phage genome is integrated into the *E. coli* chromosome at a specific site (↔ Section 8.7). This site is next to the cluster of genes that encode the enzymes for galactose utilization. After insertion, viral DNA replication

Lytic cycle

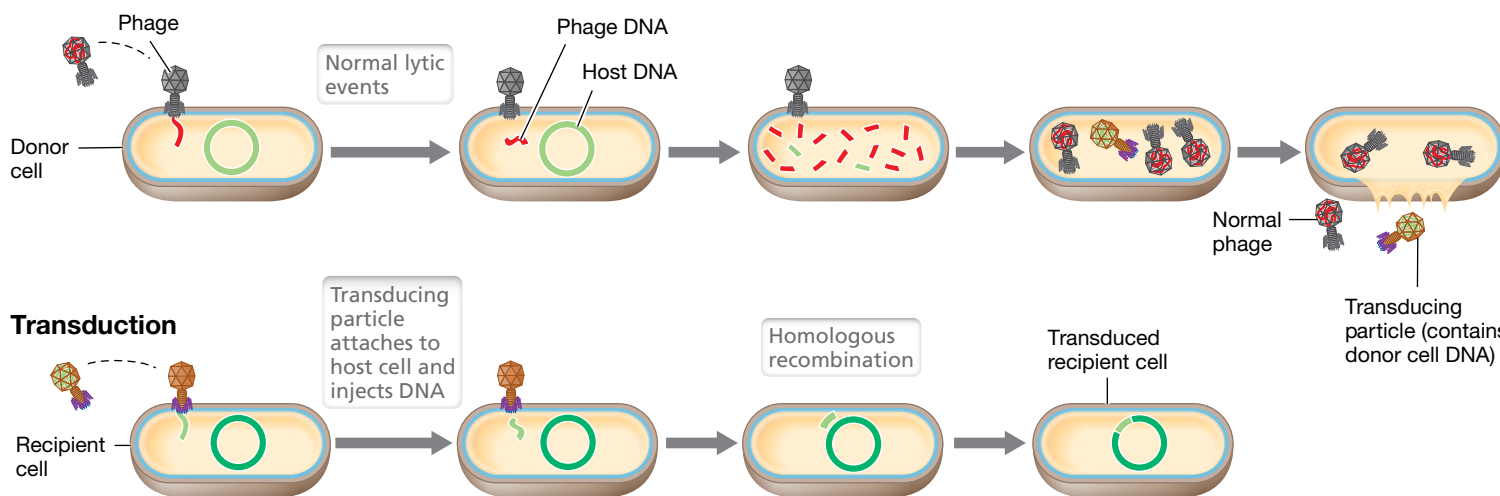


Figure 11.15 Generalized transduction. Note that “normal” virions contain phage genes, whereas a transducing particle contains host genes.

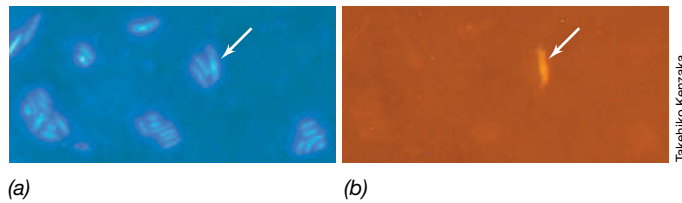


Figure 11.16 Visualization of generalized transduction. *Citrobacter freundii* cells were mixed with P1 bacteriophage carrying the β -lactamase (*bla*) gene for 10 minutes. (a) Fluorescence micrograph showing cells by DAPI staining. (b) Detection of a single *C. freundii* transductant containing the *bla* gene recombined into the genome using cycling primed in situ amplification–FISH, a modified version of fluorescence in situ hybridization (FISH, [↗](#) Section 19.5). Arrow indicates the cell that is transduced.

is under control of the bacterial host chromosome. Upon induction, the viral DNA separates from the host DNA by a process that is the reverse of integration (Figure 11.17). Usually the lambda DNA is excised precisely, but occasionally the phage genome is excised incorrectly. Some of the adjacent bacterial genes to one side of the prophage (for example, the galactose operon) are excised along with phage DNA. At the same time, some phage genes are left behind (Figure 11.17b). This transducing particle can subsequently transfer genes for galactose utilization to a recipient cell. This transfer can only be detected if a galactose-negative (Gal^-) strain is infected with such a transducing particle and Gal^+ transductants are selected.

For a lambda virion to be infectious, there is a limit to the amount of phage DNA that can be replaced with host DNA. Sufficient phage DNA must be retained to encode the phage protein coat and other phage proteins needed for lysis and lysogeny. However, if a helper phage is used together with a defective phage in a *mixed infection*, then far fewer phage-specific genes are needed in the defective phage. Only the *att* (attachment) region, the *cos* site (cohesive ends, for packaging), and the replication origin of the lambda genome ([↗](#) Figure 8.17b) are necessary.

Phage Conversion

Alteration of the phenotype of a host cell by lysogenization is called *phage conversion*. When a normal (that is, nondefective) temperate phage lysogenizes a cell and becomes a prophage, the cell becomes immune to further infection by the same type of phage. Such immunity may itself be regarded as a change in phenotype. However, other phenotypic changes unrelated to phage immunity are often observed in phage conversion of lysogenized cells.

Two cases of phage conversion have been especially well studied. One results in a change in structure of a polysaccharide on the cell surface of *S. enterica* (*anatum*) upon lysogenization with bacteriophage ϵ^{15} . The second results in the conversion of non-toxin-producing strains of *Corynebacterium diphtheriae* (the bacterium that causes the disease diphtheria) to toxin-producing (pathogenic) strains following lysogeny with bacteriophage β ([↗](#) Section 30.3). In both cases, the genes responsible for the changes are an integral part of the phage genome and hence are automatically transferred to the cell upon phage infection and establishment of the lysogenic state.

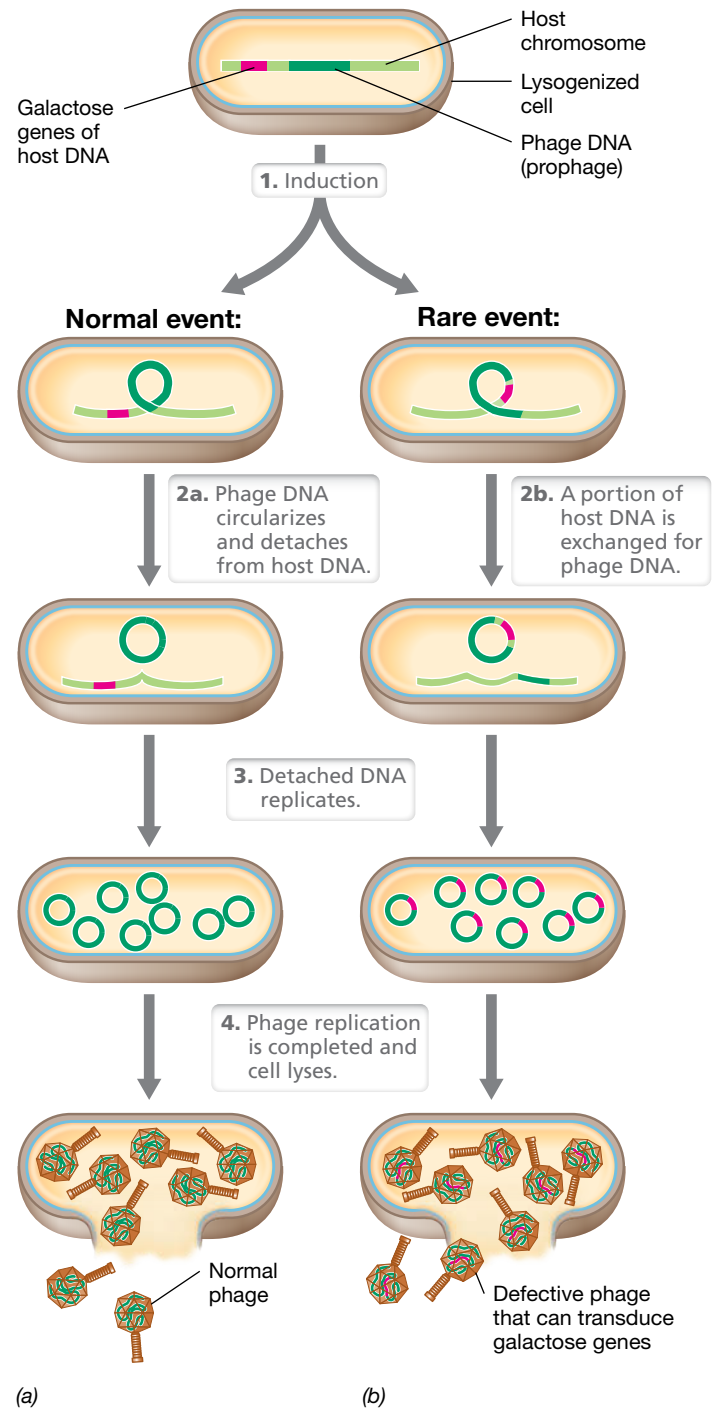
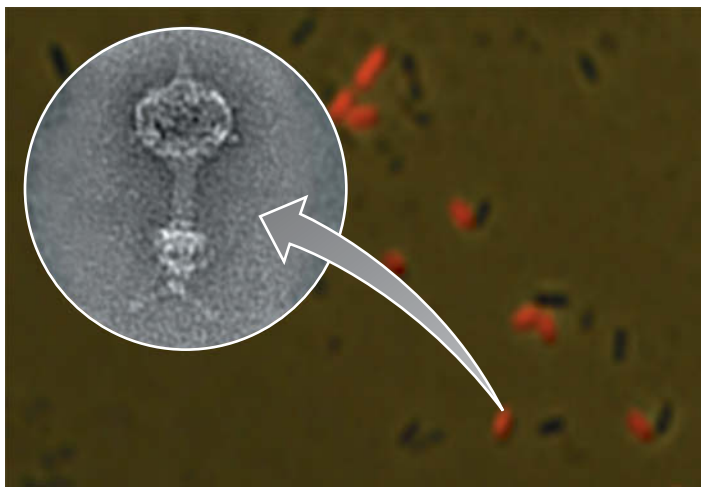


Figure 11.17 Specialized transduction. In an *Escherichia coli* cell containing a lambda prophage, (a) normal lytic events and (b) the production of particles transducing the galactose genes. Only a short region of the circular host chromosome is shown in the figure.

Lysogeny likely carries strong selective value for the host cell because it confers resistance to infection by viruses of the same type. Phage conversion may also be of considerable evolutionary significance because it results in genetic alteration of host cells. It has been found that many bacteria isolated from nature are natural lysogens, and thus it is likely that lysogeny is essential for survival of many host cells in nature.



A. Westbye, P. Fogg, C. Yip, and J.T. Beatty

Figure 11.18 Gene transfer agents. Inset: Transmission electron micrograph of a gene transfer agent (GTA) isolated from *Rhodospirillum rubrum*. Visualization of a subset of *R. capsulatus* cells producing and releasing GTAs during stationary phase using a red fluorescent reporter gene linked to the promoter of an *R. capsulatus* gene essential for GTA production.

Gene Transfer Agents

DNA can also be transferred between prokaryotic cells by defective bacteriophages. These so-called *gene transfer agents* (GTAs) are the result of prokaryotic cells hijacking defective viruses and using them specifically for DNA exchange (Figure 11.18). GTAs resemble tiny tailed bacteriophages and contain random small pieces of *host* DNA. They are not considered true viruses because they do not contain genes encoding their own production and do not produce characteristic viral plaques. Instead, the genes encoding GTAs lie within the genome of the cell that produces them.

GTAs have been isolated from a wide range of *Bacteria* including the sulfate-reducing bacterium *Desulfovibrio desulfuricans* and a variety of anoxygenic purple bacteria and other *Alphaproteobacteria*, and also from certain methanogenic *Archaea*. GTAs seem to be particularly widely produced by marine *Bacteria*, especially purple bacteria such as *Roseovarius*. Exactly what triggers GTA synthesis has just begun to be studied and may vary in different species. However, by linking the promoter of a gene essential for GTA production to a reporter gene, microbial geneticists have determined that a subpopulation of cells in a culture of the anoxygenic phototrophic bacterium *Rhodospirillum rubrum* produce and release GTAs during stationary phase and nutrient fluctuations (Figure 11.18). This suggests that GTAs may have evolved as a mechanism for a cell to disperse its genes into the environment in a protected form before cell lysis released free DNA that could be quickly degraded.

While bacteriophages are considered the most abundant microbes on Earth, the percentage of these that are actually GTAs instead of viruses is unknown but could be significant. The fact that GTAs are produced by so many different species, do not result in cell lysis, and can transfer genes between bacteria that are taxonomically distinct, points to GTAs as major vehicles for gene flow between prokaryotic cells in nature. This could be especially true of open-ocean microbial communities, where

constant low nutrient levels might trigger GTA production as a means for cells to scavenge each other's genes for improving fitness and survival.

MINIQUIZ

- How does a transducing particle differ from an infectious bacteriophage?
- What is the major difference between generalized transduction and transformation?
- Why is phage conversion considered beneficial to host cells?

11.8 Conjugation

Conjugation is a form of horizontal gene transfer in both gram-negative and gram-positive bacteria that requires cell-to-cell contact (mating). Conjugation is a plasmid-encoded mechanism that can mediate DNA transfer between closely related cells or between more distantly related cells; for example, between cells of different genera. Conjugative plasmids use this mechanism to transfer copies of themselves and the genes they encode, such as those for antibiotic resistance, to new host cells.

The process of conjugation requires a *donor* cell, which contains the conjugative plasmid, and a *recipient* cell, which does not. In addition, genetic elements that cannot transfer themselves can sometimes be *mobilized* or transferred during conjugation. These other genetic elements can be other plasmids or the host chromosome itself. Indeed, conjugation was discovered because the F plasmid of *Escherichia coli* can mobilize the host chromosome (see Figure 11.24). Transfer mechanisms may differ depending on the participating plasmid, but most plasmids in gram-negative *Bacteria* employ a mechanism similar to that used by the F plasmid.

F Plasmid

The F plasmid (F stands for “fertility”) is a circular DNA molecule of 99,159 bp. Figure 11.19 shows a genetic map of the F plasmid. One region of the plasmid contains genes that regulate DNA replication. It also contains a number of transposable elements (Section 11.11) that allow the plasmid to integrate into the host chromosome. In addition, the F plasmid has a large region of DNA, the *tra* region, containing genes that encode transfer functions. Many genes in the *tra* region participate in mating pair formation, and most of these have to do with the synthesis of the sex pilus (see Section 2.7) and a type IV secretion system (see Section 4.13) to transfer the DNA. Only donor cells produce these pili. Different conjugative plasmids may have slightly different *tra* regions, and the pili may vary somewhat in structure. The F plasmid and its relatives encode F pili.

Pili allow specific pairing to take place between the donor and recipient cells. All conjugation in gram-negative *Bacteria* is thought to depend on cell pairing brought about by pili. The pilus makes specific contact with a receptor on the recipient cell and then is retracted by disassembling its subunits. This pulls the two cells together (Figure 11.20a). Following this process, donor and recipient cells remain in contact by binding coupling proteins located in the outer membrane of each cell (Figure 11.20b). DNA is then transferred from donor to recipient cell through this conjugation junction (Figure 11.20c).

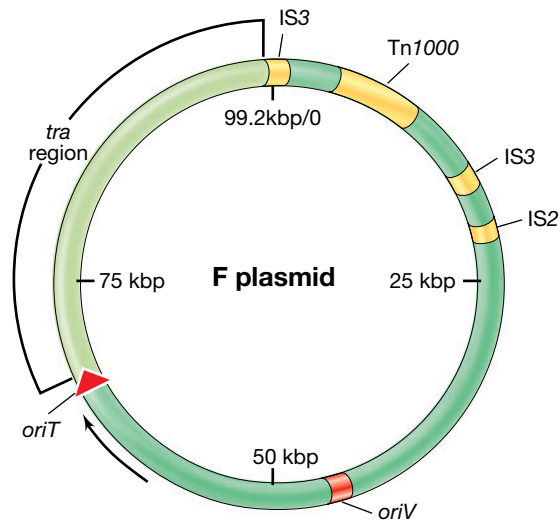
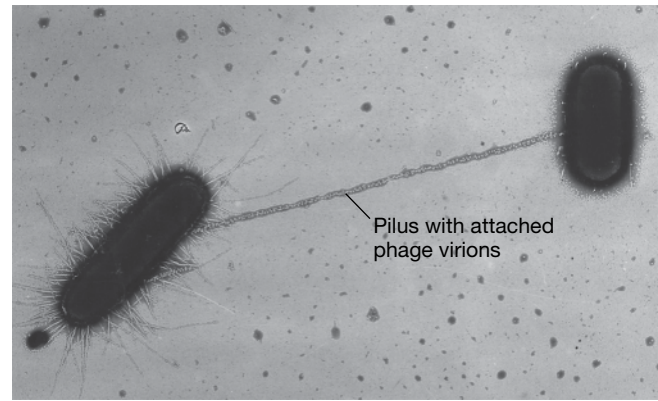


Figure 11.19 Genetic map of the F (fertility) plasmid of *Escherichia coli*. The numbers on the interior show the size in kilobase pairs (the exact size is 99,159 bp). The region in dark green at the bottom of the map contains genes primarily responsible for the replication and segregation of the F plasmid. The origin of vegetative replication is *oriV*. The light green *tra* region contains the genes needed for conjugative transfer. The origin of transfer during conjugation is *oriT*. The arrow indicates the direction of transfer (the *tra* region is transferred last). Insertion sequences are shown in yellow. These may recombine with identical elements on the bacterial chromosome, which leads to integration and the formation of different Hfr strains (Section 11.9).

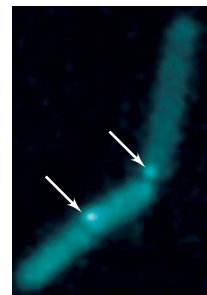
Mechanism of DNA Transfer during Conjugation

DNA synthesis is necessary for DNA transfer by conjugation. This DNA is synthesized not by normal bidirectional replication (↔ Section 4.4) but by **rolling circle replication**, a mechanism also used by some DNA viruses (↔ Section 8.7) and shown in Figure 11.21. DNA transfer is triggered by cell-to-cell contact, at which time one strand of the circular plasmid DNA is nicked and is transferred to the recipient. The nicking enzyme required to initiate the process, TraI, is encoded by the *tra* operon of the F plasmid. TraI also has helicase activity and thus also unwinds the strand to be transferred. As this transfer occurs, DNA synthesis by the rolling circle mechanism replaces the transferred strand in the donor, while a complementary DNA strand is being made in the recipient. Therefore, at the end of the process, both donor and recipient possess complete plasmids. For transfer of the F plasmid, if an F-containing donor cell, which is designated F^+ , mates with a recipient cell lacking the plasmid (F^-), the result is two F^+ cells (Figure 11.21).

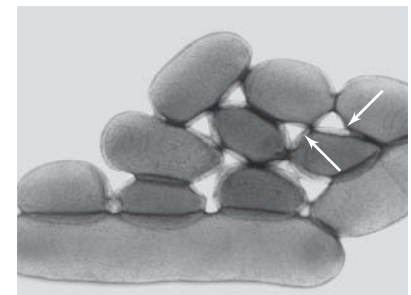
Transfer of plasmid DNA is efficient and rapid; under favorable conditions virtually every recipient cell that pairs with a donor acquires a plasmid. Transfer of the F plasmid (approximately 100 kbp) takes about 5 min. If plasmid genes can be expressed in the recipient, the recipient itself becomes a donor and can transfer the plasmid to other recipients. In this fashion, conjugative plasmids can spread rapidly among bacterial populations, behaving much like infectious agents. This is of major ecological significance because conjugative plasmids have been found in many *Bacteria* and some *Archaea* (Section 11.10), and a few plasmid-containing cells introduced into a population of potential recipients can convert the entire population into plasmid-bearing (and thus donating) cells in a short time.



(a)



(b)



(c)

Figure 11.20 Visualization of conjugation. (a) Formation of a mating pair. Direct contact between two conjugating *Escherichia coli* cells is first made via a pilus. The cells are then drawn together to form a mating pair by retraction of the pilus, which is achieved by depolymerization. Certain small phages (F-specific bacteriophages) use the sex pilus as a receptor and can be seen here attached to the pilus. (b) Coupling proteins near the cell membrane. These *Bacillus subtilis* cells contain the conjugative plasmid pL520 encoding the VirD coupling protein linked to a fluorescent reporter gene. (c) Conjugation junctions. Negatively stained transmission electron micrograph of conjugation bridges between cells of *Yersinia pseudotuberculosis*. Arrows indicate connection sites. Adapted from Lesic, B., M. Zouine, M. Ducos-Galand, C. Huon, M.-L. Rosso, M.-C. Prévost, D. Mazel, and E. Carniel. 2012 *PLoS Genetics* 8(3): e1002529.

MINIQUIZ

- In conjugation, how are donor and recipient cells brought into contact with each other?
- Explain how rolling circle DNA replication allows both donor and recipient to end up with a complete copy of plasmids transferred by conjugation.

11.9 The Formation of Hfr Strains and Chromosome Mobilization

Chromosomal genes can be transferred by plasmid-mediated conjugation. As mentioned above, the F plasmid of *Escherichia coli* can, under certain circumstances, mobilize the chromosome for transfer during cell-to-cell contact. The F plasmid is actually an *episome*, a plasmid that can integrate into the host chromosome. When the F plasmid is integrated, chromosomal genes can be transferred along with the plasmid. Following genetic recombination between donor and recipient DNA, horizontal transfer of chromosomal genes by this mechanism can be extensive.

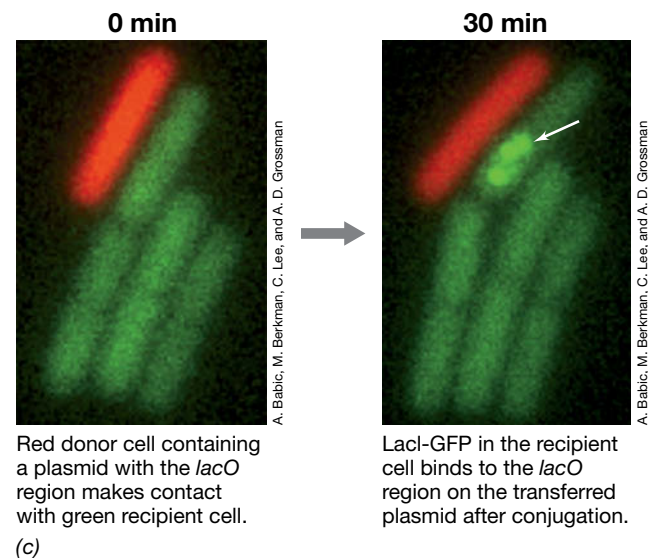
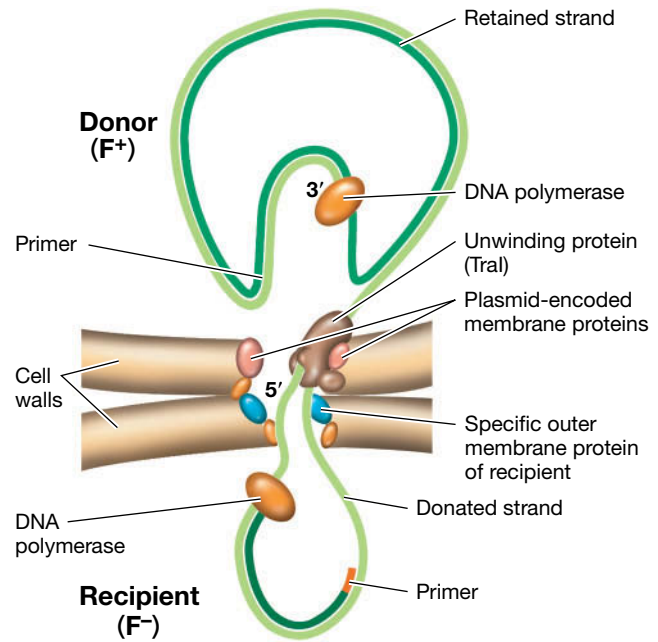
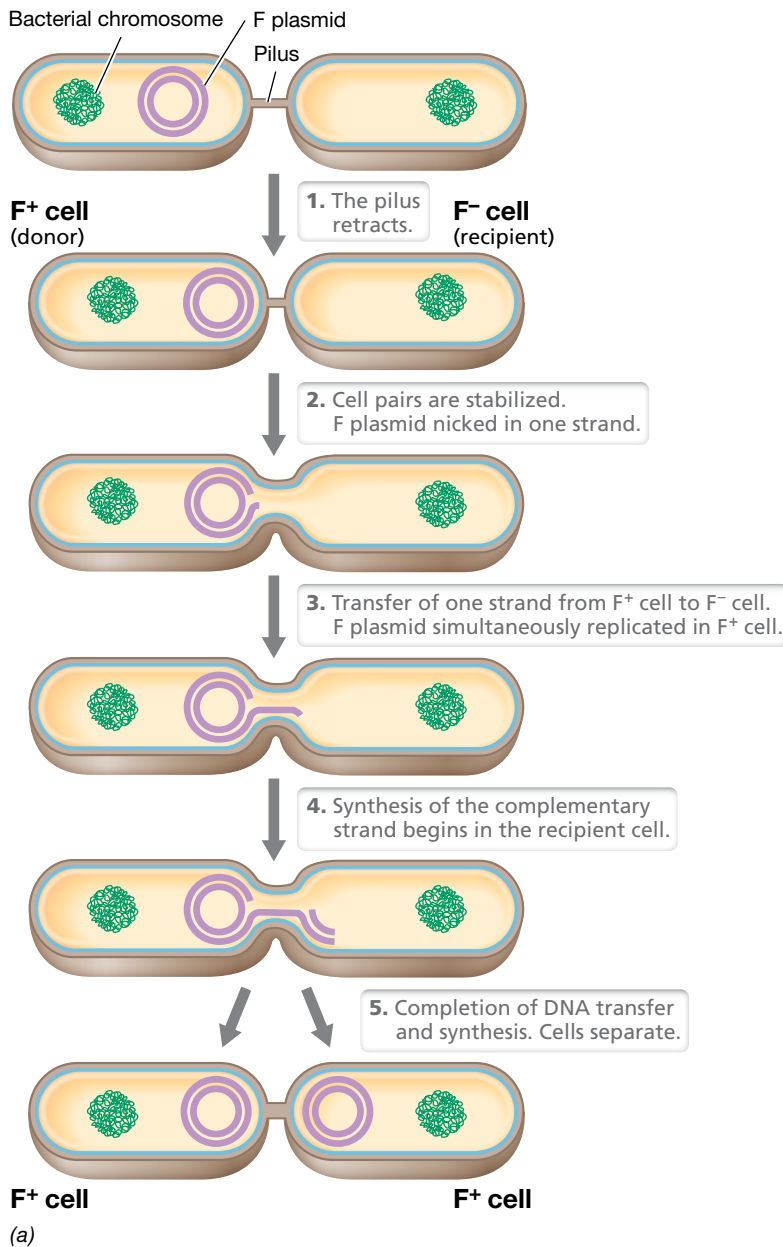


Figure 11.21 Transfer of plasmid DNA by conjugation. (a) The transfer of the F plasmid converts an F⁻ recipient cell into an F⁺ cell. Note the mechanism of rolling circle replication. (b) Details of the replication and transfer process. Note the large number of proteins

needed for successful DNA transfer. (c) Visualization of DNA transfer by conjugation in *Bacillus subtilis* using fluorescence microscopy. The donor cell constitutively expresses a red fluorescent protein, while the recipient cells fluoresce green due to green fluorescent protein

(GFP) fused to LacI (Figure 6.14). The DNA transferred from the donor contains a *lacO* operator region that binds LacI-GFP. Arrow indicates focal point in the recipient cell where LacI-GFP is bound to the *lacO* region obtained from conjugation.

Cells possessing a nonintegrated F plasmid are called F⁺, whereas those with an F plasmid integrated into the chromosome are called **Hfr cells** (for high frequency of recombination). This term refers to the high rates of genetic recombination between genes on the donor (Hfr) and recipient (F⁻) chromosomes. Both F⁺ and Hfr cells are donors, but unlike conjugation between an F⁺ and an F⁻, conjugation between an Hfr donor and an F⁻ leads to transfer of genes from the host chromosome. This is because the chromosome and plasmid now form a single molecule of DNA. Consequently, when rolling circle replication is initiated by the F plasmid, replication

continues on into the chromosome. Thus, the chromosome is also replicated and parts of it get transferred. Hence, integration of a conjugative plasmid provides a mechanism for mobilizing a cell's genome.

Overall, the presence of the F plasmid results in three distinct changes in a cell: (1) the ability to synthesize the F pilus (Figure 11.20a), (2) the mobilization of DNA for transfer to another cell, and (3) the alteration of surface receptors so the cell can no longer be a recipient in conjugation and is therefore unable to take up a second copy of the F plasmid or any genetically related plasmids.

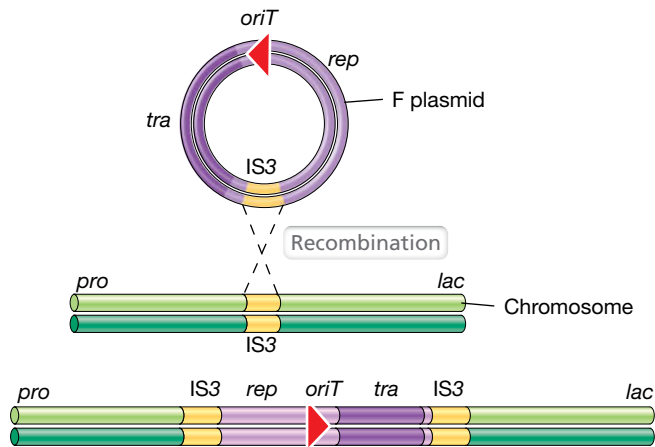


Figure 11.22 The formation of an Hfr strain. Integration of the F plasmid into the chromosome may occur at a variety of specific sites where IS elements are located. The example shown here is an IS3 located between the chromosomal genes *pro* and *lac*. Some of the genes on the F plasmid are shown. The arrow indicates the origin of transfer, *oriT*, with the arrow as the leading end. Thus, in this Hfr, *pro* would be the first chromosomal gene to be transferred and *lac* would be among the last.

Integration of F Plasmid and Chromosome Mobilization

The F plasmid and the chromosome of *E. coli* both carry several copies of mobile genetic elements called *insertion sequences* (IS; Section 11.11). These provide regions of sequence homology between chromosomal and F plasmid DNA. Consequently, homologous recombination between an IS on the F plasmid and a corresponding IS on the chromosome results in integration of the F plasmid into the host chromosome (Figure 11.22). Once integrated, the plasmid no longer replicates independently, but the *tra* operon still functions normally and the strain synthesizes pili. When a recipient cell is encountered, conjugation is triggered just

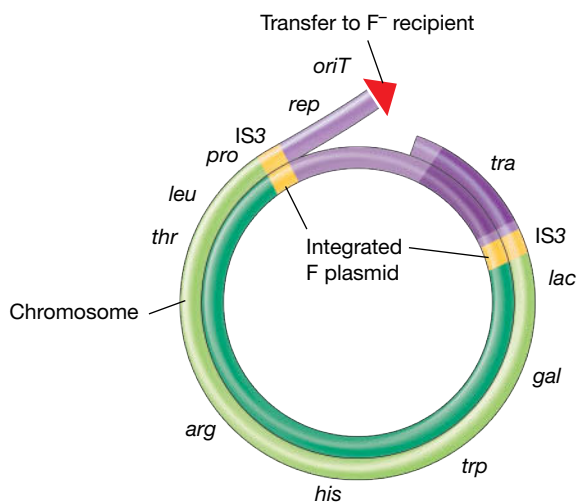


Figure 11.23 Transfer of chromosomal genes by an Hfr strain. The Hfr chromosome breaks at the origin of transfer within the integrated F plasmid. The transfer of DNA to the recipient begins at this point. DNA replicates during transfer as for a free F plasmid (Figure 11.21). This figure is not to scale; the inserted F plasmid is actually less than 3% of the size of the *Escherichia coli* chromosome.

as in an F^+ cell, and DNA transfer is initiated at the *oriT* (origin of transfer) site. However, because the plasmid is now part of the chromosome, after part of the plasmid DNA is transferred, chromosomal genes begin to be transferred (Figure 11.23). As in the case of conjugation with just the F plasmid itself (Figure 11.21), chromosomal transfer also requires DNA replication.

Because the DNA strand usually breaks during transfer, only part of the donor chromosome is typically transferred. Consequently, the recipient does not become Hfr (or F^+) because only part of the integrated F plasmid is transferred (Figure 11.24). However, after transfer, the Hfr strain remains genetically Hfr because it retains a copy of the integrated F plasmid. Following recombination, the recipient cell may express a new phenotype due to the incorporation of donor genes, but genetically it remains an F^- cell. Because a partial chromosome cannot replicate, for incoming donor DNA to survive, it must recombine with the recipient chromosome. As in transformation and transduction, genetic recombination between donor and recipient genes requires homologous recombination in the recipient cell.

Because several distinct insertion sequences are present on the *E. coli* chromosome, a number of different Hfr strains are possible. A given Hfr strain always donates genes in the same order, beginning at the same position. However, Hfr strains that differ in the

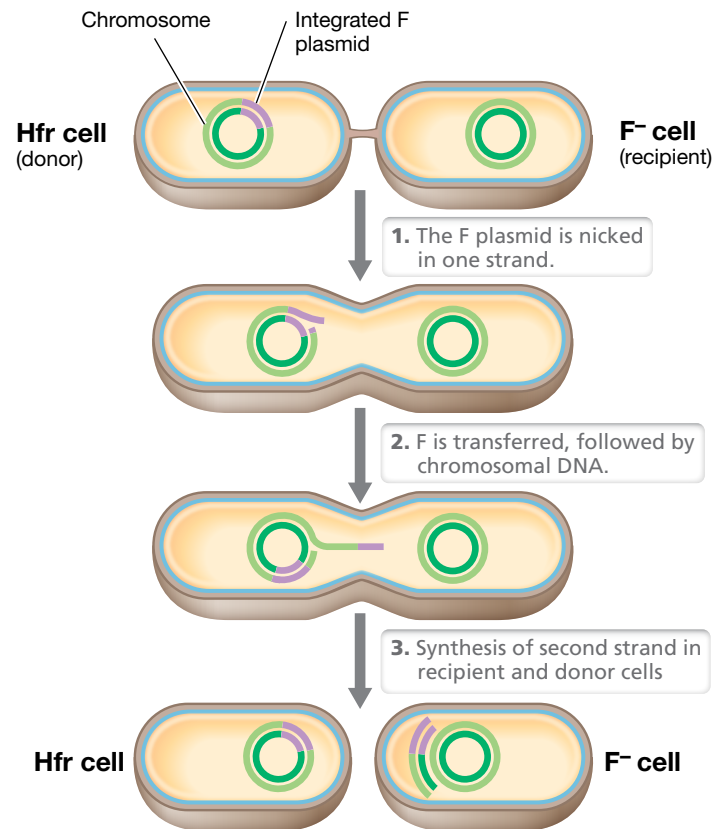


Figure 11.24 Transfer of chromosomal DNA by conjugation. Transfer of the integrated F plasmid from an Hfr strain results in the cotransfer of chromosomal DNA because it is linked to the plasmid. The steps in transfer are similar to those in Figure 11.21a. However, the recipient remains F^- and receives a linear fragment of donor chromosome attached to part of the F plasmid. For donor DNA to survive, it must be recombined into the recipient chromosome after transfer (not shown).

chromosomal integration site of the F plasmid transfer their genes in different orders (Figure 11.25). At some insertion sites, the F plasmid is integrated with its origin pointing in one direction, whereas at other sites the origin points in the opposite direction. The orientation of the F plasmid determines which chromosomal genes enter the recipient cell first and illustrate how virtually any gene on the chromosome can be mobilized by one Hfr configuration or another (Figure 11.25).

Transfer of Chromosomal Genes to the F Plasmid

Occasionally, integrated F plasmids may be excised from the chromosome. During excision, chromosomal genes may sometimes be incorporated into the liberated F plasmid. This can happen because both the F plasmid and the chromosome contain multiple identical insertion sequences where recombination can occur (Figure 11.22). F plasmids containing chromosomal genes are called *F'* (F-prime) *plasmids*. When *F'* plasmids promote conjugation, they transfer the chromosomal genes they carry at high frequency to the recipients. *F'*-mediated transfer resembles specialized transduction (Section 11.7) in that only a restricted group of chromosomal genes is transferred by any given *F'* plasmid. Transferring a known *F'* into a recipient allows one to establish diploids (two copies of each gene) for a limited region of the chromosome. Such partial diploids (merodiploids) are important for genetic complementation tests (Section 11.5).

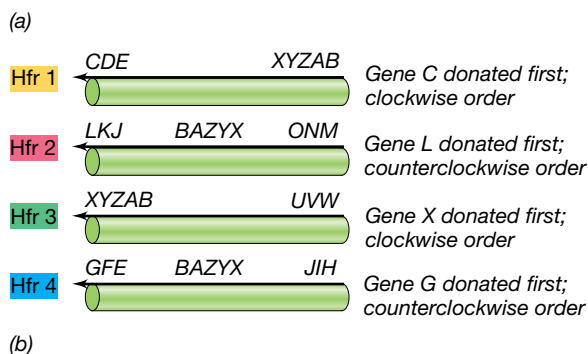
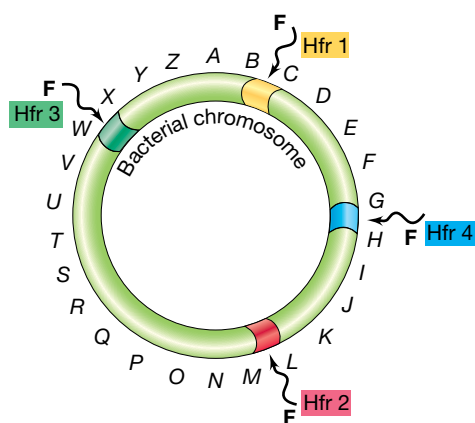


Figure 11.25 Formation of different Hfr strains. Different Hfr strains donate genes in different orders and from different origins. (a) F plasmids can be inserted into various insertion sequences on the bacterial chromosome, forming different Hfr strains. (b) Order of gene transfer for different Hfr strains. See Figures 11.22–11.24 for details of Hfr formation and DNA transfer.

MINIQUIZ

- In conjugation involving the F plasmid of *Escherichia coli*, how is the host chromosome mobilized?
- Why does an Hfr × *F*⁻ mating not yield two Hfr cells?
- At which sites in the chromosome can the F plasmid integrate?

III • Gene Transfer in Archaea and Other Genetic Events

Although studies of the genetics of *Archaea* lag behind genetics research in *Bacteria*, they are showing progress, along with archaeal versions of the tools necessary for detailed genetic analyses. In addition, some other genetic events in *Bacteria* reveal important genetic concepts even though they do not involve horizontal gene flow per se. We cover both of these topics here.

11.10 Horizontal Gene Transfer in Archaea

Although *Archaea* contain a single circular chromosome (Figure 11.26), as do most *Bacteria*, and genome analysis clearly shows that horizontal transfer of archaeal DNA also occurs in nature, laboratory-based gene transfer systems are not as well developed as those for *Bacteria*. Some problems here are of a practical nature including the fact that most well-studied *Archaea* are extremophiles, capable of growth only under conditions of high salt or high temperature (Chapter 17). The temperatures necessary to culture some hyperthermophiles, for example, will melt agar, and alternative materials are required to form solid media and obtain colonies.

Another problem is that most known antibiotics do not affect *Archaea*. For example, penicillins do not affect *Archaea* because their cell walls lack peptidoglycan. The choice of selectable markers for use in genetic crosses is therefore often limited. However, novobiocin (a DNA gyrase inhibitor) and mevinolin (an inhibitor of isoprenoid biosynthesis) have been used to inhibit growth of extreme halophiles, and puromycin and neomycin (both protein

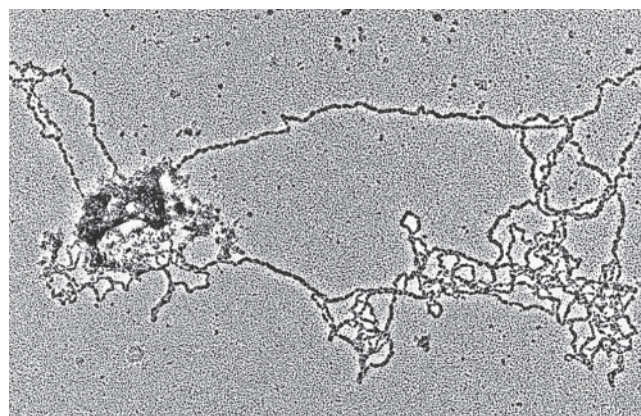


Figure 11.26 An archaeal chromosome, as shown in the electron microscope. The circular chromosome is from the hyperthermophile *Sulfolobus*, a member of the *Archaea*.

synthesis inhibitors) inhibit methanogens. Auxotrophic strains (Section 11.1) of a few *Archaea* have also been isolated for genetic selection purposes.

Examples of Archaeal Genetics

No single species of *Archaea* has become a model organism for archaeal genetics, although more genetic work has been done on select species of extreme halophiles (*Halobacterium*, *Haloferax*, [↔](#) Section 17.1) than on any other *Archaea*. Various mechanisms of gene transfer have been found scattered among a range of *Archaea*. In addition, several plasmids have been isolated from *Archaea* and some have been used to construct cloning vectors, allowing genetic analysis through gene cloning and sequencing rather than through traditional genetic crosses. Transposon mutagenesis (Section 11.11) has been well developed in certain methanogens, including species of *Methanococcus* and *Methanosarcina*, and other tools such as shuttle vectors and other in vitro methods of genetic analysis have been developed for study of the unique biochemistry of the methanogens ([↔](#) Sections 14.17 and 17.2).

Transformation occurs in the methanogen *Methanococcus voltae* and the hyperthermophiles *Thermococcus kodakarensis* and *Pyrococcus furiosus*, all of which are naturally competent (Section 11.6). *Thermococcus* species can also exchange plasmids through the budding of their cell envelope, a process that results in DNA-containing membrane vesicles. Other conditions for transformation work reasonably well in several *Archaea* although the details vary from organism to organism. One approach requires removal of divalent metal ions, which in turn results in the partial disassembly of the glycoprotein cell wall layer that surrounds many archaeal cells (S-layer, [↔](#) Section 2.6), and this allows access to transforming DNA. However, *Archaea* with rigid cell walls have proven difficult to transform, although electroporation (Section 11.6) sometimes works. One exception is in *Methanosarcina* species, organisms with a thick polysaccharide cell wall, for which high-efficiency transformation systems have been developed that employ DNA-loaded lipid preparations (liposomes) that traverse the cell wall to deliver DNA into the cell.

Although viruses that infect *Archaea* are plentiful, transduction is extremely rare. Only one archaeal virus, which infects the thermophilic methanogen *Methanothermobacter thermautotrophicus*, has been shown to transduce the genes of its host. Unfortunately the low burst size (about six phages liberated per cell) makes it impractical to use this system for gene transfer. Gene transfer agents (Section 11.7) have been found in one species of methanogen but do not appear to be widespread in *Archaea*.

Conjugation in *Archaea*

Different types of conjugation have been detected in *Archaea*. Some strains of the thermophilic and acidophilic *Sulfolobus solfataricus* ([↔](#) Section 17.9) contain plasmids that promote conjugation between two cells in a manner similar to that seen in *Bacteria*. In this process, cell pairing is independent of pili formation and DNA transfer is unidirectional. However, most of the genes encoding these functions in *S. solfataricus* seem to have little similarity to those in gram-negative *Bacteria*. The exception is a gene similar to *traG* from the F plasmid, whose protein product participates in stabilizing mating pairs. It thus seems likely

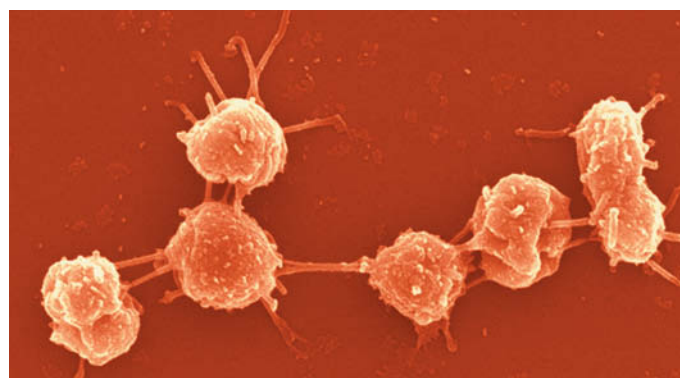


Figure 11.27 Nanotubes and *Thermococcus*. Scanning electron micrograph showing nanotubes linking cells of *Thermococcus* sp. 5-4. A single *Thermococcus* cell is about 1 μm in diameter.

that the actual mechanism of conjugation in *Archaea* is quite different from that in *Bacteria*.

Most species of *Sulfolobus* can also exchange DNA without the participation of a fertility plasmid. Such exchange is dependent on cell aggregation, a process that requires pili whose synthesis is triggered by UV radiation. While the exact mechanism of this unusual DNA exchange is unknown, these pili bring cells of the same species together through recognition of S-layer glycosylation patterns present on the cell walls.

Other *Archaea* form specialized structures between cells that allow for genetic exchange, and the formation of these structures is also independent of a fertility plasmid. For example, **Figure 11.27** illustrates the formation of DNA-transferring nanotubes between cells of a *Thermococcus* species. Interestingly, unlike the other means of horizontal gene transfer discussed in this chapter, the *Thermococcus* nanotubes allow for *bidirectional* transfer of DNA. Thus, with both cells in an exchange able to function as recipients, the *Thermococcus* nanotubes likely facilitate very dynamic gene flow. Similarly, some halobacteria also form cytoplasmic bridges between mating cells that are used for DNA transfer. Although the nanotube and cytoplasmic bridge systems are not yet in routine use, they may well be useful for developing more facile archaeal genetic transfer systems in the future.

MINIQUIZ

- Why is it usually more difficult to select recombinants with *Archaea* than with *Bacteria*?
- Why do penicillins not kill species of *Archaea*?

11.11 Mobile DNA: Transposable Elements

As we have seen, molecules of DNA may move from one cell to another, but to a geneticist, the phrase “mobile DNA” has a special meaning. Mobile DNA refers to discrete segments of DNA that move as units from one location to another *within* other DNA molecules.

Although the DNA of certain viruses can be inserted into and excised from the genome of the host cell ([↔](#) Section 8.7), most mobile DNA consists of **transposable elements**. These are

stretches of DNA that can move from one site to another. However, transposable elements are always found inserted into another DNA molecule such as a plasmid, a chromosome, or a viral genome. Transposable elements do not possess their own origin of replication. Instead, they are replicated when the host DNA molecule into which they are inserted is replicated.

Transposable elements move by *transposition*, a process that is important in both genome rearrangement and genetic analysis. Transposable elements are abundant and widespread in nature and can be found in the genomes of all three domains of life as well as in many viruses and plasmids, suggesting that the elements offer a selective advantage by accelerating genome rearrangement.

The two major types of transposable elements in *Bacteria* are *insertion sequences* (IS) and *transposons*. Both elements have two important features in common: They carry genes encoding *transposase*, the enzyme necessary for transposition, and they have short inverted terminal repeats at their ends that are also needed for transposition (the ends of transposable elements are not free but are continuous with the host DNA molecule into which the transposable element has inserted). **Figure 11.28** shows genetic maps of two well-studied transposable elements: the insertion element IS2 and the transposon Tn5.

Insertion Sequences and Transposons

Insertion sequences (IS) are the simplest type of transposable element. They are short DNA segments, about 1000 nucleotides long, and typically contain inverted repeats of 10–50 base pairs. Each different IS has a specific number of base pairs in its terminal repeats, and the only protein encoded is the transposase. Several hundred distinct IS elements have been characterized. IS elements are found in the chromosomes and plasmids of both *Bacteria* and *Archaea*, as well as in certain bacteriophages. Individual strains of the same bacterial species vary in the number and location of the IS elements they harbor. For instance, the genome of one strain of *Escherichia coli* has five copies of IS2 and five copies of IS3. Many plasmids, such as the F plasmid, also carry IS elements. Indeed,

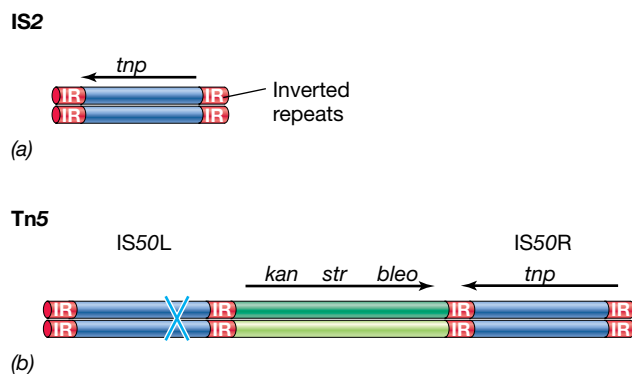


Figure 11.28 Maps of the transposable elements IS2 and Tn5. The arrows above the maps show the direction of transcription of any genes on the elements. The gene encoding the transposase is *tnp*. (a) IS2 is an insertion sequence of 1327 bp with inverted repeats of 41 bp at its ends. (b) Tn5 is a composite transposon of 5.7 kbp containing the insertion sequences IS50 L and IS50 R at its left and right ends, respectively. IS50 L is not capable of independent transposition because there is a nonsense mutation, marked by a blue cross, in its transposase gene. The genes *kan*, *str*, and *bleo* confer resistance to the antibiotics kanamycin (and neomycin), streptomycin, and bleomycin.

integration of the F plasmid into the *E. coli* chromosome is facilitated by recombination between identical IS elements on the F plasmid and the chromosome (Section 11.9 and Figure 11.22).

Transposons are larger than IS elements but have the same two essential components: inverted repeats at both ends and a gene that encodes transposase (Figure 11.28b). The transposase recognizes the inverted repeats and moves the segment of DNA flanked by them from one site to another. Consequently, any DNA that lies between the two inverted repeats is moved and is, in effect, part of the transposon. Genes included inside transposons vary widely. Some of these genes, such as antibiotic resistance genes, confer important new properties on the organism. Because antibiotic resistance is both important and easy to detect, most well-studied transposons contain antibiotic resistance genes as selectable markers. Examples include transposon Tn5, which encodes kanamycin resistance (Figure 11.28b) and Tn10, which encodes tetracycline resistance.

Because any genes lying between the inverted repeats become part of a transposon, it is possible to get hybrid transposons that display complex behavior. For example, conjugative transposons contain *tra* genes and can move between bacterial species by conjugation as well as transpose from place to place within a single bacterial genome. Even more complex is bacteriophage Mu, which is both a virus and a transposon (⇄ Section 10.4). In this case a complete virus genome is contained within a transposon.

Mechanisms of Transposition

Both the inverted repeats (located at the ends of transposable elements) and transposase are essential for transposition. The transposase recognizes, cuts, and ligates the DNA during transposition. When a transposable element is inserted into target DNA, a short sequence in the target DNA at the site of integration is duplicated during the insertion process (Figure 11.29). The duplication arises because single-stranded DNA breaks are made by the transposase. The transposable element is then attached to the single-stranded ends that have been generated. Finally, enzymes of the host cell repair the single-strand portions, which results in the duplication.

Two mechanisms of transposition are known: conservative and replicative (Figure 11.30). In *conservative* transposition, as occurs with the transposon Tn5 (Figure 11.28b), the transposon is excised

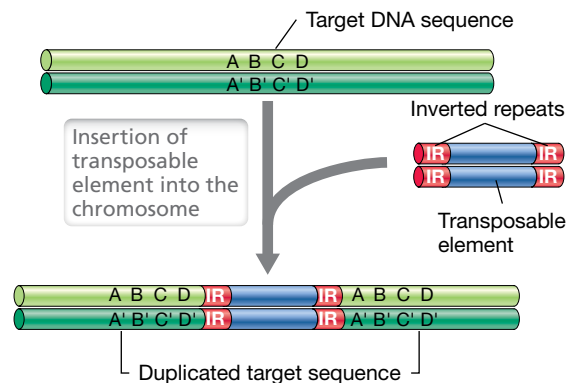


Figure 11.29 Transposition. Insertion of a transposable element generates a duplication of the target sequence. Note the presence of inverted repeats (IR) at the ends of the transposable element.

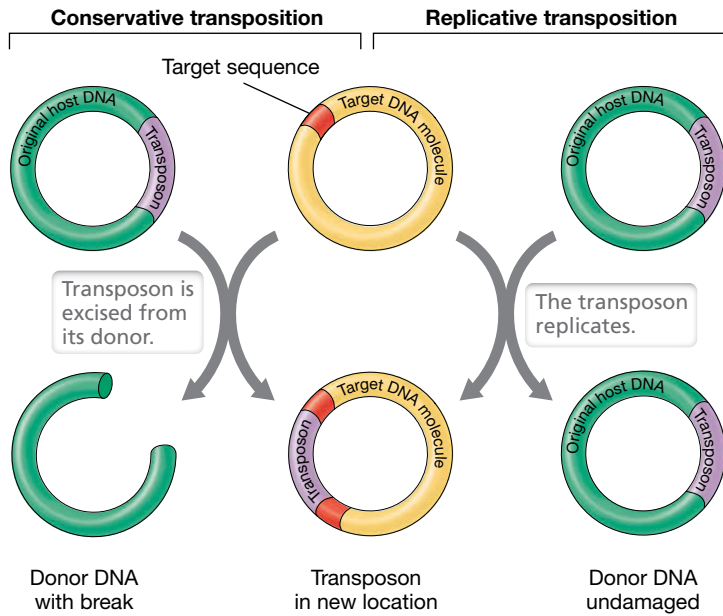


Figure 11.30 Two mechanisms of transposition. Donor DNA (carrying the transposon) is shown in green, and recipient DNA carrying the target sequence is shown in yellow. In both conservative and replicative transposition, the transposase inserts the transposon (purple) into the target site (red) on the recipient DNA. During this process, the target site is duplicated. In conservative transposition, the donor DNA is left with a double-stranded break at the previous location of the transposon. In contrast, after replicative transposition, both donor and recipient DNA possess a copy of the transposon.

from one location and is reinserted at a second location. The copy number of a conservative transposon therefore remains at one. By contrast, during *replicative* transposition, a new copy of the transposon is produced and is inserted at the second location. Thus, after a replicative transposition event, one copy of the transposon remains at the original site, while a second copy is incorporated at the new site.

Utility of Transposon Mutagenesis

When a transposon inserts itself within a gene, the DNA sequence in the gene is altered and a mutation occurs (Figure 11.31). Mutations due to transposon insertion do occur naturally. However, laboratory use of transposons has been a powerful genetic tool to create a library of bacterial mutants. To do this, transposons carrying antibiotic resistance genes are used. The transposon is introduced into the target cell on a plasmid that cannot replicate in that particular host using conjugation or transformation as transfer mechanisms. Consequently, antibiotic-resistant colonies will mostly be due to insertion of the transposon into the bacterial genome.

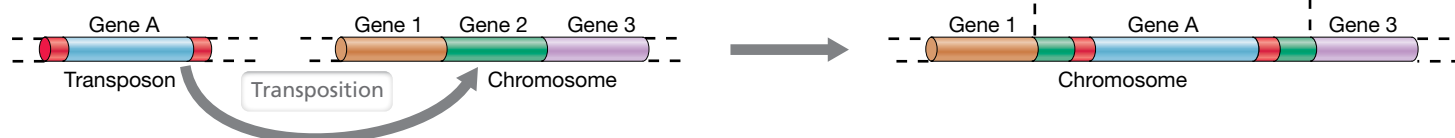


Figure 11.31 Transposon mutagenesis. The transposon moves into the middle of gene 2. Gene 2 is now disrupted by the transposon and is inactivated. Gene A from the transposon is now expressed from the chromosome.

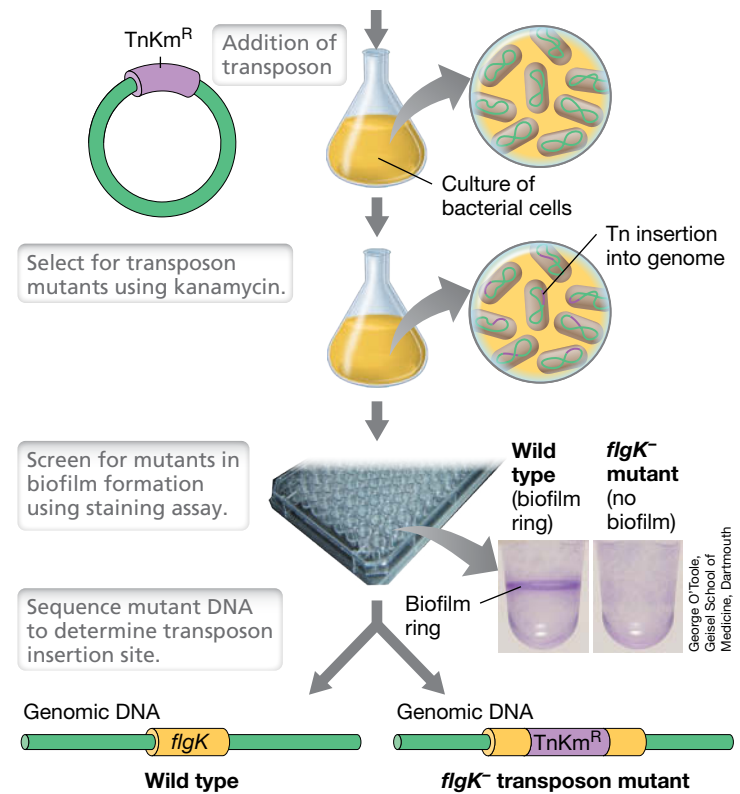


Figure 11.32 Utility of transposon mutagenesis. A transposon (Tn) conferring kanamycin resistance (Km^R) is added to a culture of wild-type *Pseudomonas aeruginosa*. After selection for integration of the transposon using the antibiotic kanamycin, the mutants are screened for biofilm formation in a microtiter plate. Cultures that produce biofilm adhere to the microtiter plate (ring in left tube of inset photo) and the resulting biofilm can be stained with crystal violet. Primers specific to the Tn can be used to determine the Tn insertion site and the identity of the interrupted gene. *flgK* encodes a protein in the flagellar hook.

Because bacterial genomes contain relatively little noncoding DNA, most transposon insertions will occur in genes that encode proteins. This technique can be used to determine the function of a novel gene if a screening method is available. For example, if a transposon inserts into a gene encoding a product required for biofilm formation (biofilms are colonies of microbes encased in an adhesive and attached to a surface, Section 5.1), the transposon mutant will no longer grow in the biofilm mode. Then, further genetic analyses can be performed to reveal which gene the transposon has disrupted. Figure 11.32 illustrates the use of transposon mutagenesis to study flagella structure and function in a notorious biofilm-producing bacterium, *Pseudomonas aeruginosa* (Section 7.9). In this research, transposon mutagenesis combined

George O'Toole,
Geisel School of
Medicine, Dartmouth

with a special staining technique was used to identify the flagellar hook protein FlgK (↗ Section 2.11) as a protein required for biofilm formation.

Two transposons widely used for mutagenesis of *E. coli* and related bacteria are Tn5 (Figure 11.28b) and Tn10. Many *Bacteria*, a few *Archaea*, and the yeast *Saccharomyces cerevisiae* have all been mutagenized using genetically engineered transposons. More recently, transposons have even been used to isolate mutations in animals, including mice.

MINIQUIZ

- Which features do insertion sequences and transposons have in common?
- What is the significance of the terminal inverted repeats of transposons?
- How can transposons be used in bacterial genetics?

11.12 Preserving Genomic Integrity: CRISPR Interference

Bacteria and *Archaea* not only produce restriction endonucleases (↗ Section 8.5) that function to destroy incoming foreign DNA, they also have an RNA-based defense system to destroy invading DNA from viral infections and some horizontally transferred

genes. This prokaryotic “immune system”—called CRISPR—was previously described as a major means for *Bacteria* and *Archaea* to evade viral destruction (↗ Section 10.13). But CRISPR also helps the cell maintain the stability and integrity of its genome by destroying certain plasmids and other genes obtained from horizontal transfers.

CRISPR Mechanism

The CRISPR region on the bacterial chromosome is essentially a memory bank of incoming nucleic acid sequences used for surveillance against foreign DNA. It consists of many different segments of foreign DNA called *spacers* alternating with identical repeated sequences (Figure 11.33). The spacer sequences correspond to pieces of foreign DNA that have previously invaded the cell. Once the spacers are recombined into the CRISPR region, the system provides resistance to any incoming DNA (and sometimes RNA) that contains the same or very closely related sequences to those in individual spacer regions.

The key to the CRISPR region’s ability to prevent horizontal transfer of some incoming DNAs is the transcription of a long RNA molecule that is then cleaved in the middle of each of the repeated sequences by the nuclease activity of CRISPR-associated (Cas) proteins. This converts the long RNA molecule into spacer segments of small RNAs called CRISPR RNAs (crRNAs). If one of these crRNAs base-pairs with an invading nucleic acid, then the foreign DNA or RNA duplex is destroyed by the nuclease activity of other Cas

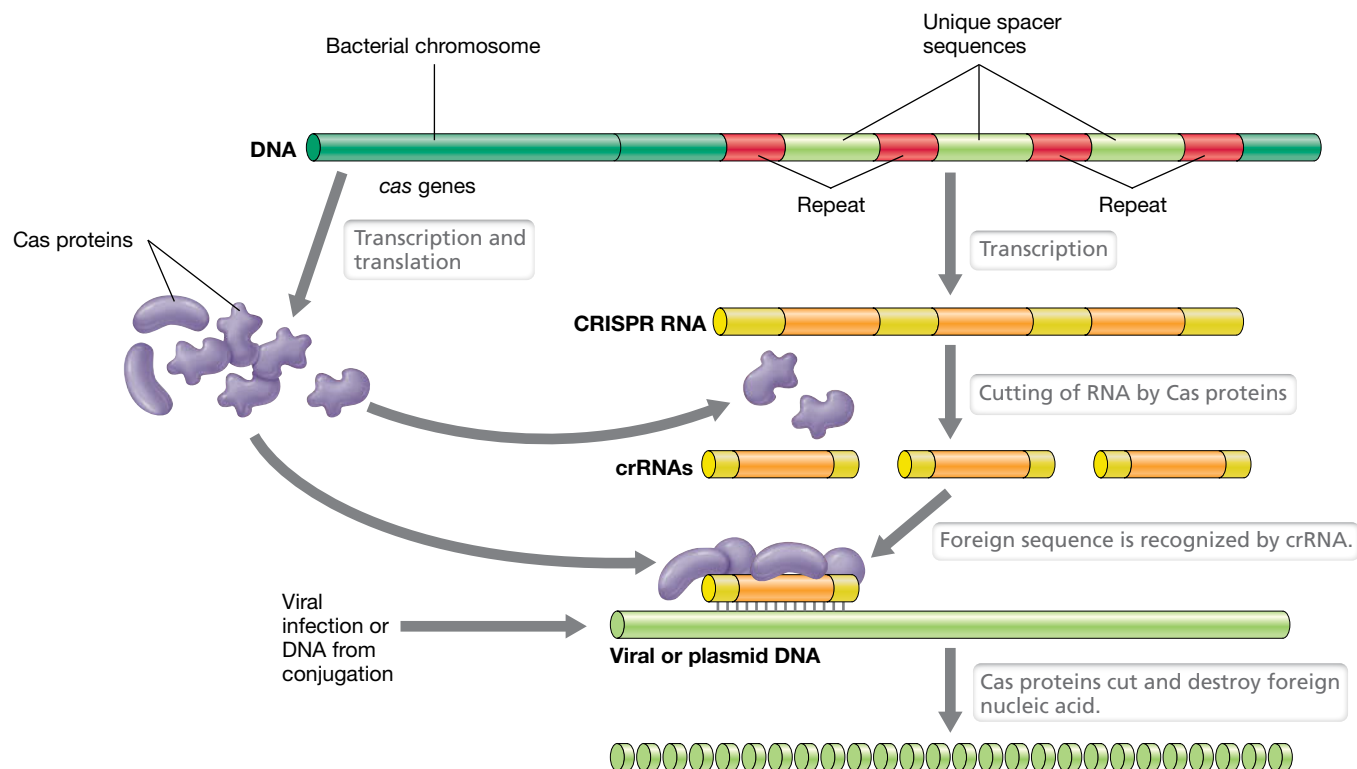


Figure 11.33 Operation of the CRISPR system. The CRISPR region on the bacterial chromosome is transcribed into a long RNA molecule that is then processed into segments by some of the Cas proteins. Each spacer segment corresponds to previous encounters with incoming foreign nucleic acid. If one of these short CRISPR RNA (crRNA) molecules (corresponding to a spacer) recognizes and base pairs with incoming nucleic acid from transduction or conjugation, other Cas proteins destroy the foreign nucleic acid.

proteins (Figure 11.33). This destruction prevents the foreign DNA from being replicated or recombined into the genome and thus keeps the genome from accumulating random bits of foreign DNA that could compromise genome integrity.

Distribution of CRISPR

The CRISPR system is widely distributed in both *Archaea* and *Bacteria*. Approximately 90% of the sequenced genomes of *Archaea* and 70% of those of *Bacteria* possess a CRISPR system. The utility of the system was first demonstrated in the dairy industry where starter cultures used for milk fermentation were susceptible to rampant bacteriophage infection. However, a strain of *Streptococcus thermophilus* was found to be resistant to virulent bacteriophage, and the difference between this *S. thermophilus* strain and those susceptible to viral infection was the spacers within its CRISPR region.

While it is unknown why some viruses and other foreign DNAs possess the initial recognition sequences (PAMs, Section 10.13)

for spacer incorporation by the CRISPR system and others do not, laboratory experiments have shown that bacteriophages can overcome recognition by the Cas proteins and crRNAs by modifying their genome through mutation. This illustrates how the CRISPR system—central as it may be to preventing attacks on a cell's genome—does have a drawback in its reliance on recognizing previously encountered DNA sequences that can continue to mutate in their host and then reappear in the future in a form not recognized by CRISPR.

In Chapter 12 we discuss the CRISPR system in a totally different context as a powerful tool in biotechnology for editing genomes and creating recombinant organisms.

MINIQUIZ

- Why is the CRISPR system considered a prokaryotic “immune system”?
- What do the spacers within the CRISPR region correspond to?

MasteringMicrobiology®

Visualize, explore, and think critically with Interactive Microbiology, MicroLab Tutors, MicroCareers case studies, and more. MasteringMicrobiology offers practice quizzes, helpful animations, and other study tools for lecture and lab to help you master microbiology.

Chapter Review

I • Mutation

11.1 Mutation is a heritable change in the nucleotide sequence of the genome and may lead to a change in phenotype. Selectable mutations are those that give the mutant a growth advantage under certain environmental conditions and are especially useful in genetic research. If selection is not possible, mutants must be identified by screening.

Q Write a one-sentence definition of the term “genotype.” Do the same for “phenotype.” Does the phenotype of an organism automatically change when a change in genotype occurs? Why or why not? Can phenotype change without a change in genotype? In both cases, give examples to support your answer.

11.2 Mutations, either spontaneous or induced, are in the base sequence of the nucleic acid in a genome. A point mutation is due to a single base-pair change. In a nonsense mutation, the codon becomes a stop codon and an incomplete polypeptide is made. Deletions and insertions cause more dramatic changes in the DNA, including frameshift mutations that often result in complete loss of gene function.

Q What are point mutations? How do silent mutations differ from missense and nonsense mutations?

11.3 Different types of mutations occur at different frequencies. For a typical bacterium, mutation rates of

10^{-6} to 10^{-7} per kilobase pair are generally seen. Although RNA and DNA polymerases make errors at about the same rate, RNA genomes typically accumulate mutations at much higher frequencies than DNA genomes.

Q What is a revertant? Explain the two common types of revertants.

11.4 Mutagens are chemical, physical, or biological agents that increase the mutation rate. Mutagens can alter DNA in many different ways. However, alterations in DNA are not mutations unless they are inherited. Some DNA damage can lead to cell death if not repaired, and both error-prone and high-fidelity DNA repair systems exist.

Q Explain the primary mechanisms of each of the three types of mutagens.

II • Gene Transfer in *Bacteria*

11.5 Homologous recombination occurs when closely related DNA sequences from two distinct genetic elements are combined together in a single element. Recombination is an important evolutionary process, and cells have specific mechanisms for ensuring that recombination takes place.

Q What are heteroduplex regions of DNA and what process leads to their formation?

11.6 Certain bacteria exhibit competence, a state in which cells are able to take up free DNA released by other bacteria.

Incorporation of donor DNA into a recipient cell requires the activity of single-strand binding protein, RecA protein, and several other enzymes. Only competent cells are transformable.

Q Explain why recipient cells do not successfully take up plasmids during natural transformation.

- 11.7** Transduction is the transfer of host genes from one bacterium to another by a bacterial virus. In generalized transduction, defective virus particles randomly incorporate fragments of the cell's chromosomal DNA, but the transducing efficiency is low. In specialized transduction, the DNA of a temperate virus excises incorrectly and takes adjacent host genes along with it; the transducing efficiency here may be very high.

Q Explain how a generalized transducing particle differs from a specialized transducing particle.

- 11.8** Conjugation is a mechanism of DNA transfer in *Bacteria* and *Archaea* that requires cell-to-cell contact. Conjugation is controlled by genes carried by certain plasmids (such as the F plasmid) and requires transfer of the plasmid from a donor cell to a recipient cell. Plasmid DNA transfer requires replication using the rolling circle mechanism.

Q What is a sex pilus and which cell type, F^- or F^+ , would produce this structure?

- 11.9** The donor cell chromosome can be mobilized for transfer to a recipient cell. This requires an F plasmid to integrate into the chromosome to form the Hfr phenotype. Because transfer of the host chromosome is rarely complete, recipient cells rarely become F^+ . F' plasmids are previously integrated F plasmids that have excised and captured some chromosomal genes.

Q What is a merodiploid and how does an F' plasmid yield a merodiploid?

III • Gene Transfer in *Archaea* and Other Genetic Events

- 11.10** Archaeal research lags behind bacterial research in the development of systems for gene transfer. Many antibiotics are ineffective against *Archaea*, making it difficult to select recombinants effectively. The unusual growth conditions needed by many *Archaea* also make genetic experimentation challenging. Nevertheless, the genetic transfer systems of *Bacteria*—transformation, transduction, and conjugation—are all known in *Archaea*.

Q Explain one type of conjugation in *Archaea* and how it differs from F-plasmid-mediated conjugation.

- 11.11** Transposons and insertion sequences are genetic elements that can move from one location on a host DNA molecule to another by transposition. Transposition can be either replicative or conservative. Transposons often carry genes encoding antibiotic resistance and can be used to identify the functions of unknown genes.

Q What are the major differences between insertion sequences and transposons?

- 11.12** The clustered regularly interspaced short palindromic repeat (CRISPR) system is an RNA-based mechanism of protecting the genome of *Bacteria* and *Archaea* from invading DNA resulting from viral infection or conjugation. If small RNA molecules resulting from the spacer regions of the CRISPR region bind to incoming complementary DNA, Cas proteins destroy the nucleic acid duplex.

Q Explain why incoming DNA recognized by a short RNA molecule expressed from the CRISPR region cannot be completely foreign to the cell.

Application Questions

1. A constitutive mutant is a strain that continuously makes a protein that is inducible in the wild type. Describe two ways in which a change in a DNA molecule could lead to the emergence of a constitutive mutant. How could these two types of constitutive mutants be distinguished genetically?
2. Although a large number of mutagenic chemicals are known, none is known that induces mutations in only a single gene (gene-specific mutagenesis). From what you know about mutagens, explain why it is unlikely that a gene-specific chemical mutagen will be found. How then is site-specific mutagenesis accomplished?
3. Why is it difficult in a single experiment to transfer a large number of genes to a recipient cell using transformation or transduction?
4. Transposable elements cause mutations when inserted within a gene. These elements disrupt the continuity of a gene. Introns also disrupt the continuity of a gene, yet the gene is still functional. Explain why the presence of an intron in a gene does not inactivate that gene but insertion of a transposable element does.

Chapter Glossary

Auxotroph an organism that has developed an additional nutritional requirement compared with the wild type, often as a result of mutation

Conjugation the transfer of genes from one prokaryotic cell to another by a mechanism requiring cell-to-cell contact

Frameshift mutation a mutation in which insertion or deletion of nucleotides changes the groups of three bases in which the genetic code is read within an mRNA, usually resulting in a faulty product

Genotype the complete genetic makeup of an organism; the complete description of a cell's genetic information

Heteroduplex a DNA double helix composed of single strands from two different DNA molecules

Hfr cell a cell with the F plasmid integrated into the chromosome

Induced mutation a mutation caused by external agents such as mutagenic chemicals or radiation

Insertion sequence (IS) the simplest type of transposable element, which carries only genes that participate in transposition

Missense mutation a mutation in which a single codon is altered so that one amino acid in a protein is replaced with a different amino acid

Mutagen an agent that causes mutation

Mutant an organism whose genome carries a mutation

Mutation a heritable change in the base sequence of the genome of an organism

Nonsense mutation a mutation in which the codon for an amino acid is changed to a stop codon

Phenotype the observable characteristics of an organism

Point mutation a mutation that involves a single base pair

Recombination a resorting or rearrangement of DNA fragments resulting in a new sequence

Reversion an alteration in DNA that reverses the effects of a prior mutation

Rolling circle replication a mechanism of replicating double-stranded circular DNA that starts by nicking and unrolling one strand and using the other (still circular) strand as a template for DNA synthesis

Screening a procedure that permits the identification of organisms by phenotype or genotype, but does not inhibit or enhance the growth of particular phenotypes or genotypes

Selection placing organisms under conditions that favor or inhibit the growth of those with a particular phenotype or genotype

Silent mutation a change in DNA sequence that has no effect on the phenotype

SOS repair system a DNA repair system activated by DNA damage

Spontaneous mutation a mutation that occurs “naturally” without the help of mutagenic chemicals or radiation

Transduction the transfer of host cell genes from one cell to another by a virus

Transformation the transfer of bacterial genes through the uptake of free DNA from the environment

Transition a mutation in which a pyrimidine base is replaced by another pyrimidine or a purine is replaced by another purine

Transposable element a genetic element able to move (transpose) from one site to another on host DNA molecules

Transposon a type of transposable element that carries genes in addition to those required for transposition

Transversion a mutation in which a pyrimidine base is replaced by a purine or vice versa

Wild-type strain a bacterial strain isolated from nature or one used as a parent in a genetics investigation

12

Biotechnology and Synthetic Biology

microbiologynow

Creation of a New Life Form: Design of a Minimal Cell

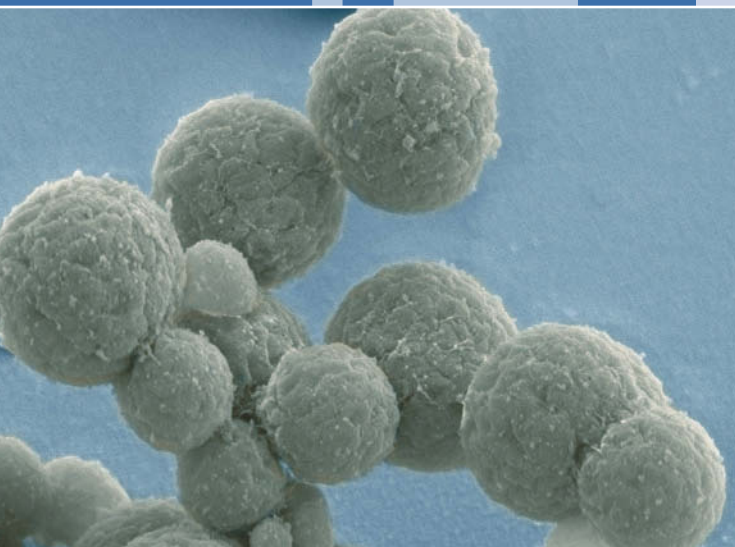
A cell's genome is its blueprint for life. However, what is the bare minimum number of genes needed to sustain a free-living cell? This is a question that microbiologists at the J. Craig Venter Institute (JCVI) have attempted to answer ever since they sequenced the genomes of several *Mycoplasma* species in the 1990s. Because *Mycoplasma* species are parasitic bacteria, their genomes are already reduced in size and hence provide an excellent foundation for creating a "minimal cell." However, little did the scientists at JCVI suspect that it would take 20 years to satisfy their scientific curiosity!

Instead of beginning by genetically manipulating a *Mycoplasma* species, microbiologists at JCVI wanted to have more control. To begin unraveling the genetic requirements for life, they first generated a synthetic self-replicating *Mycoplasma* (described in this chapter). The genome of this pioneering synthetic life form was synthesized from scratch based on its known genome sequence. The synthetic cell did not possess a "designer genome," or even a minimal one; it simply contained its own genome but one completely constructed in the laboratory. This breakthrough in synthetic biology provided the technology needed for microbiologists to create designer genomes.

Using comparative genomics and prior knowledge about specific gene sequences, microbiologists at JCVI continued their work by designing and synthesizing several minimal genomes that they hypothesized would sustain life. To their dismay, none of these resulted in a viable cell. So instead, they generated modules of DNA corresponding to a *Mycoplasma* genome and sewed different combinations together to form synthetic genomes. Once viable cells were obtained from transplanting these genomes, nonessential genes from the smallest genome were identified by transposon mutagenesis. After removing these unnecessary genes, a synthetic minimal cell coined JCVI-syn3.0 was created (see photo). This autonomous life form possesses a 531-kilobase genome encoding 473 genes; JCVI-syn3.0 thus contains a genome smaller than any other free-living cell.

While this work showcases the amazing advancements in synthetic biology and the potential for creating designer cells with novel functions, a surprising mystery surrounds this minimal cell: The roles for almost a third of JCVI-syn3.0's genes remain unknown, highlighting how much we still need to learn about the genetic foundation of a living cell.

 **Source:** Hutchison, C.A. 3rd, et al. 2016. Design and synthesis of a minimal bacterial genome. *Science* 351(6280): aad6253. Photo provided by Clyde Hutchison and J. Craig Venter, JCVI and Thomas Deerinck and Mark Ellisman, NCMIR.



- I Tools of the Genetic Engineer 369
- II Making Products from Genetically Engineered Microbes: Biotechnology 379
- III Synthetic Biology and Genome Editing 389

Industrial microbiology uses microbes on a large scale to produce desired products such as enzymes, foods, and beverages. These microbes are usually not genetically modified. Instead, naturally overproducing strains are isolated from wild-type strains and used for industrial purposes. In contrast, biotechnology uses genetically modified microorganisms to produce high-value products that the organisms do not naturally produce. In this chapter we discuss the basic techniques of genetic engineering that underlie biotechnology, in particular those used to clone, alter, and express genes efficiently in host organisms. We also explore how genetic engineering and biotechnology can be used for industrial, medical, and agricultural applications and introduce the exciting new field of synthetic biology.

I • Tools of the Genetic Engineer

Performing genetics *in vivo* (in living organisms) has many limitations that can be overcome by manipulating DNA *in vitro* (in a test tube). **Genetic engineering** refers to the use of *in vitro* techniques to alter genes in the laboratory. Such altered genes may be reinserted into the original source organism or into some other host organism. Expression of a gene from one organism in a different host organism is called **heterologous expression**.

Genetic engineering requires that specific DNA be isolated, purified, and further manipulated. We begin by considering some of the basic tools of the genetic engineer, including amplification of DNA, the separation of nucleic acids by electrophoresis, nucleic acid hybridization, and molecular cloning. We also describe methods for expressing foreign genes in bacteria and targeted mutagenesis.

12.1 Manipulating DNA: PCR and Nucleic Acid Hybridization

The first objective of genetic engineering is to isolate copies of specific genes in pure form, and the key method for doing so is the **polymerase chain reaction (PCR)** (Figure 12.1). Simply put, the polymerase chain reaction is DNA replication *in vitro*, as segments of target DNA are multiplied by up to a billionfold in the process of *amplification*. During each round of amplification, the amount of DNA *doubles*, leading to an exponential increase in the target DNA. Using an automated PCR machine called a *thermocycler*, a large amount of amplified DNA can be produced from only a few molecules of target DNA. In some cases it is desirable to quantify the initial amount of target DNA, and a variation on PCR called *quantitative PCR (qPCR)* is used for this purpose (see Figures 28.23 and 28.24). A second variation on the original PCR technique allows for amplification of RNA (following its conversion to DNA, as discussed later in this section).

PCR and Polymerases

PCR requires DNA polymerase, the enzyme that naturally copies DNA molecules (see Section 4.3), and artificially synthesized oligonucleotide primers (Section 12.4) made of DNA (rather than RNA like the primers used by cells) to synthesize DNA. PCR does not actually copy whole DNA molecules but amplifies stretches of up to a few thousand base pairs (the *target*) from within a

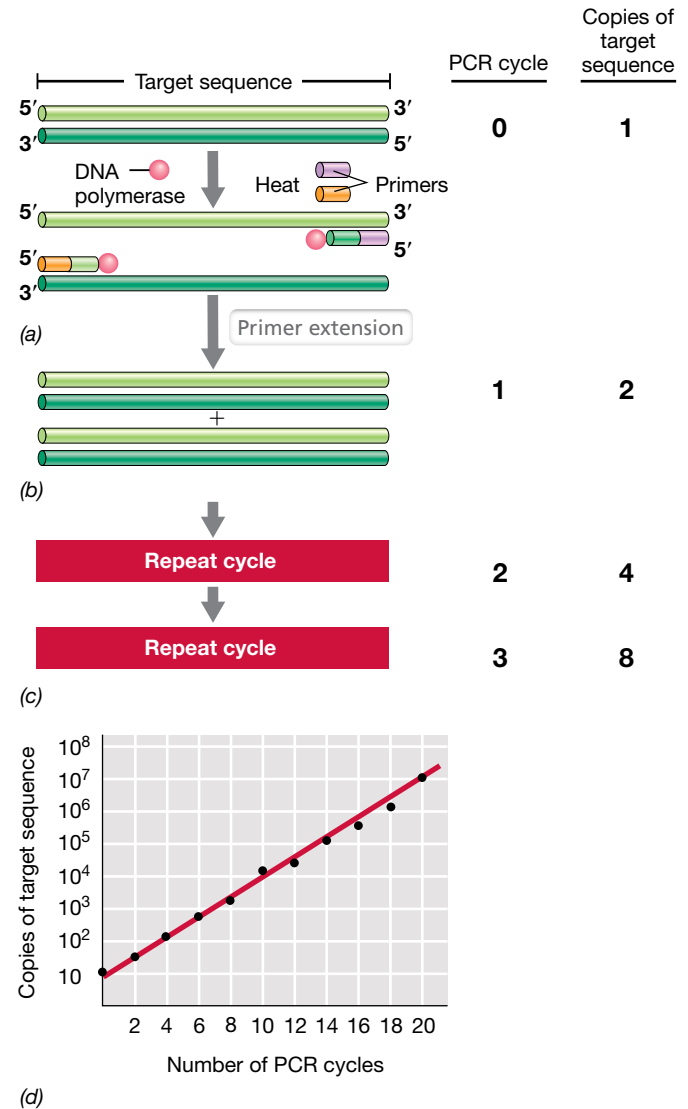


Figure 12.1 The polymerase chain reaction (PCR). The PCR amplifies specific DNA sequences. (a) Target DNA is heated to separate the strands, and a large excess of two oligonucleotide primers, one complementary to each strand, is added along with DNA polymerase. (b) Following primer annealing, primer extension yields a copy of the original double-stranded DNA. (c) Two additional PCR cycles yield four and eight copies, respectively, of the original DNA sequence. (d) Effect of running 20 PCR cycles on a DNA preparation originally containing ten copies of a target gene. Note that the plot is semilogarithmic.

larger DNA molecule (the *template*) during the following steps (Figure 12.1):

1. Template DNA is denatured by heating and then two DNA oligonucleotide primers flanking the target DNA on each strand are added in excess. This ensures that most template strands anneal to a primer, and not to each other, as the mixture cools (Figure 12.1a).
2. DNA polymerase then extends the primers using the original DNA as the template (Figure 12.1b).
3. After an appropriate incubation period, the mixture is heated again to separate the strands, but now the target gene is present in twice the original amount. The mixture is then cooled to

allow the primers to hybridize with complementary regions of newly synthesized and original DNA, and the process is repeated (Figure 12.1c). In practice, 20–30 cycles are typically run, yielding a 10^6 -fold to 10^9 -fold increase in the target sequence (Figure 12.1d).

Because high temperatures are used to denature the double-stranded copies of DNA *in vitro*, use of a thermostable DNA polymerase is critical. *Taq polymerase*, a DNA polymerase isolated from the thermophilic hot spring bacterium *Thermus aquaticus* (see Section 16.20), is stable to 95°C and thus is unaffected by the denaturation step employed in the PCR (Figure 12.1). A DNA polymerase from *Pyrococcus furiosus*, a hyperthermophilic species of *Archaea* with a growth temperature optimum of 100°C (see Section 17.4), is called *Pfu polymerase* and is even more thermostable than *Taq polymerase*. Moreover, unlike *Taq polymerase*, *Pfu polymerase* has proofreading activity (see Section 4.4), making it especially useful when high accuracy is crucial. To supply the demand for thermostable DNA polymerases, the genes encoding these enzymes have been cloned into *Escherichia coli*, allowing the enzymes to be produced commercially in large quantities.

PCR Applications and RT-PCR

PCR is extremely valuable for obtaining DNA for gene cloning or for sequencing purposes because the gene or genes of interest can easily be amplified if flanking sequences are known. PCR is also used routinely in comparative or phylogenetic studies to amplify genes from various sources. In these cases the primers are made commercially to base-pair with regions of the gene that are conserved in sequence across a wide variety of organisms. Because small ribosomal subunit (SSU) rRNA—a molecule used for phylogenetic analyses—has both highly conserved and highly variable regions (see Section 13.7 and Figure 13.15), primers specific for the SSU rRNA gene from different taxonomic groups can be used to survey habitats for their microbial contents (see Section 19.6). Also, because it is so sensitive, PCR can be used to amplify very small quantities of DNA. For example, PCR has been used to amplify DNA from sources as varied as mummified human remains, fossilized plants and animals, and even single microbial cells (see Section 9.12). It is also a common tool of medical diagnostics in clinical microbiology laboratories (see Section 28.8) and is widely used in forensic science to attach an identity to evidence from a crime scene such as blood, semen, or tissue samples.

An important extension of the standard PCR procedure is *reverse transcription PCR* (RT-PCR), used to make DNA from an mRNA template (Figure 12.2). This procedure can be used to detect if a gene is expressed or to produce an intron-free eukaryotic gene for expression in bacteria (Section 12.3). RT-PCR uses the retroviral enzyme *reverse transcriptase* to convert RNA into **complementary DNA (cDNA)** (see Section 10.11). Figure 12.2 illustrates how reverse transcriptase makes a single strand of cDNA using RNA as a template. To make cDNA, a primer complementary to the 3' end of the target RNA is used by the enzyme reverse transcriptase to initiate RNA synthesis. If the template is eukaryotic mRNA, a primer complementary to the poly(A) tail (see Section 4.6) of the mRNA can be used. The activity of reverse transcriptase results in a hybrid nucleic acid molecule containing both DNA and RNA. RNase H,

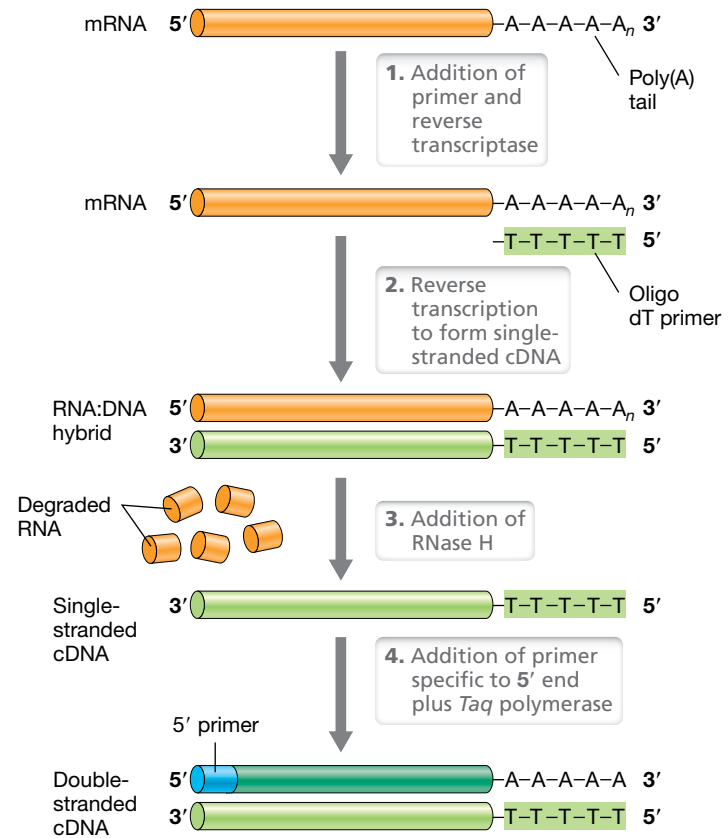


Figure 12.2 Reverse transcription PCR. Steps in the synthesis of cDNA from a eukaryotic mRNA. Reverse transcriptase synthesizes a hybrid molecule containing both RNA and DNA using the mRNA as a template and oligo-dT primer as a substrate. Next, the enzyme RNase H hydrolyzes the RNA portion of the hybrid molecule, yielding a single-stranded molecule of complementary DNA (cDNA). Following the addition of a primer complementary to the 5' end of the cDNA, *Taq polymerase* produces a double-stranded cDNA.

a ribonuclease specific for the hybrid molecule, hydrolyzes the RNA, leaving the single-stranded cDNA as template for standard PCR using an additional primer complementary to the 5' end.

Gel Electrophoresis and Nucleic Acid Hybridization

To verify that amplification of a nucleic acid was successful and for other nucleic acid manipulation steps, DNA or RNA fragments can be separated from each other by **gel electrophoresis**, a technique that employs an agarose gel to separate nucleic acid fragments based on differences in their size and charge (Figure 12.3a). When an electrical current is applied, nucleic acids move through the gel toward the positive electrode because of their negatively charged phosphate groups, and small molecules migrate more rapidly than large molecules. After the gel has been run for a time sufficient to separate the molecules, the gel is stained with *ethidium bromide*, a compound that binds to nucleic acids and makes them fluoresce (Figure 12.3b). To determine the size of the DNA or RNA of interest, the migration is compared to a standard sample consisting of nucleic acid fragments of known sizes, called a *ladder*. DNA fragments can then be purified from gels and used for a variety of purposes, such as cloning or hybridization.

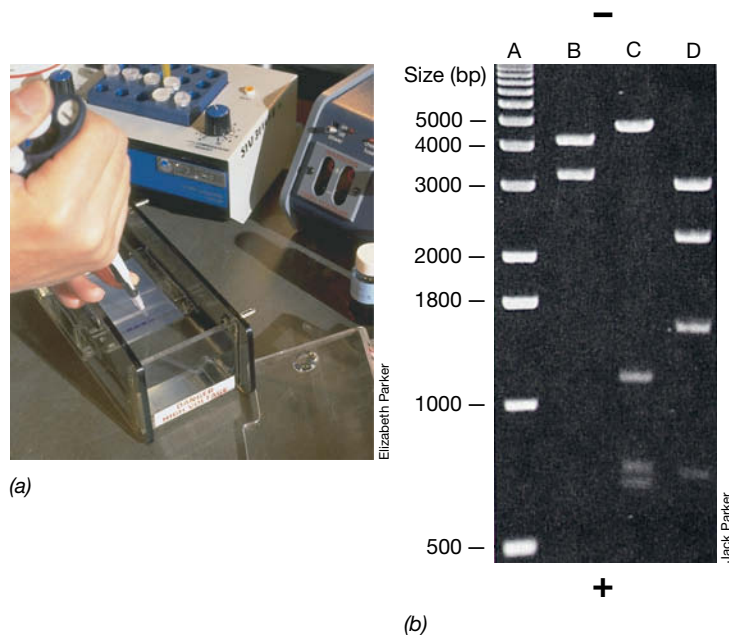
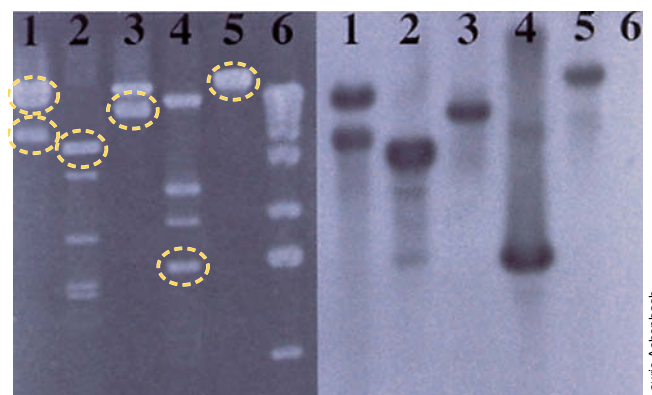


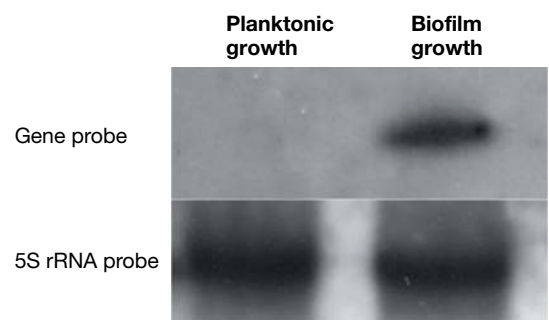
Figure 12.3 Agarose gel electrophoresis of DNA. (a) DNA samples are loaded into wells in a submerged agarose gel. (b) A photograph of a stained agarose gel. The DNA was loaded into wells toward the top of the gel (negative pole) as shown, and the positive electrode is at the bottom. The standard sample in lane A (DNA ladder) has fragments of known size that may be used to determine the sizes of the fragments in the other lanes. Bands stain less intensely at the bottom of the gel because the fragments are smaller, and thus there is less DNA to stain.

When DNA is denatured (that is, the two strands are separated), the single strands can be used to form hybrid double-stranded molecules with other single-stranded DNA (or RNA) molecules by complementary base pairing (↔ Section 4.1) in a process called *nucleic acid hybridization*, or **hybridization** for short. Hybridization is widely used in detecting, characterizing, and identifying segments of DNA and RNA. Single-stranded nucleic acids whose identity is already known and that are used in hybridization are called **nucleic acid probes**, or simply *probes*. To allow detection, probes are made radioactive or are labeled with chemicals that are colored or yield fluorescent products (↔ Section 19.5), and by varying the hybridization conditions, the “stringency” of the hybridization can be adjusted such that complementary base pairing is somewhat flexible or, alternatively, must be nearly exact.

Hybridization is useful for finding related sequences in different genomes or other genetic elements and to determine if a gene is expressed into an RNA transcript. In *Southern blotting*, probes of known sequence are hybridized to target DNA fragments that have been separated by gel electrophoresis. The hybridization procedure in which DNA is the target sequence in the gel, and RNA or DNA is the probe, is called a **Southern blot**. By contrast, a **Northern blot** uses RNA as the target sequence and DNA or RNA as the probe to detect gene expression. In both techniques, the nucleic acid fragments must be in a single-stranded form and are transferred to a synthetic membrane. The membrane is then exposed to the labeled probe. If the probe is complementary to any of the fragments, hybrids form, and the probe attaches to the membrane at the locations of the complementary fragments. **Figure 12.4** shows



(a) **Southern blot**

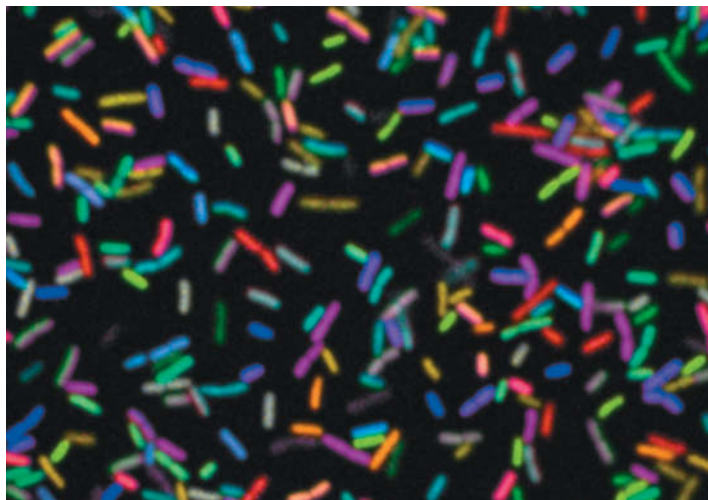


(b) **Northern blot**

Figure 12.4 Nucleic acid hybridization. (a) Southern blotting. (Left panel) Purified molecules of DNA from several different plasmids were treated with restriction enzymes and then subjected to agarose gel electrophoresis. (Right panel) Blot of the DNA gel shown to the left. After blotting, DNA in the gel was hybridized to a radioactive probe. The positions of the bands were visualized by X-ray autoradiography. Note that only some of the DNA fragments (circled in yellow in the left panel) have sequences complementary to the labeled probe. Lane 6 contained DNA used as a size marker and none of the bands hybridized to the probe. (b) Northern blotting. (Top panel) Hybridization and detection of a radioactive gene-specific probe to a blot of total RNA. The probe only bound to RNA from biofilm-grown cells, indicating that the target gene is not expressed during planktonic (suspended) growth. (Bottom panel) Hybridization and detection of a radioactive probe corresponding to the 5S rRNA to the same blot. The signal intensity indicates that equal amounts of RNA from each sample were loaded into the gel.

how a Southern blot can be used to identify fragments of DNA containing sequences that hybridize to the probe and how the intensity of a signal on a Northern blot gives a rough estimate of mRNA abundance from the target gene.

Nucleic acid hybridization has many other uses. Hybridization is the basis of the fluorescence in situ hybridization (FISH) technique (↔ Section 19.5) (**Figure 12.5**), where fluorescent probes are used to target specific DNA (or RNA) sequences in cells. This approach allows the identification of pathogens in clinical samples or bacteria of interest in environmental samples. For example, **Figure 12.5** demonstrates the simultaneous use of eight different oligonucleotide probes in combinations to distinguish between 28 different strains of *E. coli* whose SSU rRNA sequences varied only slightly from strain to strain. The variations in color give a visual indication of the specificity and power of nucleic acid probes. Hybridization is also important in various “omics,” in particular transcriptomics and metatranscriptomics, where genome-wide gene expression can be monitored in pure cultures



Alex Valm and Gary Borsy, Marine Biological Laboratory, Woods Hole, MA

Figure 12.5 Fluorescence spectral image of 28 differently labeled strains of *Escherichia coli*. Cells were labeled with combinations of fluorophore-conjugated oligonucleotides that are complementary to *E. coli* 16S rRNA.

and natural populations, respectively, using microarray technology (see Section 9.9).

MINIQUIZ

- Why is a primer needed at each end of the DNA segment being amplified by PCR?
- How does RT-PCR differ from traditional PCR?
- What are some applications of nucleic acid hybridization in molecular biology?

12.2 Molecular Cloning

The movement of desired genes from their original source to a small and manipulable genetic element (the **vector**) is called **molecular cloning**. Molecular cloning results in **recombinant DNA**, a molecule containing DNA from different sources. Once cloned, the gene(s) of interest can be manipulated, and when the recombinant vector is placed in an appropriate host, the cloned DNA is replicated, providing the foundation for much of genetic engineering.

An Overview of Gene Cloning

Following isolation of the source DNA, the major steps in gene cloning are (1) inserting the DNA into a cloning vector (Figure 12.6), and (2) inserting the vector into a host. The source DNA can be a gene or genes amplified by the polymerase chain reaction (Section 12.1), DNA synthesized from an RNA template by reverse transcriptase (Section 12.1), or even completely synthetic DNA made in vitro (Section 12.4). Cloning vectors are small, independently replicating genetic elements that can both carry and replicate cloned DNA segments (see Figure 12.8). Cloning vectors are typically designed to allow insertion of foreign DNA at a *restriction site* (Figure 12.6). *Restriction endonucleases*, or **restriction enzymes** for short, recognize specific base sequences (restriction sites) within DNA and cut

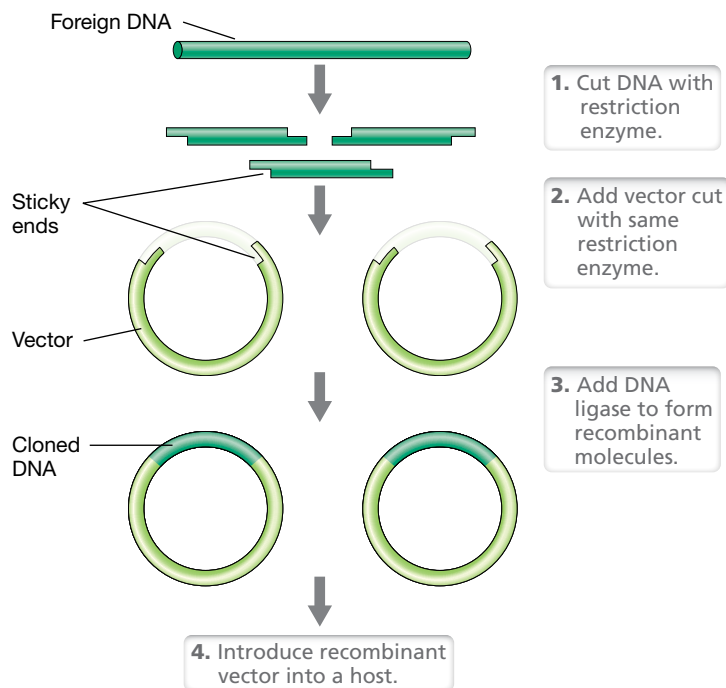


Figure 12.6 Major steps in gene cloning. By cutting the foreign DNA and the vector DNA with the same restriction enzyme, complementary sticky ends are generated that allow foreign DNA to be inserted into the vector.

the phosphodiester backbone, resulting in double-stranded breaks (Figure 12.7). The recognition sequences are typically inverted repeats and are called *palindromes*.

Restriction enzymes with different sequence specificities are widespread among *Bacteria*, where they help protect cells from attack by viral DNA. The cell is protected from its own restriction enzyme(s) by chemical modification (typically by methylation) of one of the bases in any potential restriction sites that exist in its genome. The restriction enzyme *EcoRI* makes staggered cuts,

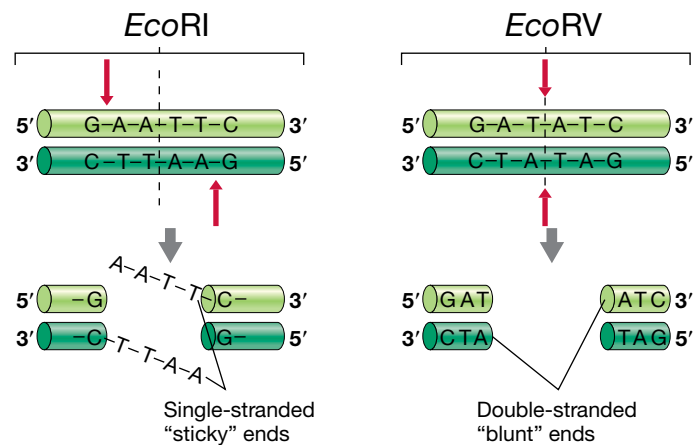


Figure 12.7 Restriction and modification of DNA. Sequences of DNA recognized by the restriction endonucleases *EcoRI* and *EcoRV*. The red arrows indicate the bonds cleaved by the enzyme, and the dashed line indicates the axis of symmetry of the sequence. After cutting DNA with these restriction enzymes, note the single-stranded “sticky” ends generated by *EcoRI* versus the “blunt” ends generated by *EcoRV*.

leaving short, single-stranded overhangs called “sticky” ends at the termini of the two fragments. Other restriction enzymes such as *EcoRV* cut both strands of the DNA directly opposite each other, resulting in blunt ends (Figure 12.7). If the source DNA and the vector are both cut with the *same* restriction enzyme that yields complementary sticky ends, the two molecules can be joined (annealed) using *DNA ligase*, an enzyme that covalently links the strands of the vector and the source DNA. If the source DNA is PCR generated, DNA ligase is used to join the amplified DNA to specialized vectors (see Figure 12.9a).

In the final step of gene cloning, recombinant DNA molecules are introduced into suitable host organisms where they can replicate. But in practice, this often yields a mixture of *recombinant* constructs, where only some of the cells contain the desired cloned gene. To identify a host colony containing the correct recombinant DNA, one can select host cells expressing a vector-encoded marker such as antibiotic resistance. Colonies can then be screened for recombinant vectors by looking for the inactivation of a vector gene due to insertion of foreign DNA (see Figure 12.8).

Cloning Vectors

Several types of cloning vectors exist, including viruses, cosmids, and artificial chromosomes, and their use is dependent on the size of the DNA fragment to be cloned and the host in which the vector will be inserted. Plasmids are widely used cloning vectors, and the plasmid pUC19 (Figure 12.8) is a good example. This plasmid possesses an ampicillin resistance gene for selection and a blue–white color-screening system to select for recombinants. It also contains a short segment of artificial DNA containing cut sites for many different restriction enzymes, called a *multiple cloning site* (MCS), inserted into the *lacZ* gene encoding the lactose-degrading enzyme β -galactosidase (see Section 6.4 and Figure 6.14). The presence of the short MCS does not inactivate *lacZ*, and cut sites for restriction enzymes present in the MCS are absent from the rest of the vector.

The use of pUC19 in gene cloning is shown in Figure 12.8. A suitable restriction enzyme with a cut site within the MCS is chosen, and both the vector and the foreign DNA to be cloned are cut with this enzyme. The vector is linearized, and segments of the foreign DNA are inserted into the open cut site and ligated into position with the enzyme DNA ligase. This insertion disrupts the *lacZ* gene—a phenomenon called *insertional inactivation*—and is used to detect the presence of foreign DNA within the vector or recombinant vector. After DNA ligation, the resulting plasmids are transformed into cells of *Escherichia coli* and the cells plated on media containing both ampicillin (to select for cells containing the plasmid) and a lactose analog called *X-gal*, to detect β -galactosidase activity. *X-gal*, which is colorless, can be cleaved by β -galactosidase to generate a blue product. Thus, cells containing the vector *without* cloned DNA form blue colonies (that is, β -galactosidase is active), whereas cells containing the vector *with* an insert of cloned DNA do not form β -galactosidase and are therefore white and are the focus of further analyses.

Plasmids developed specifically for cloning DNA products synthesized by *Taq* polymerase in a polymerase chain reaction (PCR; Section 12.1) have also been designed (Figure 12.9a). The enzymatic activity of *Taq* polymerase adds a template-independent

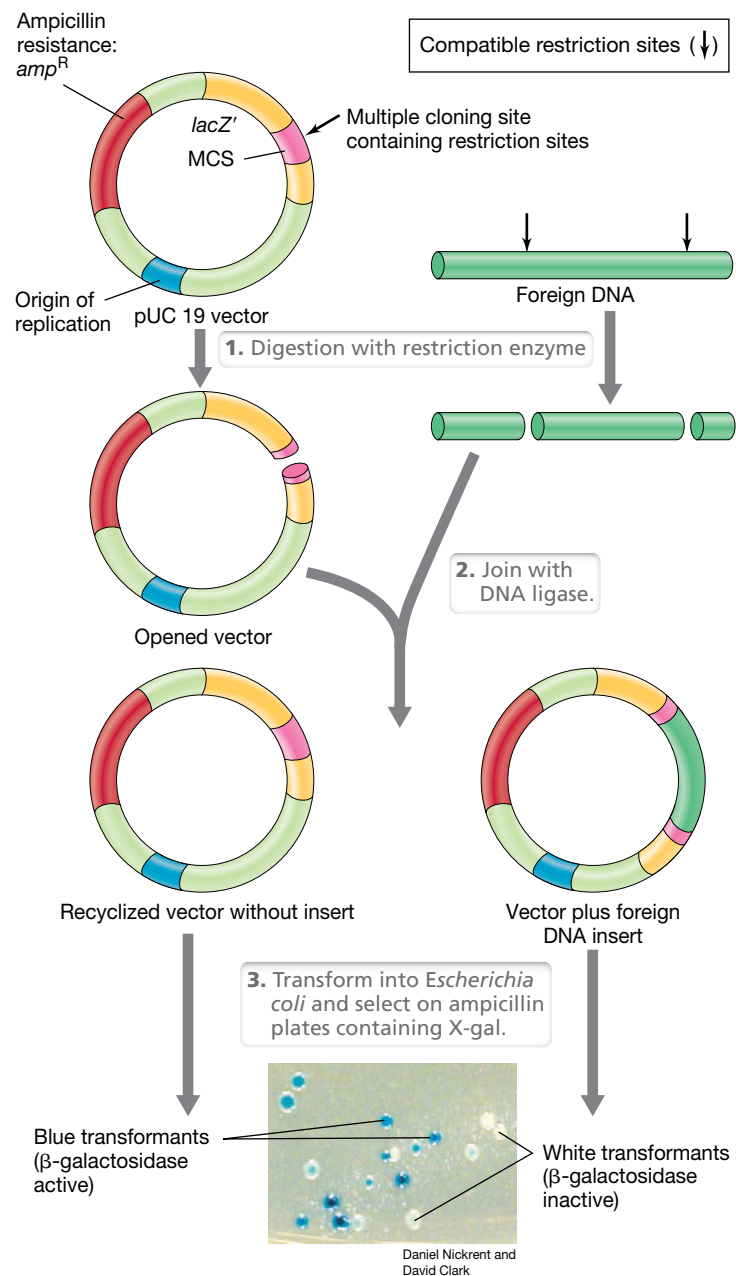


Figure 12.8 Cloning into the plasmid vector pUC19. Essential features include an ampicillin resistance marker and the multiple cloning site (MCS) with multiple restriction enzyme cut sites. The cloning vector and foreign DNA are cut with compatible restriction enzymes at positions indicated by the arrows. Insertion of DNA within the MCS inactivates β -galactosidase, allowing blue–white screening for the presence of the insert. The photo on the bottom shows colonies of *Escherichia coli* on an X-gal plate. The enzyme β -galactosidase can cleave the normally colorless X-gal to form a blue product.

adenosine residue to the 3' ends of its products. Linearized vectors are commercially available that contain overhanging thymidine residues that allow for base pairing with the *Taq* PCR product and subsequent ligation using DNA ligase (Figure 12.9a). For cloning genes into the yeast *Saccharomyces cerevisiae*, **yeast artificial chromosomes (YACs)** are frequently used (Figure 12.9b). YACs are linear vectors that replicate in yeast like normal chromosomes but have sites where very large fragments of DNA can be inserted.

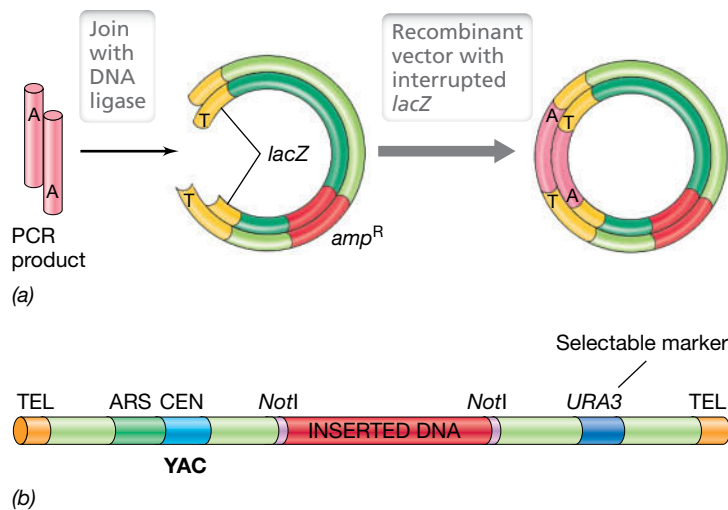


Figure 12.9 Specialized vectors. (a) PCR vector. The linearized cloning vector contains overhanging thymidine residues that base-pair with the adenosine residues present on the 3' ends of *Taq*-polymerase-generated PCR. Ligation of the two pieces of DNA yields a circular plasmid containing an interrupted *lacZ*. (b) A yeast artificial chromosome (YAC) containing foreign DNA. The foreign DNA was cloned into the vector at a *NotI* restriction site. The telomeres are labeled TEL and the centromere CEN. The origin of replication is labeled ARS (for autonomous replication sequence). The *URA3* gene is used for selection. The host into which the clone is transformed has a mutation in *URA3* and requires uracil for growth (*Ura*⁻). Host cells containing this YAC become *Ura*⁺. The diagram is not to scale; vector DNA is only 10 kbp whereas cloned DNA can be up to 800 kbp.

To function like normal eukaryotic chromosomes, YACs have an origin of DNA replication, telomeres for replicating DNA at the ends of the chromosome, and a centromere for segregation during mitosis. YACs also contain a cloning site and a gene for selection following transformation into the host (Figure 12.9b).

Hosts for Cloning Vectors

The most useful hosts for cloning are microorganisms that are easy to grow and transform with engineered DNA. They must also be genetically stable in culture and have the appropriate enzymes to allow replication of the vector. It is also helpful if considerable background information on the host and a wealth of tools for its genetic manipulation exist. Hosts that meet these conditions include the bacteria *E. coli* and *Bacillus subtilis*, and the yeast *S. cerevisiae* (Figure 12.10). Complete genome sequences are available for all of these organisms, and they are widely used as cloning hosts.

Although *E. coli* is found in the human intestine and some wild-type strains are potentially harmful (see Section 32.11), several modified *E. coli* strains have been developed specifically for cloning purposes. However, if cloned gene expression is desired, the outer membrane of this gram-negative bacterium (see Section 2.5) can hinder protein secretion. This issue can be overcome using the gram-positive bacterium *B. subtilis* as a cloning host (Figure 12.10). Cloning of DNA from eukaryotic sources into eukaryotic rather than prokaryotic cells is often done since eukaryotic hosts already possess the complex RNA and post-translational processing systems required for the production of eukaryotic proteins (see Section 4.6). Because it is easy to grow and manipulate, the

Bacteria		Eukaryote
<i>Escherichia coli</i>	<i>Bacillus subtilis</i>	<i>Saccharomyces cerevisiae</i>
Well-developed genetics Many strains available Most-studied bacterium	Easily transformed Nonpathogenic Naturally secretes proteins Endospore formation simplifies culture	Well-developed genetics Nonpathogenic Can process eukaryotic mRNAs Easy to grow
Potentially pathogenic Periplasm traps proteins	Genetically unstable Genetics less developed than in <i>E. coli</i>	Plasmids unstable Will not replicate most bacterial plasmids
Advantages		Disadvantages

Figure 12.10 Hosts for molecular cloning. A summary of the advantages and disadvantages of some common cloning hosts.

workhorse for cloning in eukaryotic cells is the yeast *S. cerevisiae*. However, some cloning applications require the use of plant tissues, insect cell lines, or cultured mammalian cells. Regardless of eukaryotic host type, it is necessary to get cloned DNA into the host, and several methods including transfection (see Figure 12.20), microinjection, and electroporation (see Section 11.6) can be used.

MINIQUIZ

- What is the purpose of molecular cloning?
- What is a multiple cloning site, and what is insertional inactivation?
- When would it be beneficial to use a eukaryotic host for molecular cloning?

12.3 Expressing Foreign Genes in Bacteria

Once genes are cloned, they can be transcribed and translated (expressed) to produce their encoded proteins. Obstacles to the expression of genes from mammalian or other eukaryotic sources include the following: (1) The genes must be placed under control of a bacterial promoter; (2) any introns (see Section 4.6) must be removed; (3) codon usage (codon bias, see Section 4.9) may require edits to gene sequences; and (4) many eukaryotic proteins require host modification after translation to yield the active form and bacteria cannot perform most such modifications. Here we consider solutions to these challenges.

Transcription and Translation of Cloned Genes Using Expression Vectors

Expression vectors are designed to allow the experimenter to control the expression of cloned genes. However, the native promoter of a cloned gene may work poorly in a new host, and the

overproduction of foreign proteins may also damage the host cell. Therefore, it is important to regulate the expression of cloned genes. Typically, the regulation is at the transcriptional level, and in practice, high levels of transcription require strong promoters that bind RNA polymerase efficiently (⚡ Section 4.5). An example of this is the use of the bacteriophage T7 promoter and T7 RNA polymerase to regulate gene expression. When T7 infects *Escherichia coli*, it encodes its own RNA polymerase that recognizes only T7 promoters (⚡ Section 10.4). In T7 expression vectors, cloned genes are placed under control of the T7 promoter. To achieve this, the gene for T7 RNA polymerase must also be present in the cell under the control of an easily regulated system, such as *lac* (⚡ Section 6.2) (Figure 12.11). This is usually done by integrating the gene for T7 RNA polymerase with a *lac* promoter into the chromosome of a specialized host strain.

The BL21 series of *E. coli* host strains are especially designed to work with the pET series of T7 expression vectors (Figure 12.11). The cloned genes are expressed shortly after T7 RNA polymerase transcription has been switched on by a *lac* inducer, such as the chemical IPTG (⚡ Section 6.2). Because it recognizes only T7 promoters, the T7 RNA polymerase transcribes only the cloned genes. The T7 RNA polymerase is so highly active that it uses most of the RNA precursors, thereby limiting transcription to the cloned genes. Consequently, host genes that require host RNA polymerase are for the most part not transcribed and thus the cells stop growing; translation in such cells then yields primarily the protein of interest. The T7 control system is thus very effective for generating large amounts of a specific protein.

Expression vectors must also be designed to ensure that the mRNA produced is efficiently translated. To synthesize protein from an mRNA molecule, it is essential for the ribosomes to bind at the correct site and begin reading in the correct frame. In bacteria

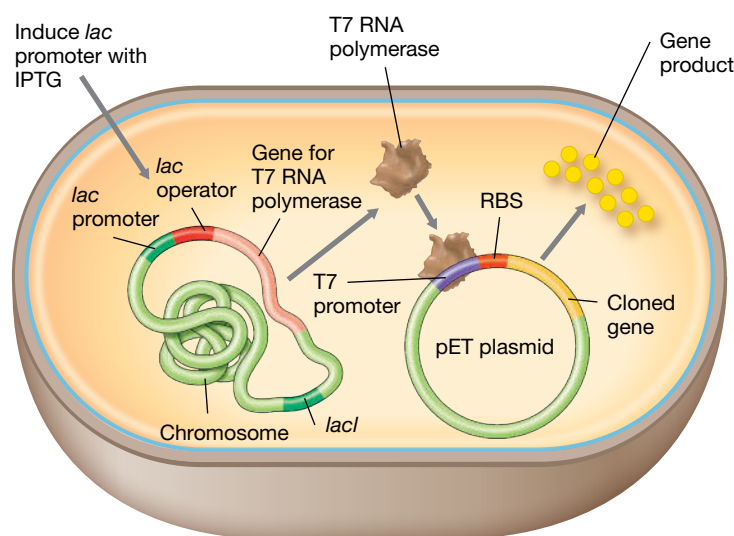


Figure 12.11 The T7 expression system. The gene for T7 RNA polymerase is in a gene fusion under control of the *lac* promoter and is inserted into the chromosome of a special host strain of *Escherichia coli* (BL21). Addition of IPTG induces the *lac* promoter, causing expression of T7 RNA polymerase. This transcribes the cloned gene, which is under control of the T7 promoter and is carried by the pET plasmid. RBS, ribosome-binding site.

this is accomplished by having a ribosome-binding site (RBS, ⚡ Section 4.9) and a nearby start codon on the mRNA. Bacterial RBSs are not found in eukaryotic genes and must be engineered into the vector if high levels of expression of the eukaryotic gene are to be obtained.

Other adjustments to a cloned gene may be necessary to ensure high-efficiency translation. For example, *codon usage* can be an obstacle. Codon usage is related to the concentration of the appropriate tRNA in the cell (⚡ Section 9.2 and Table 9.3). Because of the redundancy of the genetic code, more than one tRNA exists for most amino acids (⚡ Section 4.9). Therefore, if a cloned gene has a codon usage pattern distinct from that of its expression host, it will probably be translated inefficiently in that host. Site-directed mutagenesis (Section 12.4) can then be used to change selected codons in the gene, making it more amenable to the codon usage pattern of the host.

Cloning the Gene via mRNA or Artificial Synthesis

If a cloned gene contains introns, as eukaryotic genes typically do (⚡ Section 4.6), the correct protein product will not be made in a bacterial host unless modifications are made. This can be done via mRNA. In a typical mammalian cell, less than 5% of the total RNA is mRNA. However, eukaryotic mRNA is unique because of the poly(A) tails found at the 3' end (⚡ Section 4.6), and this makes it easy to isolate, even though it is of low abundance. If a cell extract is passed over a chromatographic column containing strands of poly(T) linked to a cellulose support, most of the mRNA separates from other RNAs by sticking to the support by specific pairing of As and Ts. The RNA is then released from the column by a low-salt buffer, which gives a preparation greatly enriched in mRNA.

Once mRNA has been isolated, the genetic information is converted into complementary DNA (cDNA) by RT-PCR as was illustrated in Figure 12.2. This double-stranded cDNA contains the coding sequence but lacks introns (Figure 12.12), and thus it can be inserted into a plasmid or other vector for cloning. However, because the cDNA contains only coding sequences, it lacks a promoter and other upstream regulatory sequences necessary for expression. Thus expression vectors containing bacterial promoters and ribosome-binding sites are used to obtain high-level expression of genes cloned in this way (see Figure 12.13).

For small proteins it is possible to artificially synthesize the entire gene (Section 12.4). Many mammalian proteins such as high-value peptide hormones are made by protease cleavage of large precursor molecules. Thus, in order to produce a short peptide such as insulin in its active form, construction and cloning of an artificial gene that encodes just the final hormone rather than the larger precursor protein from which it was derived may have several advantages. Constructed genes are naturally free of introns and thus the mRNA does not need processing. Also, promoters and other regulatory sequences can be inserted into the gene upstream of the coding sequences, and codon bias (⚡ Sections 4.9 and 9.2) can be adjusted to best suit the expression host.

Protein Stability and Purification

The synthesis of a protein in a new host may spawn additional problems. For example, some proteins are susceptible to degradation by protease enzymes and others may be toxic to their host.

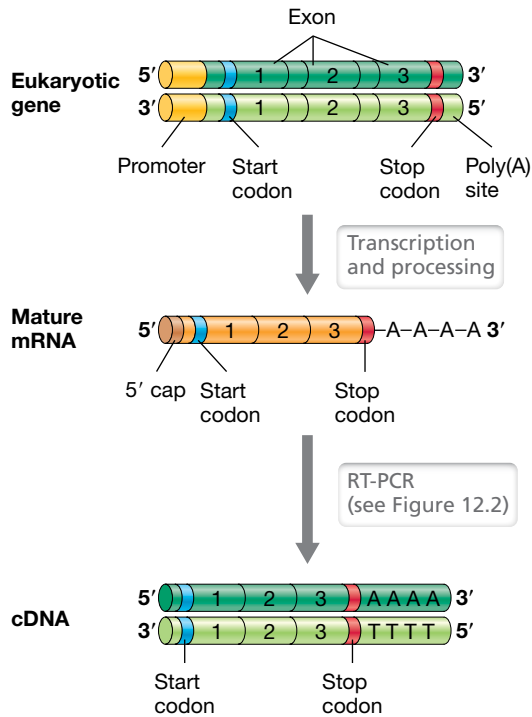


Figure 12.12 Complementary DNA (cDNA). Steps illustrating the synthesis of an intron-lacking cDNA corresponding to a eukaryotic gene generated by reverse transcription PCR (RT-PCR).

Also, when proteins are massively overproduced, they sometimes aggregate into insoluble inclusions. Although inclusions are relatively easy to purify, the protein they contain is often difficult to solubilize and may be partially denatured. Protein purification can be simplified if the target protein is made as a *fusion protein* along with a carrier protein encoded by the vector. To do this, the two genes are fused to yield a single coding sequence. A short segment that is recognized and cleaved by a commercially available protease is included between them. After transcription and translation, a single protein is made that is purified by methods designed for the carrier protein. The fusion protein is then cleaved by the protease to release the target protein from the carrier protein. Fusion proteins simplify purification of the target protein because a carrier protein is chosen that will not form inclusions and is easy to purify.

Several vectors are available to generate fusion proteins, and **Figure 12.13** shows an example of a fusion vector that is also an expression vector. In this example, the carrier protein is the *E. coli* maltose-binding protein (encoded by *malE*, **Figure 12.13**), a protein that is easily purified by methods based on its high affinity for maltose. Once purified, the two portions of the fusion protein are separated by protease or chemical treatment. One other advantage of making a fusion protein is that the carrier protein can be chosen to contain the bacterial *signal sequence*, a peptide rich in hydrophobic amino acids that enables transport of the protein across the cytoplasmic membrane (see Section 4.12). This makes possible a bacterial expression system that not only *makes* and *secretes* mammalian proteins, but also allows for the heterologously expressed protein to be separated from all of the other

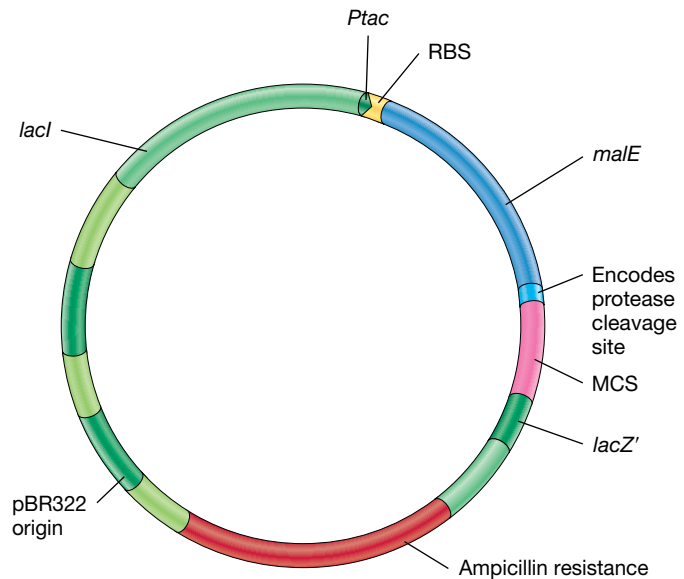


Figure 12.13 An expression vector for gene fusions. The gene to be cloned is inserted into the multiple cloning site (MCS) so it is in frame with the *malE* gene, which encodes maltose-binding protein. The insertion inactivates the gene for the alpha fragment of *lacZ*, which encodes β -galactosidase. The fused gene is under control of the hybrid *tac* promoter (*P_{tac}*) and an *Escherichia coli* ribosome-binding site (RBS). The plasmid also contains the *lacI* gene, which encodes the *lac* repressor. Therefore, an inducer must be added to turn on the *tac* promoter. The plasmid contains a gene conferring ampicillin resistance on its host.

proteins secreted by the cell using binding resins specific for the maltose-binding protein. Thus, carrier proteins can be used to save time, money, and effort in obtaining a desired product.

MINIQUIZ

- How can the bacteriophage T7 promoter be used to control expression of a eukaryotic gene in *Escherichia coli*?
- What major advantage does cloning mammalian genes from mRNA or using synthetic genes have over PCR amplification and cloning of the native gene?
- How is a fusion protein made?

12.4 Molecular Methods for Mutagenesis

Conventional mutagens introduce mutations *at random* in the DNA of the intact organism (see Section 11.4). In contrast, **site-directed mutagenesis** (also called *in vitro mutagenesis*) uses synthetic DNA plus DNA cloning techniques to introduce mutations into genes *at precisely determined sites*. In addition to changing one or just a few bases, mutations may also be engineered by inserting large segments of DNA at precisely determined locations.

Site-Directed Mutagenesis

Site-directed mutagenesis requires that short sequences of DNA (*oligonucleotides*) of precise sequence be available, and these are chemically synthesized; primers or probes for use in the polymerase chain reaction and hybridization (Section 12.1) are also made in this way. Oligonucleotides of 12–40 bases are inexpensive

and commercially available, and oligonucleotides of over 100 bases in length can be made if necessary. Site-directed mutagenesis then allows any base pair in a specific gene to be changed. When the mutated gene is expressed, a protein with an altered amino acid sequence will be produced. Site-directed mutagenesis can thus be used to manipulate proteins to test the functional importance of specific amino acids.

One procedure for site-directed mutagenesis is illustrated in **Figure 12.14**. A cloned target gene is denatured to form single-stranded DNA and then allowed to hybridize with the mutagenized oligonucleotide containing a one-base mismatch. After extension by DNA polymerase, the complementary DNA strand formed will contain the mismatch. After transformation of the vector into host cells followed by its semiconservative replication and subsequent cell division, one daughter cell will carry the mutation while the other will be wild type. Progeny bacteria are then screened for those carrying the mutation.

Site-directed mutagenesis may also be carried out using PCR. In this case, the short DNA oligonucleotide with the required mutation is used as a PCR primer. The mutation-carrying primer is designed to anneal to the target with the mismatch in the middle and must have enough matching nucleotides on both sides for binding to be stable during the PCR reaction. The mutant primer is then paired with a normal primer, and when the PCR reaction amplifies the target DNA, it incorporates the mutation(s) into the final amplified product.

Site-directed mutagenesis has many applications. The technique has been widely used by enzymologists to change a specific amino acid in the active site of an enzyme to see how the modified enzyme compares with the wild-type enzyme. In such experiments, the vector encoding the mutant enzyme is inserted into a mutant host strain unable to make the original enzyme. Consequently, the activity measured is due to the mutant version of the enzyme alone. Using *in vitro* mutagenesis, enzymologists can link virtually any aspect of an enzyme's activity—catalysis, resistance, susceptibility to chemical or physical agents, interactions with other proteins—to specific amino acids in the enzyme. In genetic engineering, site-directed mutagenesis has been used to improve the properties of specific proteins, and we discuss some examples in Section 12.6.

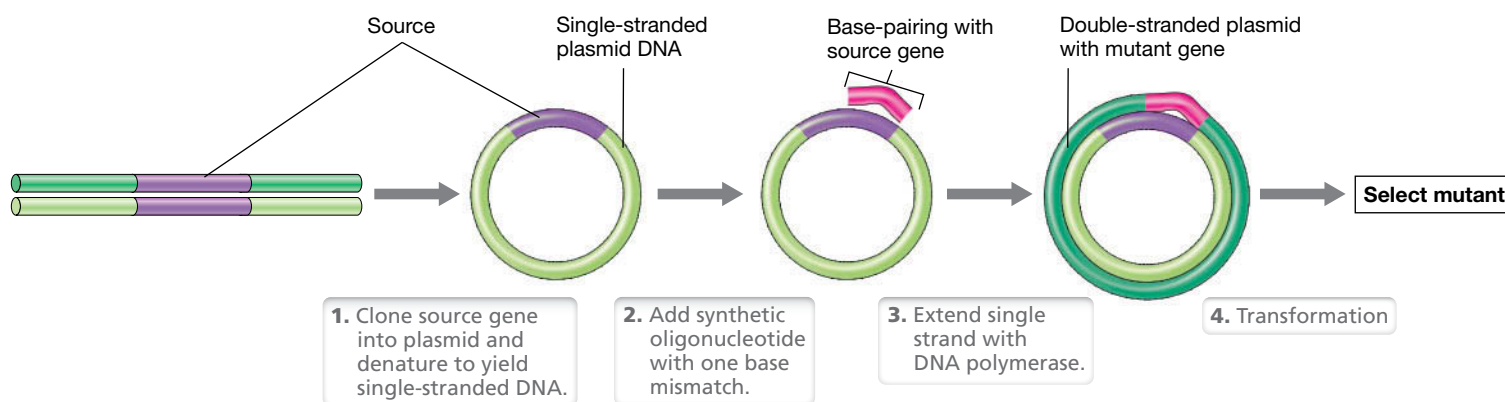


Figure 12.14 Site-directed mutagenesis using synthetic DNA. Short synthetic oligonucleotides hybridized to the cloned gene may be used to generate mutations. Cloning the source DNA into a plasmid followed by denaturation yields the single-stranded DNA needed for site-directed mutagenesis to work.

Cassette Mutagenesis and Gene Disruption

To make more than a few base-pair changes or replace sections of a gene of interest, synthetic fragments called **DNA cassettes** (or cartridges) can be used to mutate DNA in a process known as **cassette mutagenesis**. These cassettes can be synthesized using the polymerase chain reaction or by direct DNA synthesis. The cassette can then replace sections of the DNA of interest using restriction sites. However, if sites for the appropriate restriction enzyme are not present at the required location, they can be inserted by site-directed mutagenesis (Figure 12.14). Cassettes used to replace sections of genes are typically the same size as the wild-type DNA fragments they replace.

Another type of cassette mutagenesis is called **gene disruption**. In this technique, cassettes are inserted into the middle of a gene, thus disrupting the coding sequence (**Figure 12.15**). Cassettes used for making insertion mutations can be almost any size and can even carry an entire gene. To facilitate selection, cassettes that encode antibiotic resistance are commonly used. For example, a DNA cassette containing a gene conferring kanamycin resistance is inserted at a restriction site in a cloned gene. The vector carrying the disrupted gene is then converted from a circular to a linear form by cutting it with a different restriction enzyme. Finally, the linear DNA is transformed into the host and kanamycin resistance is selected. The linear plasmid cannot replicate, and so resistant cells arise mostly by homologous recombination (↔ Section 11.5) between the mutated gene on the plasmid and the wild-type gene on the chromosome (Figure 12.15).

When a cassette is inserted, the cells not only gain antibiotic resistance, but they also *lose the function of the gene* into which the cassette is inserted. Such mutations are called *knockout mutations* and are widely used in biology. Knockouts are similar to insertion mutations made by transposons (↔ Section 11.11), but here the experimenter chooses which gene will be mutated. Knockout mutations in haploid organisms yield viable cells only if the disrupted gene is nonessential. Thus, gene knockouts are commonly used for determining whether a gene of interest is essential.

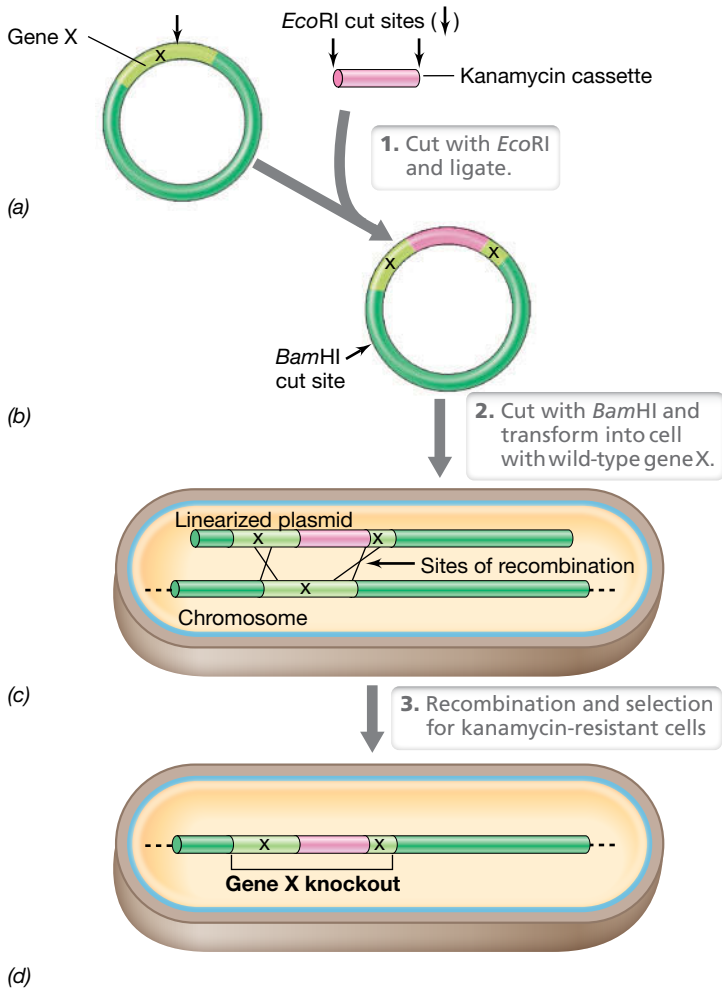


Figure 12.15 Gene disruption by cassette mutagenesis. (a) A cloned wild-type copy of gene X, carried on a plasmid, and a kanamycin cassette are cut with *Eco*RI and mixed. (b) The cut plasmid and the cassette are ligated, creating a plasmid with the kanamycin cassette as an insertion mutation within gene X. This new plasmid is cut with *Bam*HI and transformed into a cell. (c) The transformed cell contains the linearized plasmid with a disrupted gene X and its own chromosome with a wild-type copy of the gene. (d) In some cells, homologous recombination occurs between the wild-type and mutant forms of gene X. Cells that can grow in the presence of kanamycin have only a single, disrupted copy of gene X.

MINIQUIZ

- How can site-directed mutagenesis be useful to enzymologists?
- What is used to alter more than a few base pairs in a gene of interest?
- What are knockout mutations?

12.5 Reporter Genes and Gene Fusions

DNA manipulation has revolutionized the study of gene regulation, and *gene fusions* have been a major tool for studying regulatory events. In a reporter gene fusion, a coding sequence from one source (the *reporter*) is fused to a regulatory region from another source to form a hybrid gene. Regulation of gene expression is then studied by assaying for the product of the

reporter as a function of different conditions sensed by the regulator.

Reporter Genes

The key property of a **reporter gene** is that it encodes a protein that is easy to detect and assay. Reporter genes are used for a variety of purposes. They may be used to report the presence or absence of a particular genetic element (such as a plasmid) or DNA inserted within a vector. They can also be fused to other genes or to the promoter of other genes so that gene expression can be studied (see Figure 7.2).

The first widely used reporter gene was *lacZ* from *Escherichia coli*, a gene that encodes the enzyme β -galactosidase, required for lactose catabolism (see Section 6.2). Cells expressing β -galactosidase can be detected easily by the color of their colonies on indicator plates that contain the artificial substrate X-gal (5-bromo-4-chloro-3-indolyl- β -D-galactopyranoside); X-gal is cleaved by β -galactosidase to yield a blue color (see Figure 12.8).

The **green fluorescent protein (GFP)** is widely used as a reporter (Figure 12.16). Although the gene for GFP was originally cloned from the jellyfish *Aequorea victoria*, GFP may be expressed in most cells as it is stable and causes little or no disruption of host cell metabolism. If expression of a cloned gene is linked to the expression of GFP, the latter signals (reports) that the cloned gene has also been expressed (Figure 12.16). Since the advent of GFP, many similar but differently colored fluorescent proteins have been developed as reporters (see Section 7.1).

Gene Fusions

Gene fusions are genetic constructs that consist of segments from two different genes. If the promoter that controls a coding sequence is removed, the coding sequence can be fused to a different regulator to place the gene under the control of a different promoter. Alternatively, the promoter region can be fused to a gene

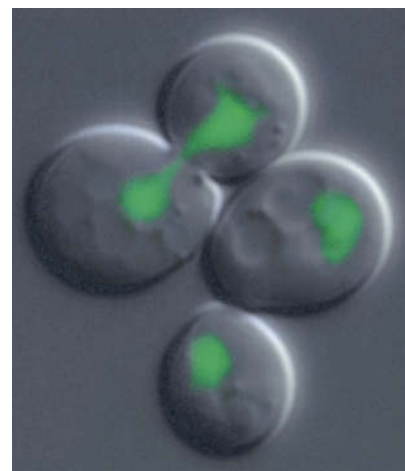


Figure 12.16 Green fluorescent protein (GFP). GFP can be used as a tag for protein localization in vivo. In this example, the gene encoding Pho2, a DNA-binding protein from the yeast *Saccharomyces cerevisiae*, was fused to the gene encoding GFP and photographed by fluorescence microscopy. The recombinant gene was transformed into budding yeast cells. These expressed the fluorescent fusion protein localized in the nucleus.

whose product is easy to assay. There are two different types of gene fusions. In **operon fusions**, a coding sequence that retains its own translational start site and signals is fused to the transcriptional signals of another gene. In **protein fusions**, genes that encode two different proteins are fused together so that they share the same transcriptional and translational start and stop signals. Following translation, protein fusions yield a single hybrid polypeptide (Section 12.3).

Gene fusions are often used in studying gene regulation, especially if measuring the levels of the natural gene product is difficult, expensive, or time consuming. The regulatory region of the gene of interest is fused to the coding sequence for a reporter gene, such as that for β -galactosidase or GFP. The reporter is then made under the conditions that would trigger expression of the target gene (Figure 12.17). The expression of the reporter is assayed under a variety of conditions to determine how the gene of interest is regulated (Chapter 6). *Transcriptional control* is assayed by fusing the transcriptional start signals of the gene of interest to a reporter gene, whereas *translational control* is assayed by fusing translational start signals of a gene of interest to a reporter gene under the control of a known promoter.

Gene fusions may also be used to test for the effects of regulatory genes. Mutations that affect regulatory genes are introduced into cells carrying gene fusions, and expression is measured and compared to cells lacking the regulatory mutations. This allows the rapid screening of multiple regulatory genes that are suspected of controlling the target gene. Besides the use of fusions to monitor for the presence or expression of a gene, proteins that are easily purified can also be fused to proteins of interest to aid in purification of the latter (Section 12.3).

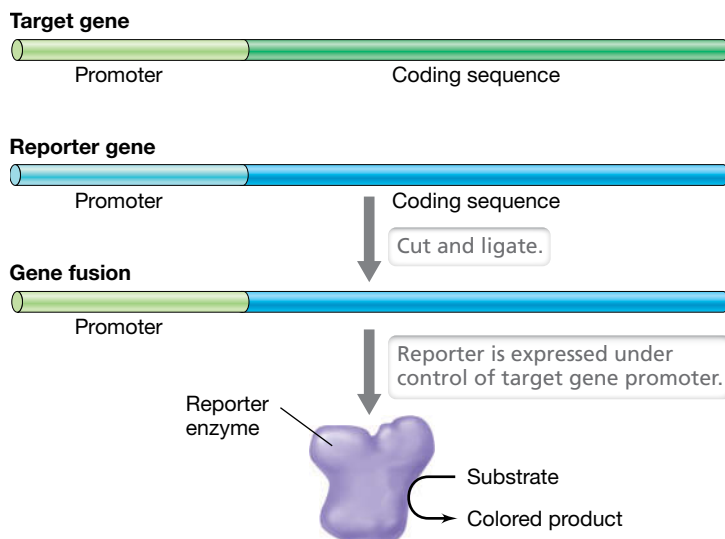


Figure 12.17 Construction and use of gene fusions. The promoter of the target gene is fused to the reporter coding sequence. Consequently, the reporter gene is expressed under those conditions where the target gene would normally be expressed. The reporter shown here is an enzyme (such as β -galactosidase) that converts a substrate to a colored product that is easy to detect. This approach greatly facilitates the investigation of regulatory mechanisms.

MINIQUIZ

- What is a reporter gene? The product of which reporter gene yields a green color?
- Why are gene fusions useful in studying gene regulation?

II • Making Products from Genetically Engineered Microbes: Biotechnology

Genetic engineering can transform microorganisms into tiny factories for the production of valuable products including fuels, chemicals, drugs, and human hormones, such as insulin. This is the science of **biotechnology**. Up to this point we have only considered the techniques used for manipulating, cloning, and expressing DNA. We now consider how these techniques are applied in biotechnology to produce valuable proteins and genetically modified plants, animals, vaccines, and metabolic pathways.

12.6 Somatotropin and Other Mammalian Proteins

One of the most economically profitable areas of biotechnology has been the production of human proteins. Many mammalian proteins have high pharmaceutical value but are typically present in very low amounts in normal tissue, and it is therefore extremely costly to purify them. Even if the protein can be produced in cell culture, this is much more expensive and difficult than growing microbial cultures that produce the protein in high yield. Therefore, the biotechnology industry has developed genetically engineered microorganisms to produce many different mammalian proteins.

Somatotropin

Although insulin was the first human protein to be produced by bacteria, the genetic engineering required was complicated because insulin consists of two short polypeptides held together by disulfide bonds. A more straightforward example is *human somatotropin* (growth hormone), which consists of a single polypeptide encoded by a single gene; a deficiency of somatotropin in the body results in hereditary dwarfism. Because the human somatotropin gene has been successfully cloned and expressed in bacteria, children showing stunted growth can be treated with *recombinant human somatotropin* to correct this. However, some forms of dwarfism are caused by a lack of the somatotropin receptor, and in such cases, administration of somatotropin has no effect.

The human somatotropin gene was cloned as complementary DNA (cDNA) from mRNA as described in Section 12.3 (see Figure 12.18). The cDNA was then expressed in a bacterial expression vector. The main problem with producing relatively short polypeptide hormones such as somatotropin is their susceptibility to protease digestion, but this problem was overcome by using bacterial host strains lacking key protease enzymes. Today recombinant

human growth hormone taken by injection is marketed under several brand names in the United States and has successfully treated thousands of children afflicted with any of several different syndromes that result in short stature. Recombinant somatotropin has also been used to treat some cases of tissue atrophy in adults. However, use in adults is not a common practice, and growth hormone is banned by the International Olympic Committee and by some professional sports leagues for its alleged performance-enhancing capabilities.

Recombinant bovine somatotropin (rBST) is used in the dairy industry (Figure 12.18). Injection of rBST into cows does not make them grow larger but instead stimulates milk production. This is because somatotropin has two binding sites; one is the somatotropin receptor that stimulates growth while the other is the prolactin receptor that promotes milk production. Thus, cows treated with rBST produce more milk. However, when human somatotropin is used to treat short stature conditions, it is desirable to avoid any side effects from the hormone's prolactin activity. To alleviate this problem, site-directed mutagenesis (Section 12.4) of the human somatotropin gene was used to alter those amino acids of somatotropin that bind to the prolactin receptor, thus ensuring that the hormone would only target growth. As this example

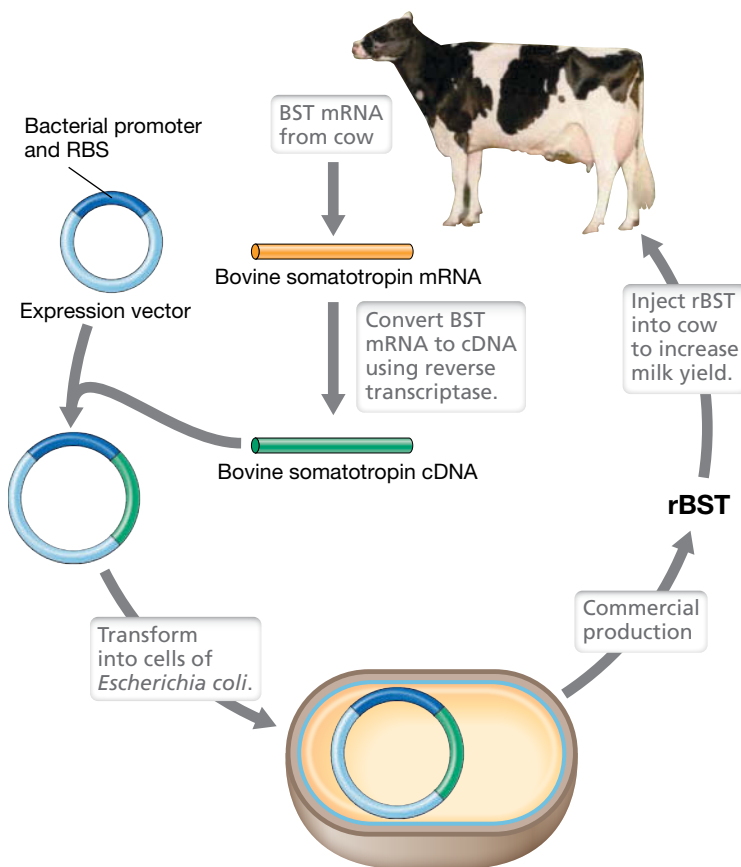


Figure 12.18 Cloning and expression of bovine somatotropin. The mRNA for bovine somatotropin (BST) is obtained from a cow, and the mRNA is converted to cDNA by reverse transcriptase. The cDNA version of the somatotropin gene is then cloned into a bacterial expression vector that has a bacterial promoter and ribosome-binding site (RBS). The construct is transformed into cells of *Escherichia coli*, and recombinant bovine somatotropin (rBST) is produced. Milk production increases in cows treated with rBST.

shows, it is possible not only to make genuine human hormones but also to alter their specificity and activity to make them better pharmaceuticals.

Other Mammalian Proteins

Many other mammalian proteins are produced today by genetic engineering (Table 12.1). These include, in particular, an assortment of hormones and proteins for blood clotting and other blood processes. For example, *tissue plasminogen activator* (TPA) is a protein that dissolves blood clots in the bloodstream that may form in the final stages of the healing process. TPA is primarily used in heart patients or others suffering from poor circulation to prevent the development of clots that can be life-threatening. Heart disease is a leading cause of death in many developed countries, especially in the United States, so microbially produced TPA is in high demand.

In contrast to TPA, the blood clotting factors VII, VIII, and IX are critically important for the *formation* of blood clots. Hemophiliacs suffer from a deficiency of one or more clotting factors and can

TABLE 12.1 A few human medical products made by genetic engineering

Product	Function
Blood proteins	
Erythropoietin	Treats certain types of anemia
Factors VII, VIII, IX	Promotes blood clotting
Tissue plasminogen activator	Dissolves blood clots
Urokinase	Promotes blood clotting
Human hormones	
Epidermal growth factor	Wound healing
Follicle-stimulating hormone	Treatment of reproductive disorders
Insulin	Treatment of diabetes
Nerve growth factor	Treatment of degenerative neurological disorders and stroke
Relaxin	Facilitates childbirth
Somatotropin (growth hormone)	Treatment of some growth abnormalities
Immune modulators	
α -Interferon	Antiviral, antitumor agent
β -Interferon	Treatment of multiple sclerosis
Colony-stimulating factor	Treatment of infections and cancer
Interleukin-2	Treatment of certain cancers
Lysozyme	Anti-inflammatory
Tumor necrosis factor	Antitumor agent, potential treatment of arthritis
Replacement enzymes	
β -Glucocerebrosidase	Treatment of Gaucher disease, an inherited neurological disease
Therapeutic enzymes	
Human DNase I	Treatment of cystic fibrosis
Alginate lyase	Treatment of cystic fibrosis

therefore be treated with microbially produced clotting factors. In the past hemophiliacs have been treated with clotting factor extracts from pooled human blood, some of which was contaminated with viruses such as HIV and hepatitis C, putting hemophiliacs at high risk for contracting AIDS, hepatitis, or liver cancer. Recombinant clotting factors have eliminated this problem.

Some mammalian proteins made by genetic engineering are enzymes rather than hormones (Table 12.1). For instance, *human DNase I* is used to treat the buildup of DNA-containing mucus in the lungs of patients with cystic fibrosis. The mucus forms because cystic fibrosis is often accompanied by life-threatening lung infections by the bacterium *Pseudomonas aeruginosa*. The bacterial cells form biofilms (↔ Sections 7.9 and 20.4) within the lungs that make drug treatment difficult. DNA is released when the bacteria lyse, and this fuels mucus formation, making it difficult to breathe. DNase digests the DNA and greatly decreases the viscosity of the mucus.

MINIQUIZ

- What is the advantage of using genetic engineering to make insulin?
- What are the major problems when manufacturing proteins in bacteria?
- Explain how an enzyme can be useful in treating a bacterial infection, such as that which occurs with cystic fibrosis.

12.7 Transgenic Organisms in Agriculture and Aquaculture

Genetic improvement of plants and animals by traditional selection and breeding has a long history, but recombinant DNA technology has led to revolutionary changes. Although the genetic engineering of higher organisms is not truly microbiology, much of the DNA manipulation is carried out using bacteria and their plasmids and genes. Hence, we consider the genetic manipulation

of plants and animals here with a focus on the microbiology that supported it.

Because genetically engineered plants or animals contain a gene from another organism—called a *transgene*—they are **transgenic organisms**. The public knows these as **genetically modified organisms (GMOs)**. Strictly speaking, the term *genetically modified* refers to genetically engineered organisms whether or not they contain foreign DNA. In this section we discuss how foreign genes are inserted into plant and fish genomes and how transgenic organisms may be used.

The Ti Plasmid and Transgenic Plants

While recombinant DNA can be transformed into plant cells by electroporation or transfection (see Figure 12.20), the **Ti plasmid** from the gram-negative bacterium *Agrobacterium tumefaciens*, a plant pathogen, can be used to transfer DNA directly into the cells of certain plants. This plasmid is responsible for *A. tumefaciens* virulence and encodes genes that mobilize DNA for transfer to the plant, which as a result contracts crown gall disease (↔ Section 23.5). The segment of the Ti plasmid DNA that is actually transferred to the plant is called **T-DNA**. The sequences at the ends of the T-DNA are essential for transfer, and the foreign DNA to be transferred must reside between these ends.

One common Ti-vector system that has been used for the transfer of genes to plants is a two-plasmid system called a *binary vector*, which consists of a cloning vector plus a helper plasmid. The cloning vector contains the two ends of the T-DNA flanking a multiple cloning site, two origins of replication so that it can replicate in both *Escherichia coli* (the host for cloning) and *A. tumefaciens*, and two antibiotic resistance markers, one for selection in plants and the other for selection in bacteria. The foreign DNA is inserted into the vector, which is transformed into *E. coli* and then moved to *A. tumefaciens* by conjugation (Figure 12.19).

This cloning vector lacks the genes needed to transfer T-DNA to a plant. However, when placed in an *A. tumefaciens* cell that contains a suitable helper plasmid, the T-DNA can be transferred to a plant. The “disarmed” helper plasmid, called *D-Ti*, contains the

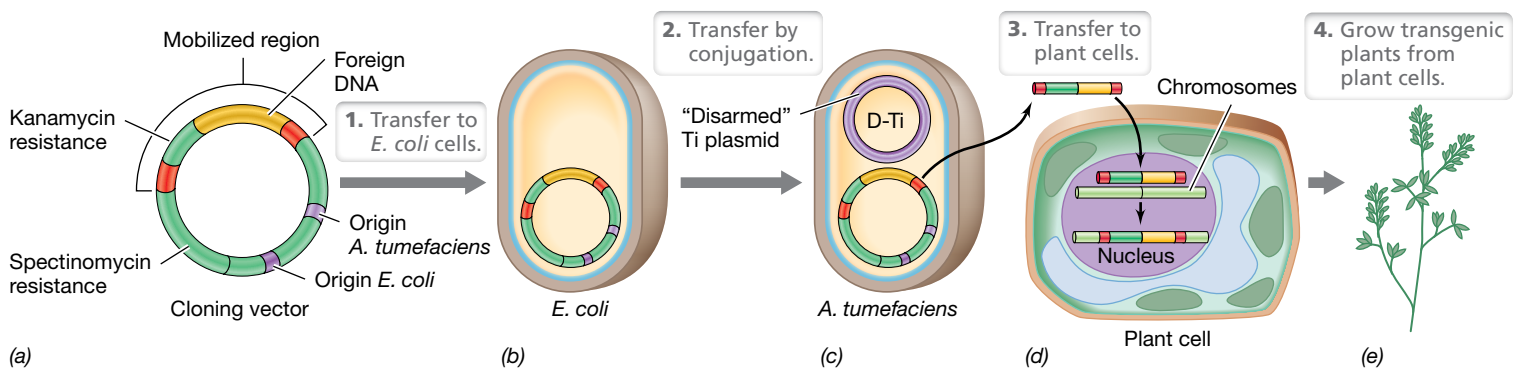


Figure 12.19 Production of transgenic plants using a binary vector system in *Agrobacterium tumefaciens*. (a) Plant cloning vector containing ends of T-DNA (red), foreign DNA, origins of replication, and resistance markers. (b) The vector is put into cells of *Escherichia coli* for cloning and then (c) transferred to *A. tumefaciens* by conjugation. The resident Ti plasmid (D-Ti) has been genetically engineered to remove key pathogenesis genes. (d) D-Ti can still mobilize the T-DNA region of the vector for transfer to plant cells grown in tissue culture. (e) From the recombinant plant cell, a whole plant can be grown. Details of Ti plasmid transfer from bacterium to plant are shown in Figure 23.25.

virulence (*vir*) region of the Ti plasmid but lacks the T-DNA. It also lacks the genes that initiate disease but supplies all the functions needed to transfer the T-DNA from the cloning vector. The cloned DNA and the kanamycin resistance marker of the vector are mobilized by D-Ti and transferred into a plant cell where they enter the nucleus (Figure 12.19*d*). Following integration into a plant chromosome, the foreign DNA can be expressed and confer new properties on the plant.

A number of transgenic plants have been produced using the Ti plasmid of *A. tumefaciens*. The Ti system works well with broad-leaf plants (dicots), including crops such as tomato, potato, tobacco, soybean, alfalfa, and cotton. It has also been used to produce transgenic trees, such as walnut and apple. The Ti system does not work with plants from the grass family (monocots, including the important crop plant corn), but other methods of introducing DNA, such as transfection by microprojectile bombardment with a particle gun (Figure 12.20), have been used successfully for them.

Herbicide- and Insect-Resistant Plants

Major areas targeted for genetic improvement in plants include herbicide, insect, and microbial disease resistance, as well as improved product quality. The main genetically modified (GM) crops today are soybeans, corn, cotton, and canola. Almost all the GM soybeans and canola planted were herbicide resistant, whereas the corn and cotton were herbicide resistant or insect resistant, or both.

Herbicide resistance is genetically engineered into a crop plant to protect it from herbicides applied to kill weeds. Many herbicides inhibit a key plant enzyme or protein necessary for growth. For example, the herbicide *glyphosate* (Roundup™, made by



Stephen R. Padgett, Monsanto Company

Figure 12.21 Transgenic plants: herbicide resistance. The photograph shows a portion of a field of soybeans that has been treated with Roundup™, a glyphosate-based herbicide manufactured by Monsanto Company (St. Louis, Missouri, USA). The remnants of plants on the right are normal soybeans; the plants on the left have been genetically engineered to be glyphosate resistant.

Monsanto) kills plants by inhibiting an enzyme necessary for making aromatic amino acids. Some bacteria contain an equivalent enzyme and are also killed by glyphosate. However, mutant bacteria were selected that were resistant to glyphosate and contained a resistant form of the enzyme. The gene encoding this resistant enzyme from *A. tumefaciens* was cloned, modified for expression in plants, and transferred into important crop plants, such as soybeans. When sprayed with glyphosate, plants containing the bacterial gene are not killed (Figure 12.21). Thus glyphosate can be used to kill weeds that compete for water and nutrients with the growing crop plants. Herbicide-resistant soybeans are now widely planted in the United States.

Transgenic plants resistant to damage by certain insects have been produced by genetic engineering (Figure 12.22). One widely used approach is based on introducing genes encoding the toxic proteins of the gram-positive, endospore-forming bacterium *Bacillus thuringiensis* into plants. As it sporulates, *B. thuringiensis* produces a crystalline protein called *Bt toxin* (see Section 16.8) that is toxic to moth and butterfly larvae. Many variants of Bt toxin exist that are specific for different insects. Certain strains of *B. thuringiensis* produce additional proteins toxic to beetle and fly larvae and mosquitoes.

The Bt transgene is normally inserted directly into the plant genome. For example, a natural Bt toxin gene was cloned into a plasmid vector under control of a chloroplast ribosomal RNA promoter and then transfected into tobacco plant chloroplasts by microprojectile bombardment (Figure 12.20). This yielded transgenic plants that expressed Bt toxin at levels that were extremely toxic to larvae of several insect species. Binding Bt triggers a change in the toxin's conformation that disrupts the insect digestive system, causing death. Bt toxin is harmless to mammals (including humans) because any toxin ingested is destroyed in the stomach and the specific Bt receptors in the insect intestine are absent from the intestines of other groups of organisms.

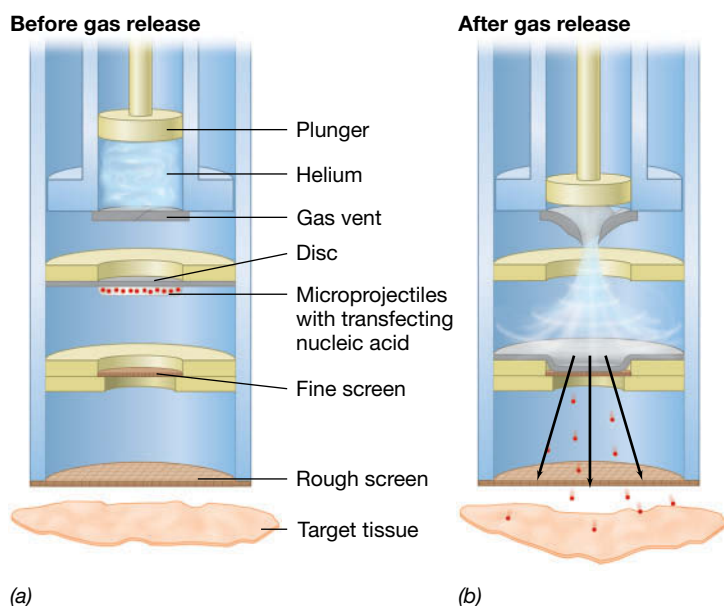


Figure 12.20 DNA gun for transfection of eukaryotic cells. The inner workings of the gun show how metal pellets coated with nucleic acids (microprojectiles) are projected at target cells. (a) Before firing and (b) after firing. A shock wave due to gas release throws the disc carrying the microprojectiles against the fine screen. The microprojectiles continue on into the target tissue.

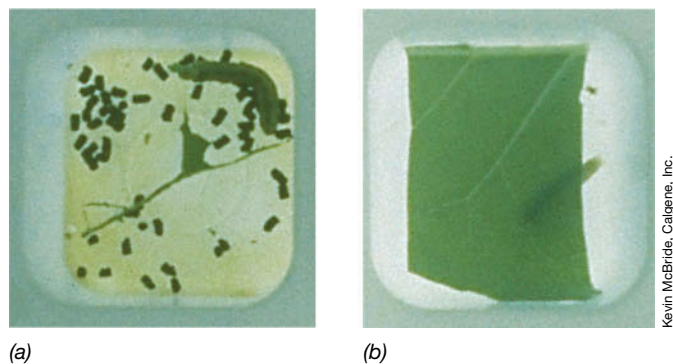


Figure 12.22 Transgenic plants: insect resistance. The results of an assay to determine the effect of beet armyworm larvae on tobacco leaves. (a) Leaf from wild-type plant. (b) Leaf from transgenic plant that expresses Bt toxin in its chloroplasts.

Transgenic Fish

Many foreign genes have been incorporated and expressed in laboratory research animals and in commercially important animals. The genetic engineering uses microinjection to deliver cloned genes to fertilized eggs; genetic recombination then incorporates the foreign DNA into the genomes of the eggs. More recently, farm animals and fish have been genetically modified to improve yields.

An interesting practical example of a transgenic animal is the *AquaAdvantage* salmon developed by AquaBounty Technologies (Figure 12.23). These transgenic salmon do not grow to be larger than normal salmon but simply reach market size much faster—18 months versus 3 years. The gene for growth hormone in native salmon is activated by light. Consequently, salmon grow rapidly only during the summer months. In the genetically engineered salmon, the growth promoter for the growth hormone gene was replaced with the promoter from another fish that grows at a more or less constant rate all year round. The result was salmon that make growth hormone continuously and thus grow faster. Such transgenic salmon can be grown commercially in aquaculture operations and harvested more quickly than with non-GMO farm-raised salmon.



Figure 12.23 Fast-growing transgenic salmon. The *AquaAdvantage*[™] salmon (top) was engineered by AquaBounty Technologies (Maynard, Massachusetts, USA). The transgenic and control fish are both 18 months old but weigh 4.5 kg and 1.2 kg, respectively.

In 1995 AquaBounty applied to the U.S. Food and Drug Administration for approval to distribute the fast-growing salmon. After two decades of debate regarding the potential risks of consuming genetically modified fish, final approval occurred in 2015. Thus the *AquaAdvantage* salmon is the first genetically engineered animal to be heading to the supermarket and is licensed to be sold without any GMO label.

MINIQUIZ

- What is a transgenic plant?
- Give an example of a genetically modified plant and describe how its modification benefits agriculture.
- How have transgenic salmon been engineered to reach market size faster?

12.8 Engineered Vaccines and Therapeutics

Genetic engineering is used to manufacture certain vaccines and medical therapeutics. Vaccines are substances that elicit immunity to a disease when injected into an animal (Section 27.2). Moreover, many pathogenic bacteria cause disease through their ability to infiltrate cells and release virulence factors such as toxins and destructive enzymes. Through genetic engineering, some of these activities have been harnessed to specifically target cancer cells. We consider both of these “medical miracles”—vaccines and engineered pathogens—here.

Recombinant Vaccines, Vaccinia Virus, and Subunit Vaccines

Genetic engineering can modify a pathogen by deleting genes that encode virulence factors (Section 25.3) while retaining those whose products elicit an immune response. This yields a recombinant and infective (but attenuated) vaccine. Conversely, one can add genes from a pathogenic virus to the genome of a relatively harmless virus, called a *carrier virus*. Such vaccines are called **vector vaccines** and induce immunity to the pathogenic virus. Indeed, one can even combine the two approaches by disarming one pathogen and adding back to it immunity-inducing genes from a second pathogen. This yields a **polyvalent vaccine**, a vaccine that immunizes against two different diseases at the same time.

Vaccinia virus (Section 10.6) is widely used to prepare recombinant vaccines for human use; however, cloning into vaccinia requires a selective marker, which is provided by the gene encoding the enzyme thymidine kinase. Vaccinia virus contains a gene encoding thymidine kinase, an enzyme that converts thymidine into thymidine triphosphate. However, this enzyme also converts the base analog 5-bromodeoxyuridine (BrdU) into a nucleotide that is incorporated into DNA, causing a lethal reaction. Thus, cells that express either host- or virus-encoded thymidine kinase are killed when treated with BrdU.

Genes to be put into vaccinia virus are first inserted into an *Escherichia coli* plasmid that contains part of the vaccinia thymidine kinase (*tdk*) gene (Figure 12.24). The foreign DNA is inserted into the *tdk* gene, which is therefore disrupted. This recombinant

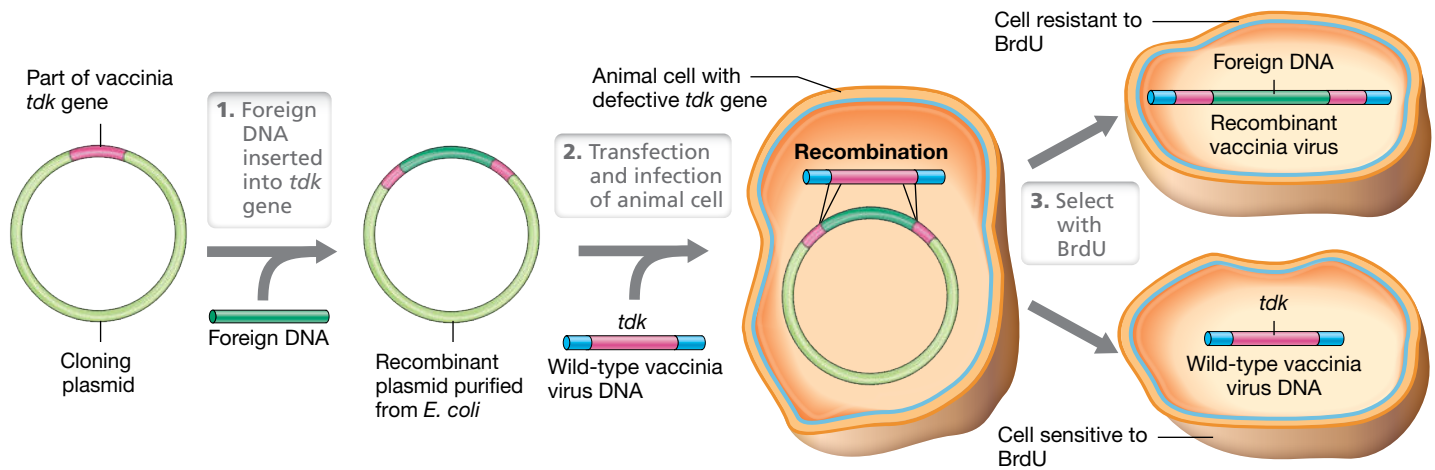


Figure 12.24 Production of recombinant vaccinia virus. Foreign DNA is inserted into a short segment of the thymidine kinase gene (*tdk*) from vaccinia virus carried on a plasmid. Following replication of this plasmid in *Escherichia coli*, both the recombinant plasmid and wild-type vaccinia virus are put into the same animal host cell to promote recombination. The animal cells are treated with 5-bromodeoxyuridine (BrdU), which kills only cells with an active thymidine kinase. Only recombinant vaccinia viruses whose *tdk* gene is inactivated by insertion of foreign DNA survive.

plasmid is then transformed into animal cells whose own *tdk* genes have been inactivated. These cells are also infected with wild-type vaccinia virus. The two versions of the *tdk* gene—one on the plasmid and the other on the virus—then recombine. Some viruses gain a disrupted *tdk* gene plus its foreign insert (Figure 12.24). Cells infected by wild-type vaccinia virus (with an active thymidine kinase) are killed by BrdU. By contrast, cells infected by recombinant vaccinia virus (with a disrupted *tdk* gene) grow long enough to yield a new generation of virions (Figure 12.24). The protocol thus selects for viruses whose *tdk* gene contains a cloned insert of foreign DNA. Vaccinia viruses can also be engineered to carry genes from multiple viruses, forming polyvalent vaccines. Currently, several vaccinia vector vaccines have been developed and licensed for veterinary use, including one for rabies, while many other vaccinia vaccines are at the clinical trial stage.

Subunit vaccines, vaccines that contain only a specific protein or two from a pathogen, are also produced by recombinant means. For a pathogenic virus, the gene encoding its coat protein is often the best vaccine candidate because coat proteins are typically highly immunogenic. Subunit vaccines are popular because large amounts of immunogenic proteins are produced that can be administered at high dosage without the risk that exists with attenuated or killed-cell vaccines that may inadvertently contain viable pathogen cells or viruses. However, some subunit vaccines, such as that prepared against a surface protein of human hepatitis B, require that the immunogenic proteins be glycosylated by the host before they are immunologically active. To solve this problem, the subunit hepatitis B vaccine was produced in a eukaryotic host (yeast), which generated the glycosylated and immunologically active form of the vaccine.

Pathogens and Antibodies as Engineered Anticancer Therapeutics

While many cancers are treatable with radiation and chemotherapy, how to specifically target the drugs or radiation to tumor cells has been a long-standing problem, and biotechnology may

have a solution. *Listeria monocytogenes* is a pathogenic bacterium that causes listeriosis, a serious foodborne illness (see Section 32.13). *L. monocytogenes* grows within human cells, which allows wild-type strains to evade the immune system. By contrast, weakly pathogenic recombinant strains of *L. monocytogenes* can be cleared by the immune system of healthy cells but not by tumor cells. This observation hinted that weakened strains of *L. monocytogenes* might be turned into anticancer vehicles to deliver toxic drugs or radioisotopes specifically to tumor cells. This was accomplished by coupling the radioisotope $^{188}\text{rhenium}$ to a recombinant strain of *L. monocytogenes*. In experiments with mice, this therapeutic strain infected and multiplied in pancreatic tumor cells (Figure 12.25) without harming normal pancreatic cells.

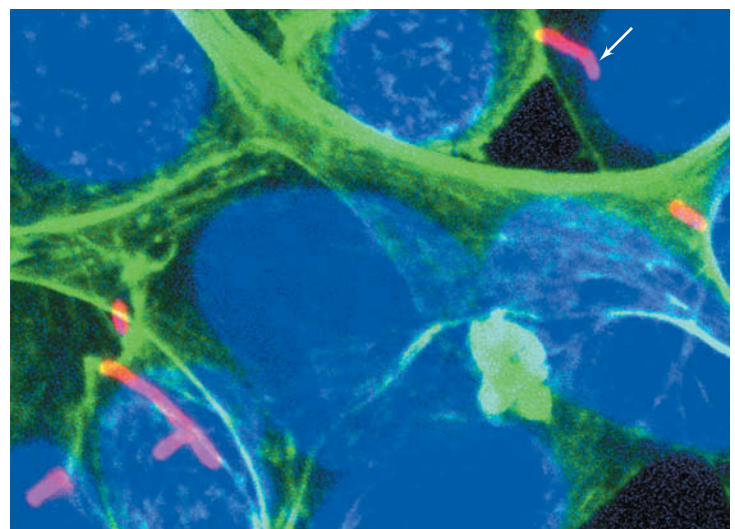


Figure 12.25 Therapeutic *Listeria*. *L. monocytogenes* cells (pink) linked to the radioisotope $^{188}\text{rhenium}$ enter and multiply inside mouse pancreatic tumor cells (blue, cell nuclei; green, cytoplasm) that have spread from the primary tumor. Radiation from the $^{188}\text{rhenium}$ slowly kills the tumor cells.

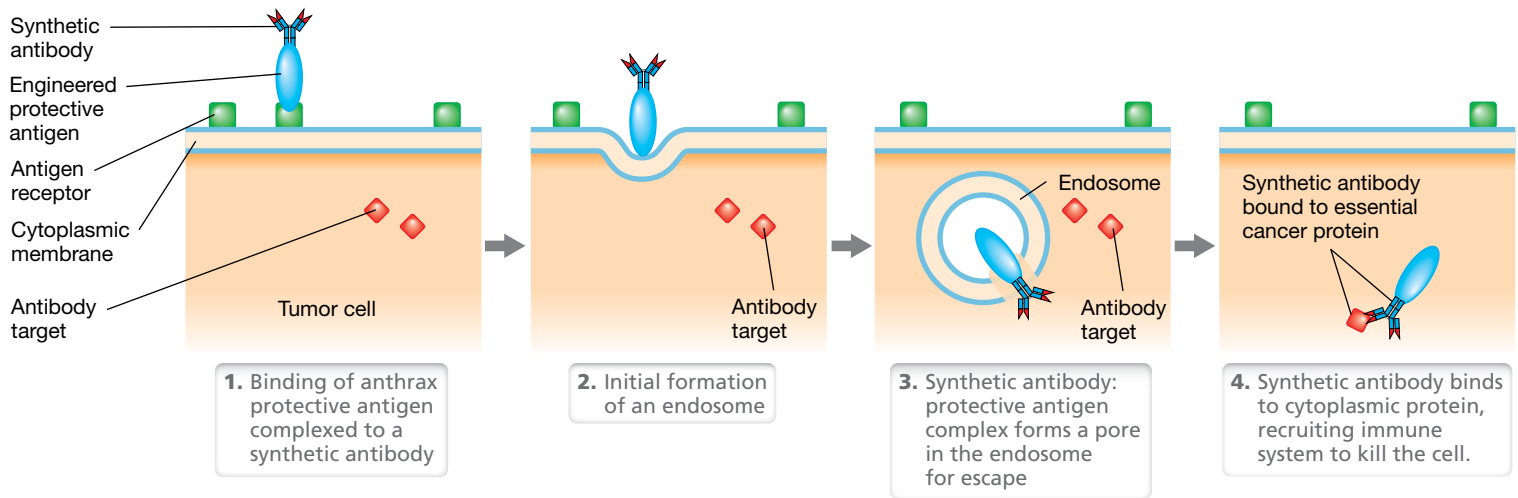


Figure 12.26 Engineered anthrax toxin. The protective antigen component of the anthrax toxin is engineered to carry a synthetic antibody. This engineered protective antigen specifically binds to a cell receptor on target cancer cells. After binding to the receptor, the engineered complex is taken up into the cell through an endosome. Following release into the cytoplasm, the synthetic antibody binds to an essential cellular protein, triggering cell death through the host's immune response. Anthrax toxin is discussed in more detail in Sections 29.9 and 31.8.

Another mechanism for treating cancer is by using antibodies, proteins produced by the immune system to attack foreign substances (see Section 27.3). It has been found that the binding of antibodies to specific targets inside cancer cells can trigger the host's immune system to kill the cancer cell. However, antibodies do not freely enter cells and thus a transport mechanism has been genetically engineered using a toxin produced by *Bacillus anthracis*, the bacterium that causes anthrax (see Sections 29.9 and 31.8) (Figure 12.26). Anthrax toxin contains three components: edema factor, lethal factor, and protective antigen; the latter is essential for carrying the toxic edema and lethal factors into the cell.

Using genetic engineering, scientists have modified the *B. anthracis* protective antigen to carry a *synthetic anticancer antibody* instead of the toxic edema and lethal factors. Injection of the modified and now harmless toxin results in the protective antigen recognizing and binding to a receptor on the outside of the cancer cell (Figure 12.26). The protective antigen: antibody complex is then taken up into the cancer cell through the formation of an endosome. Once the antibody is released from the endosome into the cytoplasm, it specifically binds to a protein essential to tumor viability. This binding then triggers the cell's immune system, which recognizes the antibody: cellular protein complex and kills the cell (Figure 12.26). Any antigen:antibody complex incorporated by normal cells is harmless because both the toxic portions of the toxin and the specific cancer proteins targeted by the antibody are missing in normal cells.

Stimulating the immune system to fight cancer may well turn out to be a mechanism in a whole new line of anticancer therapies. The novel system devised here—combining antitumor antibodies with a delivery system crafted from a highly toxic bacterium—shows both the power and the promise of genetic engineering to accelerate the war on cancer.

MINIQUIZ

- Explain why recombinant vaccines might be safer than some vaccines produced by traditional methods.
- What are the important differences among a recombinant live attenuated vaccine, a vector vaccine, and a subunit vaccine?
- What feature of some pathogenic bacteria make them attractive for engineered cancer treatments?

12.9 Mining Genomes and Engineering Pathways

Complex environments, such as fertile soil, contain vast numbers of uncultured microbes and their genes that are ripe for harvesting by genetic engineering. In addition, microbial metabolic pathways can be altered, embellished, or otherwise modified to change their characteristics and improve their efficiency. We explore how genetic engineering can both mine environmental genomes and alter metabolic pathways here.

Environmental Gene Mining

Just as the total gene content of an organism is its *genome*, the collective genomes of an environment are its *metagenome* (see Sections 9.8 and 19.8). *Gene mining* is the process of identifying and isolating potentially useful genes from the environment without the need to culture the organisms that contained them. In gene mining, DNA (or RNA) isolated directly from environmental samples is cloned into suitable vectors to construct a *metagenomic library* (Figure 12.27). If RNA is isolated, it must first be converted to cDNA by reverse transcriptase (Figure 12.2).

Screening of environmental metagenomic libraries has identified novel genes that encode enzymes that can degrade various pollutants and enzymes that make novel antibiotics. Retrieval of gene

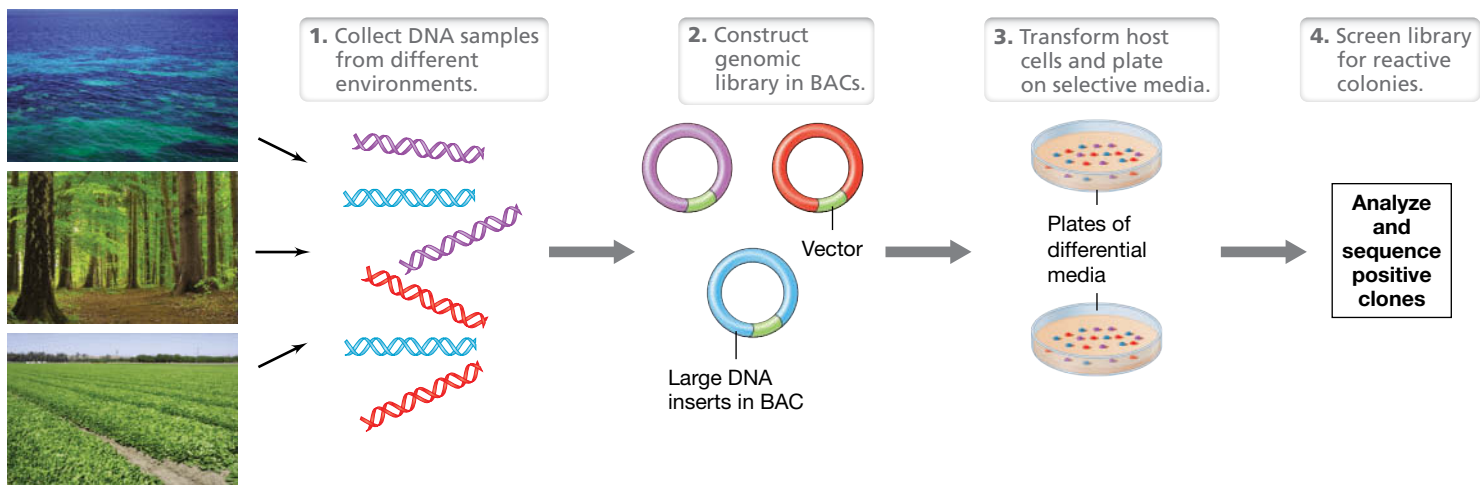


Figure 12.27 Metagenomic search for useful genes in the environment. DNA samples are obtained from different environments, such as seawater, forest soil, and agricultural soil. A metagenomic library is constructed using bacterial artificial chromosomes (BACs) and screened for genes of interest. Possibly useful clones are analyzed further.

clusters encoding entire metabolic pathways—such as for antibiotic synthesis—requires vectors such as **bacterial artificial chromosomes (BACs)**. BACs are similar to plasmids except that they can carry large inserts of DNA. BACs are especially useful for screening samples from rich environments, such as soil, where vast numbers of unknown genomes are present and correspondingly large numbers of genes are available to screen (Figure 12.27).

Several lipases, chitinases, esterases, and other degradative enzymes with novel substrate ranges and other properties have been isolated by this approach, and such enzymes have many industrial applications. Enzymes with improved resistance to industrial production conditions, such as high temperature, high or low pH, and oxidizing conditions, are especially valuable and desirable. Metagenomics can also target products with a particular combination of properties, such as a heat-stable lipase. Lipases hydrolyze fats, but their industrial production and use often requires that they remain active at high temperatures. To isolate a thermostable lipase, a metagenomic library was prepared from a hot spring sample and the DNA transformed into cells of *Escherichia coli*. Recombinant colonies expressing lipase activity were then selected and analyses indicated that certain of them remained active at 90°C. The gene encoding the heat-stable lipase was then introduced into an expression vector for commercial production of the enzyme.

By metagenomic mining of extreme environments, several useful heat- and acid-stable enzymes have been isolated for cleaning food-processing equipment in the food industry (Figure 12.28). To prevent foodborne infections, food-processing equipment must be rigorously cleaned, and cleaning protocols typically employ rigorous acid and base treatments, detergents, and sanitizers, all of which consume large amounts of chemicals and generate large volumes of wastewater that must be treated. By contrast, cleaning equipment with enzymes that function optimally near the boiling point of water in dilute acids (Figure 12.28) requires fewer chemicals and less water and more effectively removes microbial biofilms than do standard cleaning practices.

Pathway Engineering: Indigo Synthesis

Pathway engineering is the process of assembling a new or improved biochemical pathway using genes from one or more organisms. Engineered microbes are used to make alcohols, solvents, food additives, dyes, antibiotics, and many other products. They may also be used to degrade agricultural waste, pollutants, herbicides, and other toxic or undesirable materials. Here we discuss improving or modifying *existing pathways*, and later we explore the use of synthetic biology to create *entirely new pathways* (Section 12.11).

An interesting example of pathway engineering is the production of indigo by *E. coli* (Figure 12.29). Indigo is an important dye used for treating wool and cotton; blue jeans, for example, are made of cotton dyed with indigo. Although indigo can be synthesized chemically, the heavy demand for indigo by the textile industry has spawned new approaches for its synthesis, including a biotechnological approach using pathway engineering.



Figure 12.28 Application of CinderBio hyperstable enzymes for the cleaning of creamery equipment. These heat-stable enzymes clean industrial food-processing equipment as well as or better than traditional cleaning methods and do not generate large amounts of toxic wastewater.

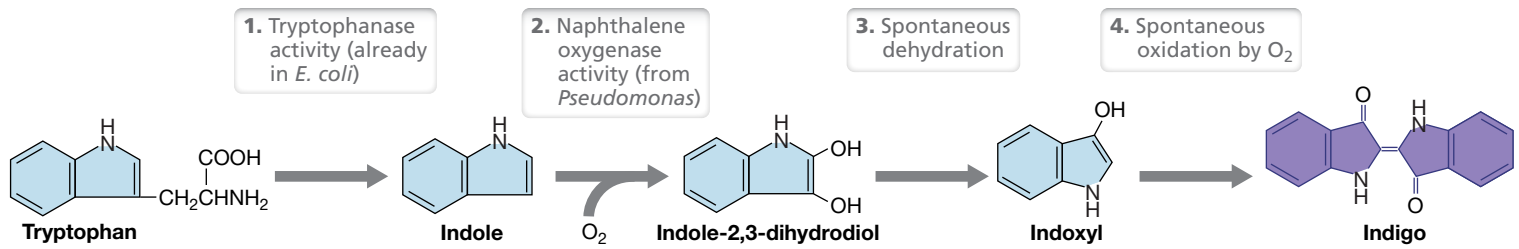


Figure 12.29 Engineered pathway for production of the dye indigo. *Escherichia coli* naturally expresses tryptophanase, which converts tryptophan into indole. Naphthalene oxygenase (originally from *Pseudomonas*) converts indole to dihydroxy-indole, which spontaneously dehydrates to indoxyl. Upon exposure to air, indoxyl dimerizes spontaneously to form indigo.

Because the structure of indigo is very similar to that of the aromatic hydrocarbon naphthalene, enzymes that oxygenate naphthalene also oxidize indole to its dihydroxy derivative; the latter oxidizes spontaneously in air to yield indigo, a bright blue pigment. Enzymes for oxygenating naphthalene are encoded by several plasmids found in *Pseudomonas* and other soil bacteria. When genes from such plasmids were cloned into *E. coli*, the cells turned blue because they had incorporated the genes encoding the enzyme naphthalene oxygenase.

The indigo pathway consists of four steps, two enzymatic and two spontaneous (Figure 12.29). *E. coli* naturally synthesizes the enzyme tryptophanase that carries out the first of these steps, the conversion of tryptophan to indole. In the engineered *E. coli*, a second step converts indole to the product that converts to indigo spontaneously (Figure 12.29). For indigo production, tryptophan must be supplied to the recombinant *E. coli* cells, and for commercial application, this was accomplished by affixing cells to a solid support in a bioreactor and then continuously trickling a tryptophan solution obtained from waste protein sources over the cells. If the tryptophan solution is recirculated over the cells several times, indigo levels steadily increase until the dye can be harvested.

Although bioproduction of indigo is clearly possible, it is currently difficult to compete with the chemical production of more than 20 kilotons per year. Indeed, some of the major challenges of pathway engineering are controlling the metabolic pathway and producing the desired compound at the yields necessary to be cost effective.

MINIQUIZ

- Explain why metagenomic cloning gives large numbers of novel genes.
- What types of environments are often sampled to prospect for industrial enzymes and why?
- How was *Escherichia coli* modified to produce indigo?

12.10 Engineering Biofuels

With the global supply of fossil fuels limited and environmental concerns growing about how climate change will affect our planet, renewable energy sources such as biologically produced fuels—*biofuels*—are in demand. The major biofuels in use today

are ethanol, biodiesel, hydrogen, and methane. Select microorganisms can produce biofuels; however, the yield from wild-type organisms is often hindered by toxic by-products and missing enzymes for critical steps. Thus, to enhance the production of biofuels, microorganisms have been genetically modified to optimize production. Here we discuss how genetic engineering has allowed for the use of alternative biofuel feedstocks, how enzyme replacement can yield new biofuels, and how phototrophic microorganisms can be harnessed as biofuel factories.

Bacterial Conversion of Switchgrass to Ethanol

Over 14 billion gallons of ethanol are produced per year in the United States from the fermentation of corn sugar by yeast (Figure 12.30a). However, because corn requires considerable cost and energy inputs to grow and harvest and is a major food source for both humans and domesticated animals, alternative nonedible and low-resource-input plant materials are more desirable biofuel feedstocks. Much attention has been focused on fast-growing grasses such as *switchgrass* (*Panicum virgatum*) (Figure 12.30b) as a source of cellulose for the production of ethanol. However, switchgrass cellulose is integrated with other plant polymers such as hemicellulose and lignin and requires high temperature, chemical, and enzymatic pretreatment to break the polymers down to fermentable sugars.

By leveraging both microbial diversity and genetic engineering, a bacterium has been discovered and genetically modified not only to break down switchgrass cellulose to fermentable sugars but also to ferment this sugar to ethanol. *Caldicellulosiruptor*, a gram-positive anaerobic and thermophilic bacterium, naturally produces a cellulase enzyme that can convert cellulose and hemicellulose to glucose. Unique proteins called *tāpirins* that extrude from the outer layer of the bacterium's peptidoglycan allow the cell to directly bind to raw switchgrass during the conversion process (Figure 12.30c). While *Caldicellulosiruptor bescii* grows optimally at 80°C and can ferment sugars, it naturally yields mostly acetate, lactate, and hydrogen as fermentation products. To directly convert switchgrass to ethanol, genetic engineers altered the terminal steps of the *C. bescii* glycolytic pathway (Figure 3.14) by replacing genes encoding lactate dehydrogenase and other acidic fermentation products with a bifunctional acetaldehyde/alcohol dehydrogenase from another thermophile, *Clostridium thermocellum*. This shifted the fermentation in *C. bescii* from mainly acidic products to 70% ethanol.



Figure 12.30 Biofuels. (a) A bioethanol plant in Nebraska (USA). In the plant, glucose from cornstarch is fermented by yeast to ethanol plus CO_2 . The large tank in the foreground is the ethanol storage tank, and the pipes in the background are for distilling the alcohol from the fermentation broth. (b) Switchgrass, a source of cellulose as a feedstock for ethanol production. (c) Fluorescence photomicrograph of acridine orange–stained cells of the thermophilic and cellulolytic bacterium *Caldicellulosiruptor kronotskyensis* growing on switchgrass (a single cell measures about $0.6 \mu\text{m} \times 3 \mu\text{m}$). The green color represents the switchgrass.

Use of thermophilic microorganisms for biofuel production has several advantages such as reduced risk of contamination with mesophiles and improved substrate solubility. It also makes the collection of any volatile products easier. For example, separating small amounts of desired products from large amounts of growth media (such as collecting ethanol during yeast fermentation of corn sugar) requires significant energy inputs to cool the bioreactors for growth of the organism and then later to heat and distill off the ethanol (Figure 12.30a). By contrast, since *C. bescii* grows optimally at 80°C —just above the boiling point of ethanol—commercial production of alcohol from cellulose by this genetically engineered bacterium requires little or no cooling and saves energy by the continuous emission of the desired product, ethanol.

Engineered Alkenes and Alkanes

Petroleum contains a mixture of hydrocarbons of varying chain length. Propane (C_3H_8), produced from natural gas processing and petroleum refining, is a widely used heating and cooking fuel and a key fuel for agricultural applications. Using genetic engineering, scientists have modified strains of *Escherichia coli* to convert glucose into propane and some other petroleum hydrocarbons.

Hydrocarbon production in *E. coli* begins with the synthesis of a fatty aldehyde. This was done by heterologously expressing in *E. coli* the *Photorhabdus luminescens luxCED* genes, which encode enzymes that reduce fatty acids to their corresponding aldehydes (Figure 12.31). The activity of these fatty acid reductase, synthetase, and transferase enzymes yields a fatty aldehyde that can be converted to hydrocarbons by the enzyme aldehyde decarbonylase. However, *E. coli* lacks this enzyme as well. To overcome this limitation, genetic engineers cloned the aldehyde decarbonylase gene from the cyanobacterium *Nostoc punctiforme* into *E. coli*, allowing the engineered *E. coli* to convert linear fatty acids added to its growth medium into linear hydrocarbons (Figure 12.31).

Because branched-chain hydrocarbons yield higher octane numbers (which is good for gasoline engine performance), scientists expanded this engineered pathway by heterologously expressing enzymes that participate in the first step of the fatty acid elongation cycle (↻ Figure 3.30) from *Bacillus subtilis*. This allowed

the engineered *E. coli* to use more branched-chain fatty acids as starting substrates and by doing so generate higher-octane fuels.

Microalgae and Biodiesel

Microalgae are unicellular phototrophic eukaryotes that produce an abundance of bioactive compounds including lipids, fatty

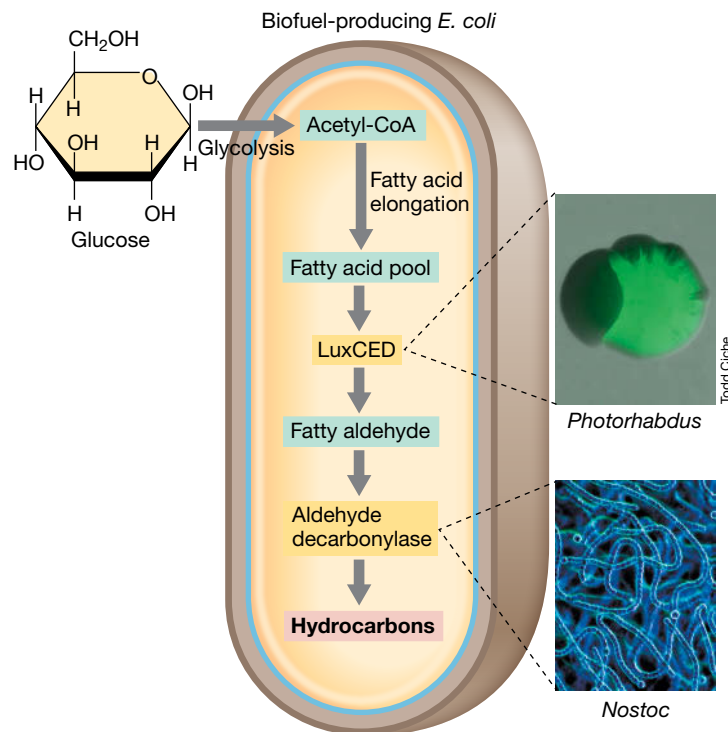


Figure 12.31 Hydrocarbon-producing *Escherichia coli*. *E. coli* naturally produces fatty acids from acetyl-CoA. By engineering a strain to express the LuxC (fatty acid reductase), LuxE (fatty acid synthetase), and LuxD (fatty acid transferase) proteins from the fluorescent bacterium *Photorhabdus luminescens* (inset: fluorescent colony), a fatty aldehyde intermediate is produced. This fatty aldehyde can then be converted to a hydrocarbon if the same strain has also been engineered to express the aldehyde decarbonylase enzyme from the cyanobacterium *Nostoc punctiforme* (inset photo).

acids, and carotenoids. These products are made using only sunlight, CO₂, a few minerals, and water. Microalgae of interest to biotechnology include the green algal genera *Chlorella* and *Chlamydomonas*. These organisms produce significant amounts of storage lipids known as *triacylglycerides* (TAG), substances that can be chemically treated to yield *biodiesel*, a fuel for use in the diesel engines present in many trucks and heavy transport vehicles.

Improving TAG synthesis in the microalgal biosynthetic pathway required some genetic engineering breakthroughs, and **Figure 12.32a** illustrates the successful design of vectors that allow proteins to be targeted to the nucleus or chloroplast of cells of *Chlamydomonas*; other vectors have been developed that allow the mitochondrion or the endoplasmic reticulum to be targeted. Organelle targeting is facilitated through the fusion of *signal sequences* (↔ Section 4.12) to the protein of interest. A vector for the simultaneous expression of

two foreign genes in separate cellular locations has also been generated (Figure 12.32b). These targeting vectors are critical for manipulating compartmentalized TAG biosynthetic activities such as control of gene expression by transcription factors (encoded by the nuclear genome), ATP-producing enzymes (encoded by the mitochondrial genome), and enzymes for initial fatty acid synthesis reactions (encoded by the chloroplast genome). Translation of proteins by ribosomes on the endoplasmic reticulum is also important for protein secretion.

Using microalgal triacylglycerides as a feedstock for biodiesel increases the possibilities for biofuel production, and the fact that the process is driven by the energy of sunlight makes it an environmentally attractive process. However, the major obstacle facing all biofuel production schemes—especially those that require sunlight—is the expense of the equipment and engineering necessary to scale production up to levels necessary to compete with the petroleum industry. Currently, about 80 million barrels of oil a day arrive on the highly volatile world energy market, and until oil supplies diminish to the point where a significant price hike is encountered, any biofuel will have a difficult time competing.

MINIQUIZ

- How has *Caldicellulosiruptor* been modified to produce ethanol directly from switchgrass?
- What is the advantage of using thermophiles to produce biofuels?
- What has been the limiting factor in engineering microalgae to produce greater amounts of lipids?

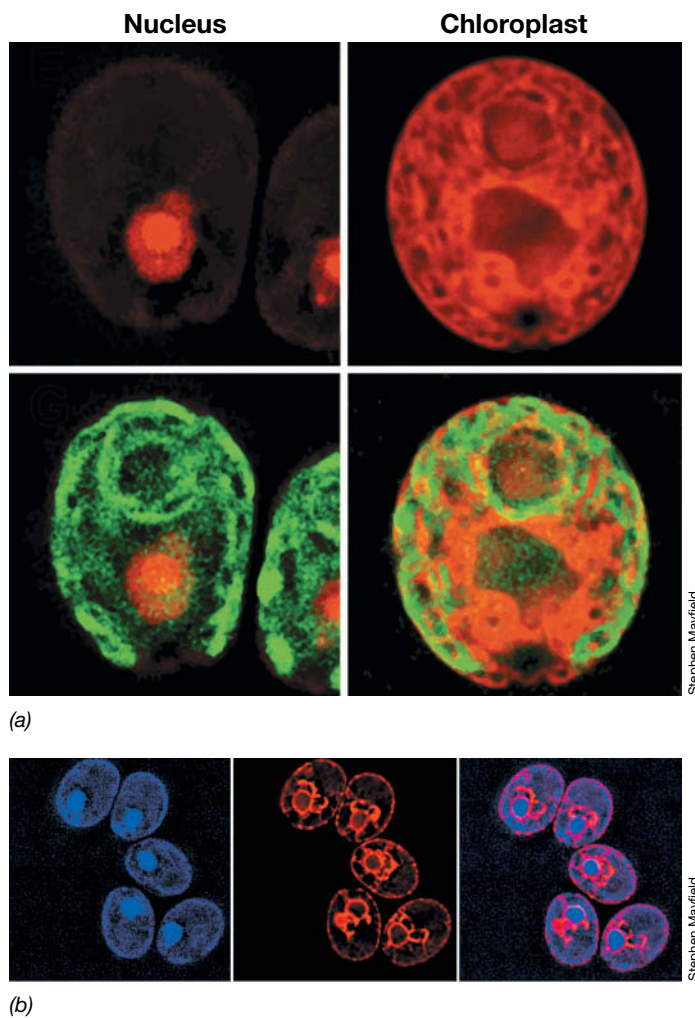


Figure 12.32 Genetic tools for engineering microalgae. (a) Fluorescent micrographs showing the expression of reporter genes in targeted regions of cells of *Chlamydomonas*. Top panel: A reporter gene encoding a red fluorescent protein that targets the nucleus (left) and chloroplasts (right). Bottom panel: both images overlaid with a green fluorescent stain for the chloroplast. (b) Dual expression of two reporter genes to different cellular locations using a single vector. A gene encoding a blue fluorescent protein is localized to the nucleus, while a gene encoding a red fluorescent protein is targeted to the endoplasmic reticulum in cells of *Chlamydomonas*. Adapted from Rasala, B.A., S-S. Chao, M. Pier, D.J. Barrera, and S.P. Mayfield. 2014 *PLoS ONE* 9(4): e94028.

III • Synthetic Biology and Genome Editing

The term *synthetic biology* refers to the use of genetic engineering to create novel biological systems out of available biological parts, often taken from several different organisms. These biological parts (promoters, enhancers, operators, riboswitches, regulatory proteins, enzyme domains, signal receivers, etc.) have been termed *biobricks*. Synthetic biology links these biobricks together in various combinations to form modules capable of generating complex behaviors. While we will discuss some amazing examples of synthetic biology (including the formation of synthetic cells) in Section 12.11, college students are also trying their hand at synthetic biology. An undergraduate competition called the *International Genetically Engineered Machine* (iGEM; <http://igem.org/>) occurs annually worldwide. Teams in this competition have used synthetic biology to engineer products ranging from biodegradable Styrofoam-like material to a strain of *Escherichia coli* that can protect honeybees from parasites.

Another powerful new and rapidly developing technology allows the precise editing of genomes in living cells, a technique that has revolutionized biotechnology. In Section 12.12 we explore how the microbial immune system has been leveraged to edit any genome and the remarkable applications of this genome editing technology.

12.11 From Synthetic Metabolic Pathways to Synthetic Cells

A major focus of synthetic biology thus far has been the construction or modification of metabolic pathways. By using various enzyme and regulator biobricks, synthetic biologists can construct artificial pathways to convert cheap and abundant substrates into high-value products. Such products are often expensive because purifying them from their original source, typically plants, is costly. While genetically modified organisms (GMOs) are used for their production, no foreign DNA is present in the products.

Engineering a Major Food Product

Vanillin, one of the most popular flavoring agents in the world, is a secondary metabolite extracted from the seedpods (vanilla beans) of orchids of the genus *Vanilla*. Natural vanillin is expensive because of the slow growth of orchids, their relatively low output, and the high production costs of cultivating and harvesting the beans. By analyzing the natural metabolic pathway for the production of vanillin, genetic engineers have synthesized strains of *Escherichia coli* and yeast that can synthesize the flavoring agent from glucose. Vanillin synthesis in *E. coli* requires five heterologously expressed enzymes, and once the necessary biobricks were incorporated into the bacterium, it produced vanillin identical in structure and taste to naturally produced vanillin. (However, some argue that naturally produced vanilla contains additional compounds extracted from the vanilla beans that contribute to its unique taste.)

“Synbio vanillin,” as the *E. coli* product has been called, is commercially available as an inexpensive source of vanillin and has been used primarily for flavoring ice cream and baked goods. As the second most expensive spice in the world (behind saffron), natural vanilla is an especially costly ingredient for bulk use such as in ice cream. Synbio vanillin is a good example of how synthetic biology can be used to convert inexpensive feedstocks (corn sugar) into high-value products.

Synthetic Pharmaceuticals: Artemisinin and Malaria

Pharmaceuticals are often derived from natural products; aspirin, for example, was originally obtained from willow bark. While aspirin is now chemically synthesized, not all pharmaceuticals can be

economically synthesized. One example is the antimalarial drug *artemisinin*. Malaria, which is caused by protozoans of the genus *Plasmodium*, is transmitted by mosquitoes and infects nearly 500 million people each year, primarily in tropical and subtropical countries. While various traditional antimalarials are in use, the parasite has evolved resistance to many of these and thus new antimalarial drugs are constantly needed. Artemisinin is an alternative antimalarial that is produced in limited amounts by the cultivated sweet wormwood plant (*Artemisia annua*). To ensure availability of the drug, the *Semi-Synthetic Artemisinin Project* was initiated with the goal of engineering a microorganism for the synthesis of artemisinic acid, used for production of artemisinin (Figure 12.33).

A. annua naturally produces artemisinic acid by converting acetyl-CoA to farnesyl diphosphate (FPP; a 15-carbon intermediate) using the mevalonate biosynthetic pathway. FPP is then oxidized to artemisinic acid and dihydroartemisinic acid through an intermediate called *amorphaadiene*. Initially, the plan was to have *E. coli* produce artemisinic acid through fermentation. Synthetic biologists made numerous attempts to engineer *E. coli* with the correct biobricks to produce artemisinic acid through permutations of the natural pathway, inhibiting natural competing enzymes, mutating genes for codon optimization (Section 12.3), and modifying fermentation conditions, but no strategy emerged that could yield the amorphaadiene intermediate at sufficient levels. However, by transferring the necessary metabolic pathway biobricks to a modified strain of the baker’s yeast *Saccharomyces cerevisiae* along with metabolic adjustments to divert carbon flux toward the final product, synthetic biologists designed a yeast strain that produces large amounts of artemisinic acid that can then be chemically converted to artemisinin (Figure 12.33).

Synthetic biology has also been successful at synthesizing powerful painkilling drugs. For example, genetic engineers rewired yeast to produce the chemical *thebaine*, a precursor to morphine and hydrocodone, using glucose as feedstock. This was accomplished by the heterologous expression in yeast of 21 different genes that originated from sources as diverse as plants, a rat, and the gram-negative bacterium *Pseudomonas*. The synthesized thebaine can then be chemically converted to a suite of pain-relieving drugs and marketed by pharmaceutical companies.

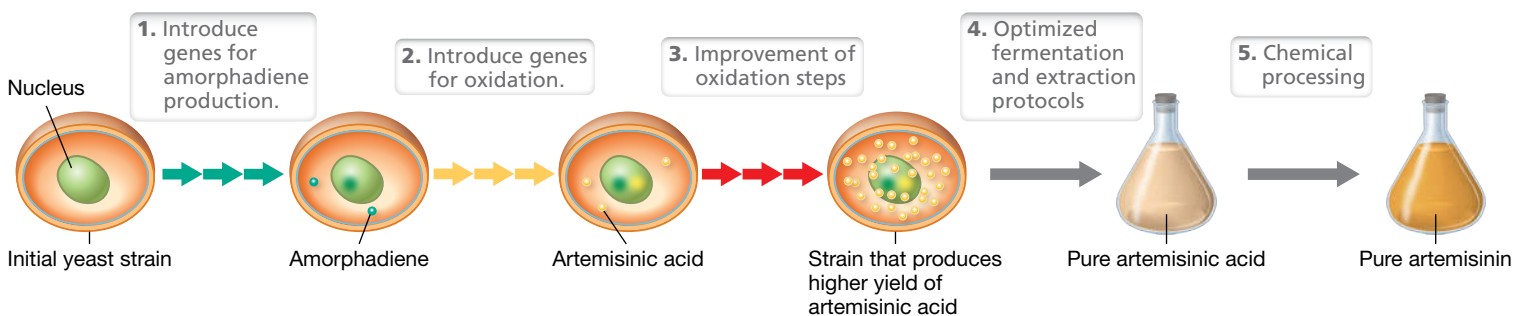
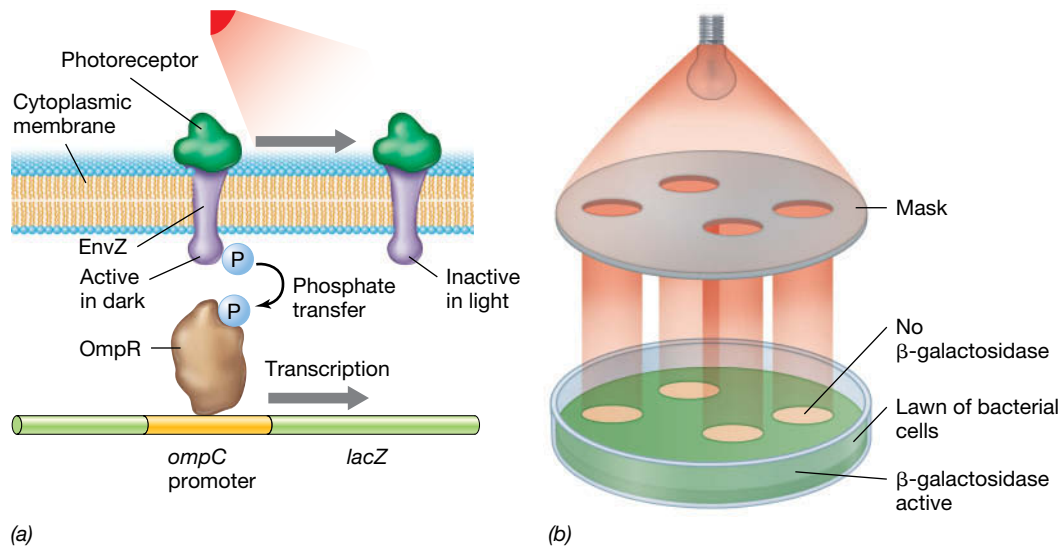


Figure 12.33 Artemisinin synthesis through synthetic biology. A summary of the engineering steps used to modify a *Saccharomyces cerevisiae* yeast strain to produce artemisinic acid. Colored arrows represent expression of genes from the plant *Artemisia annua* and other genetic modifications.



Aron Chevalier and Matt Levy

Figure 12.34 Bacterial photography. (a) Light-detecting *Escherichia coli* cells were genetically engineered using components from cyanobacteria and *E. coli* itself. Red light inhibits phosphate (P) transfer to the DNA-binding protein OmpR; phosphorylated OmpR is required to activate *lacZ* transcription (*lacZ* encodes β -galactosidase). (b) Setup for making a bacterial photograph. The opaque portions of the mask correspond to zones where β -galactosidase is active and thus to the dark regions of the final image. (c) A bacterial photograph of a portrait of Charles Darwin.

Photographic *Escherichia coli*

An early excursion into the world of synthetic biology is the use of genetically modified *E. coli* to produce photographs. The engineered bacteria are grown as a lawn on agar plates, and when an image is projected onto the lawn, unilluminated bacteria make a dark pigment while illuminated bacteria do not. The result is a primitive photograph of the projected image (Figure 12.34).

Construction of the photographic *E. coli* required synthetic biologists to create three key biobricks: (1) a light detector and signaling module; (2) a pathway to convert heme (already present in *E. coli*) into the photoreceptor pigment phycocyanobilin (an accessory light-harvesting pigment of cyanobacteria, [↔](#) Section 14.2); and (3) an enzyme encoded by a gene whose transcription can be switched on and off to make the dark pigment (Figure 12.34a). The photoreceptor is a fusion protein in which the sensing half is the light-detecting part of the phytochrome protein from the cyanobacterium *Synechocystis*. This required phycocyanobilin, which is not naturally made by *E. coli*, hence the need to install the biobrick that contained the pathway to make phycocyanobilin.

The other half of the fusion protein is the signal transmission domain of the EnvZ sensor protein from *E. coli*. EnvZ is part of a two-component regulatory system, its partner being OmpR ([↔](#) Section 6.6). Normally, EnvZ activates the DNA-binding protein OmpR, and the latter in turn activates target genes by binding to the promoter. The hybrid protein was designed to activate OmpR in the dark but not in the light. This is because phosphorylation of OmpR is required for activation, and red light converts the sensor to a state in which phosphorylation is inhibited. Consequently, the target gene is *off* in the light and *on* in the dark. When a mask is placed over the Petri plate containing a lawn of the engineered *E. coli* cells (Figure 12.34b), cells in the dark make a pigment that cells in the light do not, and in this way a “photograph” of the masked image develops (Figure 12.34c).

The pigment made by the *E. coli* cells results from the activity of the lactose-degrading enzyme β -galactosidase, naturally present in *E. coli*. The target gene, *lacZ*, encodes this enzyme. In the dark, *lacZ* is expressed and β -galactosidase is made. The enzyme cleaves the lactose analog X-gal (Section 12.2) present in the growth medium to release galactose and a black dye. In the light, the *lacZ* gene is not expressed, no β -galactosidase is made, and so no dye is released. Contrast in the photograph is controlled by how much light cells see, which is governed by the nature of the mask that is used (Figure 12.34c).

Although bacterial photographs can hardly compare with digital photographs, the knowledge gained from assembling the biobricks required for bacterial photographs—work that is now many years old—helped form a foundation for the deployment of synthetic approaches in more complex biological systems, including entire cells, to which we turn now.

Synthetic Cells

The synthesis of an entire cell from scratch can be considered the pinnacle of synthetic biology (see page 368 for more on this). This has not happened as of 2017, but a related feat was announced in 2010: A group of synthetic biologists from California (USA) had produced a “synthetic” bacterium. However, the organism produced was not the result of assembling various biobricks to form a living organism—true *de novo* cell synthesis—but instead was the product of the artificial construction of a bacterial genome from a known genome sequence and the insertion of this synthetic genome into a different bacterial species to yield viable cells (Figure 12.35).

This feat was accomplished by artificially synthesizing a 1.08-million-base-pair (Mbp) genome based on the genome sequence of the bacterium *Mycoplasma mycoides*. This circular chromosome was pieced together from linear fragments of DNA containing homologous ends and the homologous recombination

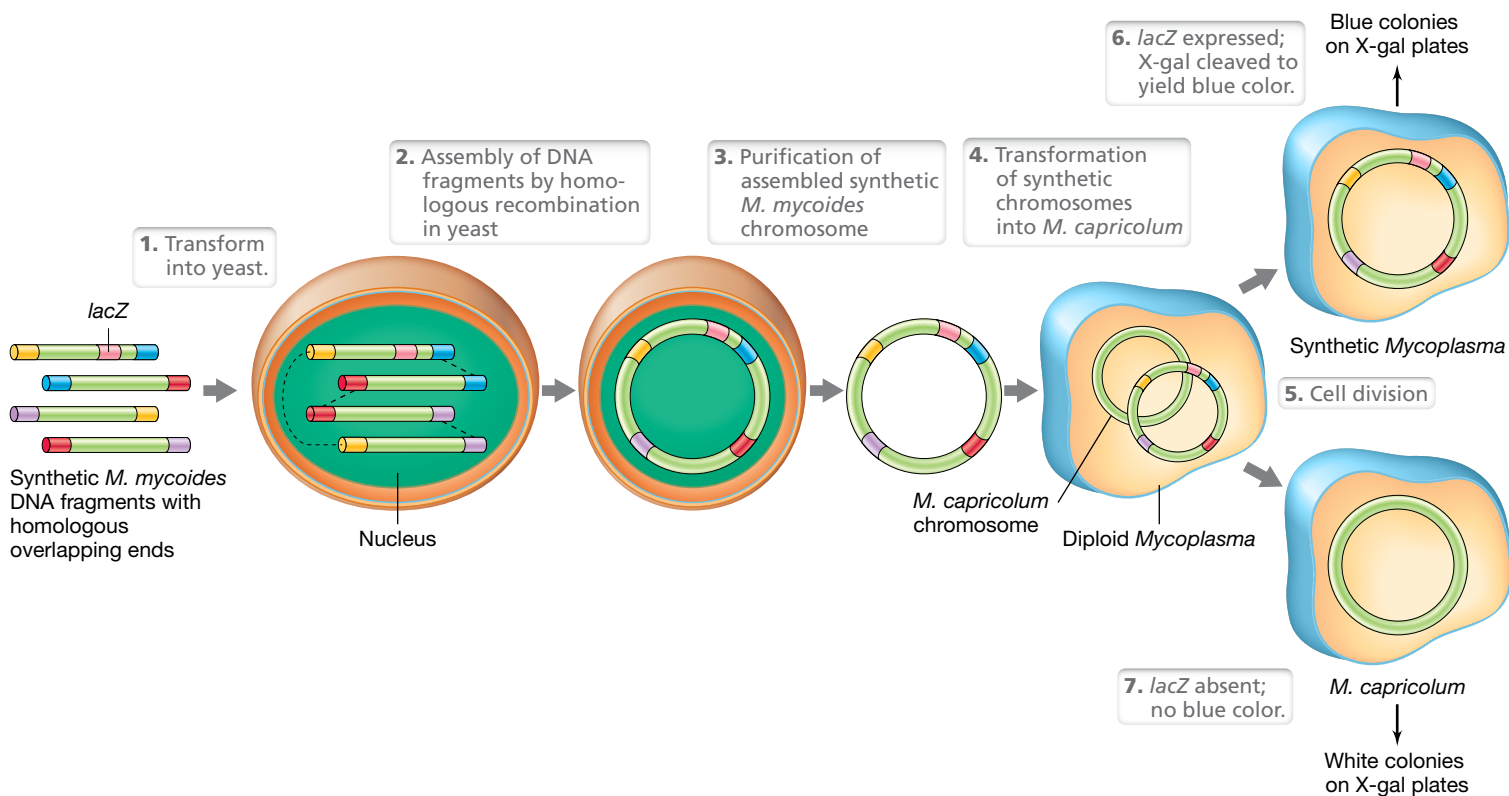


Figure 12.35 Formation of a synthetic *Mycoplasma* cell. DNA fragments corresponding to the desired *M. mycoides* chromosome are synthesized with overlapping ends (represented by the same colors). These DNA fragments are linked together using the homologous recombination machinery of *Saccharomyces cerevisiae*,

with sites of homologous recombination indicated by lines. Once assembled, the synthetic chromosome is transformed into a cell of *M. capricolum*. This results in a cell with two separate chromosomes. After cell division, one cell possesses the synthetic chromosome (considered the synthetic *M. mycoides*) and the other remains a

wild-type *M. capricolum*. The cell containing the synthetic chromosome was identified because it contained the reporter gene *lacZ* and thus produced blue colonies (see Figure 12.8). For more coverage of synthetic cells including research to define the minimalist genome, see page 368.

complex of yeast cells. The fully assembled synthetic chromosome was then purified and transformed into a cell of *Mycoplasma capricolum* (Figure 12.35). The synthetic chromosome contained a reporter gene, *lacZ*, absent from the *M. capricolum* genome, which caused colonies containing the synthetic chromosome to turn blue (see Figure 12.8); this was needed to distinguish *M. capricolum* cells containing the *M. mycoides* genome from those containing the *M. capricolum* genome after cell division (Figure 12.35). When cells in the blue colonies were examined, they showed all of the properties of the original *M. mycoides* cell.

Although this synthetic *M. mycoides* was not constructed solely from biobricks (the *M. capricolum* host cell contained ribosomes, various enzymes, and other important cytoplasmic components), the experiment did prove that an entire genome could be transplanted from one species to another. And just as importantly, the experiment offered a glimpse of the possibilities that lie on the horizon for synthetic biology. The possibilities are indeed endless, and the benefits of this science for humans and our planet will likely be significant.

MINIQUIZ

- What are biobricks?
- What organism has been genetically modified to produce precursors to the drugs artemisinin and morphine?
- How was *Escherichia coli* modified to produce a photograph?

12.12 Genome Editing and CRISPRs

Earlier in this book we discussed clustered regularly interspaced short palindromic repeat (CRISPR) systems and their role in protecting *Bacteria* and *Archaea* from foreign DNA and maintaining genome integrity (see Sections 10.13 and 11.12). Microbiologists studying a CRISPR system called *CRISPR/Cas9* (*Cas9* refers to *CRISPR associated protein 9*) in the bacterium *Streptococcus pyogenes* discovered that the system could also recognize and cleave a specific DNA sequence within other cells and that foreign DNA could be inserted into the cut site.

Optimization of the CRISPR/Cas9 system has revolutionized biotechnology by providing the most powerful and precise tool yet for altering eukaryotic genomes in living cells. Indeed, *genome editing*, as it has come to be known, has been used successfully to edit the genomes of plants, animal embryos, and human cell lines. Here we explore how this system works and its benefits to synthetic biology.

Sequence Targeting by the Cas9 Protein

As illustrated in Figure 10.28, CRISPR systems possess Cas proteins that function as endonucleases when guided to a piece of nucleic acid by the complementary binding of CRISPR RNAs (crRNAs). By designing a synthetic RNA molecule that both recruits the *Streptococcus* Cas9 protein and binds to the desired target DNA sequence,

genetic engineers have harnessed the power of the CRISPR–Cas system to cut specific DNA sequences in the genome of virtually any cell. At the cut site, the DNA can either be ligated (yielding a gene deletion) or used for inserting new DNA (Figure 12.36a, b).

The synthetic RNA molecule used for gene editing is called a *synthetic guide RNA* (sgRNA), and the Cas9 protein from *S. pyogenes* binding both the sgRNA and target DNA can be visualized in Figure 12.36c. For complete Cas9 endonuclease activity, a short *protospacer adjacent motif* (PAM; see Figure 10.28) must also occur on the target DNA. Without this PAM sequence, the Cas9 protein will bind to the region where the sgRNA binds, but will not cut. When the Cas9 protein does cut DNA, its two endonuclease domains (Figure 12.36c) cooperate to cut both DNA strands, and this generates double-stranded breaks (Figure 12.36a, b). Depending on the target cell, various methods of delivering the CRISPR system can be used. Genes corresponding to the designed sgRNA and Cas9 protein are often cloned into a plasmid under regulatory control

of a strong promoter. Alternatively, the sgRNA and mRNA corresponding to the Cas9 protein can be generated *in vitro*. In either case, the materials are injected directly into target cells to trigger the gene editing process.

The presence of a PAM sequence close to the desired cleavage site is often the only limitation to cutting a specific DNA sequence. However, the PAM sequence is just three nucleotides in length and thus occurs with frequency in most genomes. CRISPR systems from other bacteria also recognize different PAMs, so alternative Cas proteins may be employed if needed. To delete a region of DNA, two target cleavage sites flanking the DNA sequence to be deleted must be identified and corresponding sgRNAs designed (Figure 12.36b). By contrast, only one cleavage site and corresponding sgRNA are needed to insert DNA (Figure 12.36a). To edit a region of a chromosome, sgRNAs are designed to bind to target DNA. This binding stimulates the Cas9 protein to cleave the target site if a PAM sequence is nearby (Figure 12.36a, b).

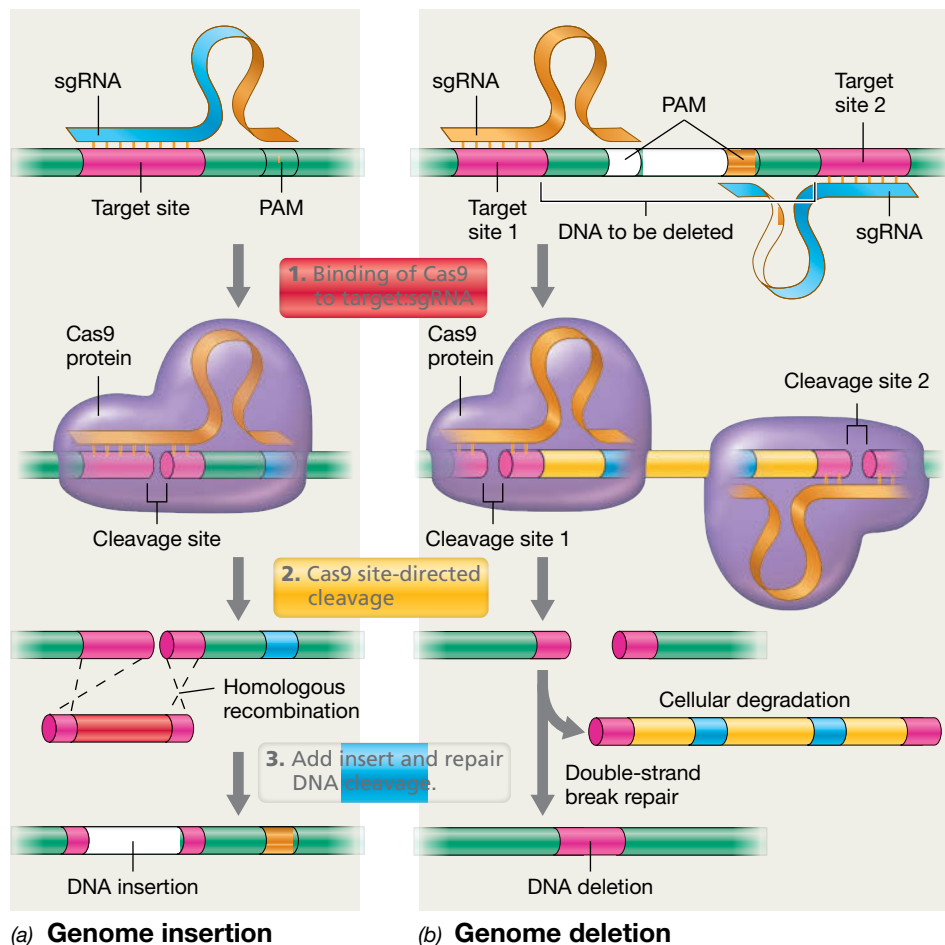
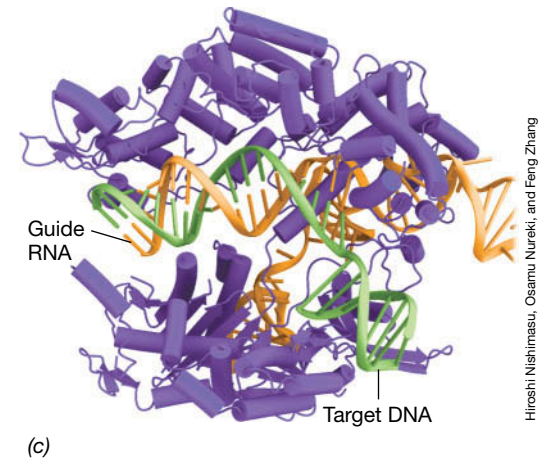


Figure 12.36 CRISPR/Cas9 genome editing. sgRNA represents the synthetic guide RNA, while PAM represents a protospacer adjacent motif. Note that each genome target site must possess a PAM sequence for DNA cleavage to occur. (a) Insertion of foreign DNA into a targeted site of the genome. An sgRNA is synthesized to bind to a single target site on the genome through complementarity. This binding of the sgRNA to the DNA

stimulates the Cas9 protein to cleave the genome at the target site. Foreign DNA with ends homologous to the cleavage site can be incorporated into the cut site through homologous recombination. This results in a genomic insertion. (b) Deletion of a genomic region. Two separate target sites flanking the DNA to be deleted are selected. After the design, addition, and binding of sgRNAs corresponding to these regions, Cas9



protein-dependent DNA cleavage occurs. This results in a double-strand break in the target chromosome and a free piece of DNA. The double-strand break is then ligated by the cell's DNA double-strand break repair pathway, while the free piece of genomic DNA is degraded. This results in a genomic deletion. (c) Crystal structure of the *Streptococcus pyogenes* Cas9 protein. The target DNA is shown in green and the sgRNA in orange.

While a Cas9 protein and sgRNA can be used to cut DNA at specific sites (Figure 12.36), how is new DNA *inserted* at the cleavage site and how does the DNA get ligated back together? These tasks are accomplished by harnessing the cell's own DNA repair machinery. If a piece of DNA containing sequences with homology to the cut site is added to the system, homologous recombination will be used to incorporate the DNA (↔ Section 11.5), yielding a genomic insertion (Figure 12.36a). If the goal is only to delete the chromosomal region between two cut sites, the nonhomologous double-strand DNA break repair pathway will be employed to ligate the DNA following the deletion event (Figure 12.36b).

CRISPR Editing in Practice

Applications of CRISPR genome editing have seen a meteoric rise since its discovery in 2013. By designing a Cas9-encoding gene to possess codons optimized for the organism of interest (Section 12.3) and engineering sgRNA to target the gene of interest, almost any DNA can be edited. While the genomes of rice, sorghum, and wheat have been edited using CRISPR, the CRISPR/Cas9 system has also been employed in tomato plants (a dicot) by targeting a gene region in which a mutation leads to leaves that are needle-like or wiry (Figure 12.37). In this study, DNA encoding the Cas9 protein and sgRNA was introduced into the plant cells using *Agrobacterium* and the Ti plasmid system (Figure 12.19). Because the resulting mutations were stable and heritable, CRISPR genome editing is poised for testing the functions of unknown genes in dicots and may possibly be used in the near future to improve the nutritional and other properties of fruits and vegetables.

While a Cas9 system has been used to excise the genome of the retrovirus HIV (the causative agent of AIDS, ↔ Section 10.11) from the genome of infected human cells *in vitro*, Cas proteins from other bacteria can be used to target HIV or other specific viral RNAs. For example, the Csy4 CRISPR-associated protein from *Pseudomonas* is an endoribonuclease that processes the long CRISPR transcript (↔ Figure 11.33). The Csy4 protein has been modified to recognize and destroy free HIV RNA in infected cells (Figure 12.38). Similarly, the CRISPR/Cas9 system from the bacterium *Francisella novicida* has been reprogrammed to target the

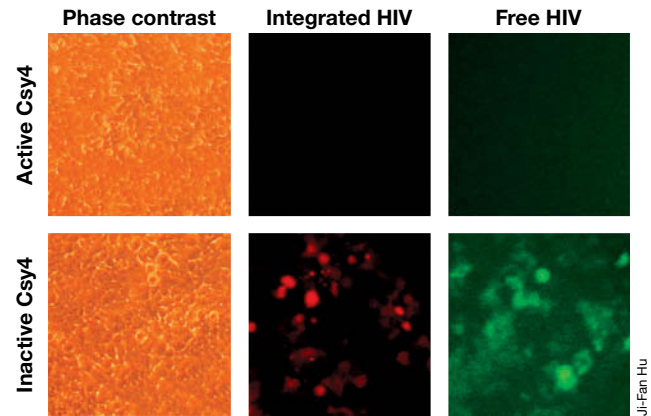


Figure 12.38 CRISPR-mediated inhibition of HIV infection. The results of using the *Pseudomonas* Csy4 endoribonuclease CRISPR protein to target RNA-based HIV in infected human embryonic kidney cells. Red indicates HIV provirus integration into infected cells, while free HIV virions are indicated in green. Top panel: expression of an HIV–Csy4 targeting vector. Bottom panel: expression of a mutated vector containing a nonfunctional Csy4 protein. Adapted from Guo, R., H. Wang, J. Cui, G. Wang, W. Li, and J-F Hu. 2015. *PLoS ONE* 10(10): e0141335.

single-stranded RNA genome of hepatitis C virus (a liver pathogen and a cause of liver cancer) in a human cell line.

Not only has CRISPR genome editing been used to delete, interrupt, and insert DNA sequences into a single location, it can also be used to target multiple genetic loci. An impressive example of this is the removal of 62 copies of the porcine endogenous retrovirus from swine cells. The presence of this retrovirus is one of the factors preventing the use of swine organs for transplants in humans (swine are anatomically very similar to humans). While there are other swine proteins that provoke an immune response in humans, genetic engineers envision editing away all of these factors to produce immune-friendly pig embryos for human organ production in the next few years!

Currently, CRISPR genome editing appears to have very few limitations. In fact, fertility clinic human embryos that were nonviable and could not result in a live birth have even been modified. While this landmark accomplishment has raised serious ethical questions regarding the use of CRISPR editing in humans, the technique may be the key to eradicating a host of devastating genetic diseases before a baby carrying the genes for one or more of them is born.

MINIQUIZ

- What is it about the Cas9 protein that makes it an efficient DNA editing tool?
- What is the role of the sgRNA in genome editing?
- How is recombinant DNA inserted into a genome using CRISPR editing?

12.13 Biocontainment of Genetically Modified Organisms

Throughout this chapter we have focused on genetically engineering microbes as factories for the synthesis of high-value products. While these applications of genetic engineering are clearly



Figure 12.37 CRISPR editing of tomato genome. Interruption of the tomato *SlAGO7* gene encoding an argonaute homolog results in plants that have wiry or spindly leaves (left) compared to a normal tomato plant (right).

beneficial, environmental concerns remain that genetically modified organisms (GMOs) may spread their modified genes to wild-type populations with adverse consequences. How can synthetic biology help solve this problem?

Early Containment Schemes

Through the years, various schemes have been proposed for containing GMOs. For example, both the use of auxotrophic strains and the induction of genes encoding self-toxins have been attempted with GMOs, but for both strategies, successful and reliable implementation has been a problem. Recall that an auxotroph is a mutant derivative of a microbial species that has a nutritional requirement; without the nutrient, the auxotroph cannot grow (↔ Section 11.1), and this is the theory behind using auxotrophs to contain GMOs. However, in nature, auxotrophic strains can often survive by cross-feeding off the metabolites of other organisms, and the possibility always remains that an auxotrophic GMO could revert back to the wild type by back mutation and lose its nutritional dependencies.

In Section 7.11 we saw how certain bacteria produce a growth-inhibiting toxin that slows cell growth to help ensure survival of the population under stressful conditions. This has also been explored as a mechanism for biocontainment. That is, if a GMO were to escape confinement within a bioreactor (where all growth conditions are ideal for production of the desired product), the toxin would kick in and trigger the dormant state, from which the escaped GMO could not recover. However, in nature, where a cell has to compete not only with cells of its own kind but also with other microbial species, survival strongly selects for toxin gene mutations; if such a mutation occurred quickly, the toxin system could be disarmed and the GMO might survive.

Neither the auxotroph nor the toxin approaches adequately solve the problem of GMO containment. Moreover, neither of these mechanisms address the possibility of engineered DNA being released through industrial waste streams and finding its way into other microorganisms by horizontal gene transfer (Chapter 11). Thus, to more thoroughly and safely address the containment problem, genetic engineers have tapped into synthetic biology itself to devise novel methods of controlling GMOs in the environment, and we consider one now.

Biocontainment by Recoding the GMO Genome

A novel approach to prevent genetically modified bacteria from surviving outside of their bioreactor is to recode the genome of the GMO so the bacterium *can only grow if supplied with a synthetic amino acid* (Figure 12.39). This involves rewiring the organism's genetic code and translational machinery (Chapter 4) to synthesize proteins containing the synthetic amino acid. To accomplish this significant feat, synthetic biologists first replaced all the TAG (UAG on the RNA) stop codons associated with open reading frames on the *Escherichia coli* chromosome with the TAA stop codon. They also deleted the gene for release factor 1 that terminates translation when the ribosome encounters a UAG on the mRNA. Because this manipulation placed an alternative stop codon at the end of genes that had TAG codons, proteins of the correct length are still produced during translation and the recoded cell grows normally. This genetic manipulation

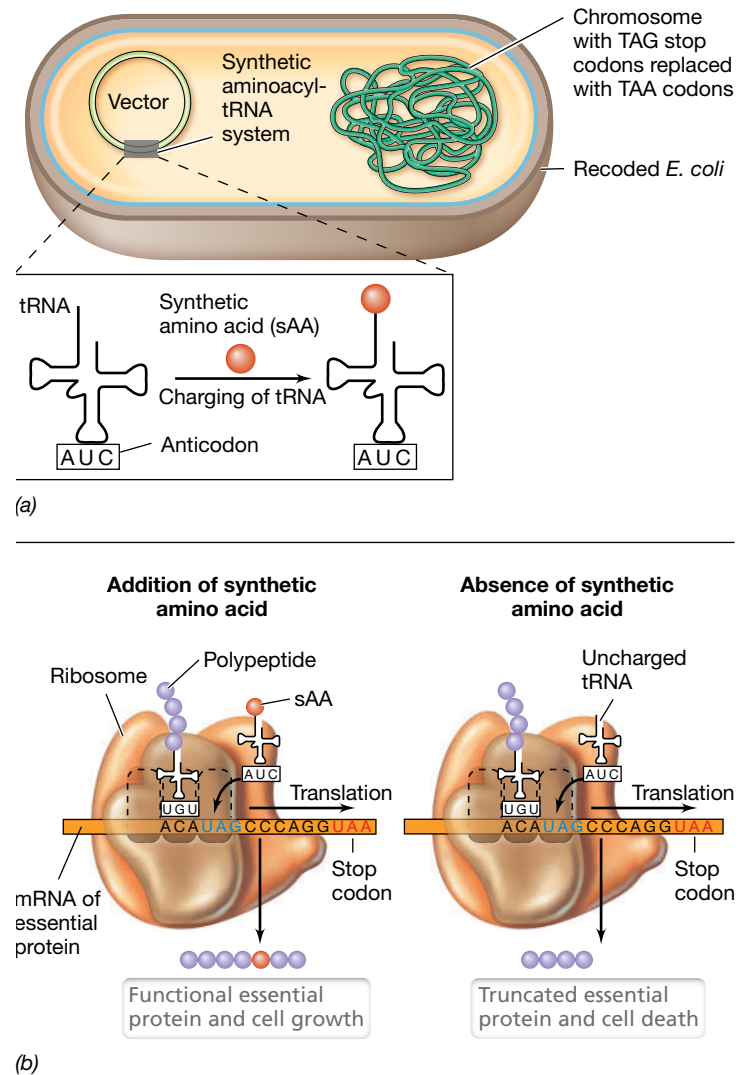


Figure 12.39 Recoding and control of genetically modified *Escherichia coli*. (a) The chromosome of an *E. coli* cell is genetically modified (recoded) to replace all TAG stop codons with TAA stop codons. The recoded *E. coli* stably maintains a vector expressing a tRNA with an AUC anticodon and an aminoacyl-tRNA synthetase (↔ Figure 4.33) that charges the tRNA with a synthetic amino acid (sAA). (b) Control of cell growth by inserting a UAG codon into mRNAs of essential proteins. If the synthetic amino acid is added to the growth medium, the recoded *E. coli* will translate functional essential proteins. If the synthetic amino acid is not present, an uncharged tRNA will bind to UAG codons engineered into essential proteins. This results in truncated essential proteins and ultimately cell death.

also freed up the UAG codon to be reassigned another translational function.

To engineer dependence on a synthetic amino acid, genes for an aminoacyl-tRNA synthetase (↔ Section 4.8) that recognizes the synthetic amino acid (sAA) and a corresponding tRNA with an AUC anticodon were expressed from a vector (Figure 12.39a). This resulted in tRNAs with the AUC anticodon carrying the sAA. A set of essential genes was then modified to contain the TAG codon in positions where incorporation of the sAA would not affect protein activity. Thus for the recoded bacterium to translate mRNAs possessing the UAG codon, the cell must be fed the artificial amino acid (Figure 12.39a). If the growth environment does not have the

sAA (as would be the case if the GMO escaped to nature), uncharged tRNAs will enter into the ribosome when a UAG codon is encountered; if this occurs during the translation of an essential protein, translation will stall and a truncated (and nonfunctional) protein will result (Figure 12.39b). This will lead to death of the cell and puts ultimate control of the recoded GMO organism into human hands. Because the TAG codon was placed in three essential genes, it is extremely unlikely that sufficient mutations could arise to remove the GMO's dependence on the presence of the sAA.

While we have focused on how CRISPRs can be used to genetically modify organisms, synthetic biologists are even tinkering with the genome editing system itself to have it specifically clip out target DNA such as recombinant genes under specific conditions. Continued advances in synthetic biology and the control of GMOs not only will allow for more widespread use of these

organisms for synthesizing desired products and therapeutics in carefully controlled production settings, but may also trigger the more extensive use of GMOs to solve the urgent problems that remain in medicine, agriculture, and the environment.

MINIQUIZ

- Why is the use of auxotrophy not a good method for controlling the growth of a genetically modified organism?
- How can a tRNA be engineered to encode for a synthetic amino acid?
- Why is it unlikely that GMOs recoded to depend on a synthetic amino acid will mutate to no longer depend on the exogenously supplied synthetic amino acid?

MasteringMicrobiology®

Visualize, explore, and think critically with Interactive Microbiology, MicroLab Tutors, MicroCareers case studies, and more. MasteringMicrobiology offers practice quizzes, helpful animations, and other study tools for lecture and lab to help you master microbiology.

Chapter Review

I • Tools of the Genetic Engineer

- 12.1** The polymerase chain reaction is a procedure for amplifying DNA in vitro and employs heat-stable DNA polymerases. This amplified DNA is often used for cloning purposes and can be visualized by gel electrophoresis. Complementary nucleic acid sequences may be detected by hybridization.
- Q** How is DNA amplified in a polymerase chain reaction (PCR) procedure? Why do DNA polymerases from thermophilic microbes significantly improve the PCR procedure?
- 12.2** The isolation of a specific gene or region of a chromosome by molecular cloning is done using a cloning vector. Plasmids are useful cloning vectors because they are easy to isolate and purify and are often able to multiply to high copy numbers in bacterial cells. The choice of a cloning host depends on the final application. In many cases the host can be a prokaryote, but in others, it is essential that the host be a eukaryote.
- Q** How does the insertional inactivation of β -galactosidase allow the presence of foreign DNA in a plasmid vector such as pUC19 to be detected?
- 12.3** Many cloned genes are not expressed efficiently in a foreign host. Expression vectors have been developed that both increase transcription of the cloned gene and control the level of transcription. To achieve very high levels of expression of eukaryotic genes in prokaryotes, the expressed gene must be free of introns. This can be accomplished by synthesizing cDNA from the mature mRNA encoding the protein of interest or by making an

entirely synthetic gene. Protein fusions are often used to stabilize or solubilize the cloned protein.

Q What is the significance of reverse transcriptase in the cloning of animal genes for expression in bacteria?

- 12.4** Synthetic DNA molecules of desired sequence can be made in vitro and used to construct a mutated gene directly or to change specific base pairs within a gene by site-directed mutagenesis. Also, genes can be disrupted by inserting DNA fragments, called cassettes, into them, generating knockout mutants.
- Q** What does site-directed mutagenesis allow you to do that normal mutagenesis does not?
- 12.5** Reporter genes are genes whose products are easy to assay or detect. They are used to simplify and increase the speed of genetic analysis. In gene fusions, segments from two different genes, one of which is usually a reporter gene, are spliced together.
- Q** What is the key property of a reporter gene?
- ### II • Making Products from Genetically Engineered Microbes: Biotechnology
- 12.6** The first human protein made commercially using engineered bacteria was human insulin. Recombinant bovine somatotropin is widely used in the United States to increase milk yield in dairy cows.
- Q** What classes of mammalian proteins are produced by biotechnology? How are the genes for such proteins obtained?

12.7 Genetic engineering can make plants resistant to disease and improve product quality. The Ti plasmid of the bacterium *Agrobacterium tumefaciens* can transfer DNA into plant cells. Genetically engineered commercial plants are called genetically modified organisms (GMOs).

Q What is the Ti plasmid and how has it been of use in genetic engineering?

12.8 Many recombinant vaccines have been produced or are under development. These include live recombinant, vector, and subunit vaccines. Properties associated with pathogens can be used to develop cancer-treating therapeutics.

Q What is a subunit vaccine and why are subunit vaccines considered a safer way of conferring immunity to viral pathogens than attenuated virus vaccines?

12.9 Genes for useful products may be cloned directly from DNA or RNA in environmental samples without first isolating the organisms that carry them. In pathway engineering, genes that encode the enzymes for a metabolic pathway are assembled. These genes may come from one or more organisms, but the engineering must achieve regulation of the coordinated sequence of expression required in the pathway.

Q How has metagenomics improved the discovery of novel, useful products in biotechnology?

12.10 While select microorganisms can produce biofuels in small amounts, others can be modified to produce various biofuels through pathway engineering. This modification

often requires genes from multiple microorganisms. New tools for genetically manipulating microalgae have also been developed to facilitate biofuel production.

Q How can biodiesel be produced from microalgae-biosynthesis products?

III • Synthetic Biology and Genome Editing

12.11 Instead of modifying or improving a single existing pathway, synthetic biology focuses on engineering novel biological systems by linking known biological components together in various combinations. These modifications can result in the production of high-value products.

Q How does synthetic biology differ from engineering of *Escherichia coli* to produce indigo?

12.12 Not only can CRISPR systems be used as a prokaryotic immune system, they can also be modified to edit the genomes of eukaryotes.

Q How has the CRISPR editing technology been applied to targeting virus-infected eukaryotic cells?

12.13 With the advancements in genetic engineering, methods for controlling genetically modified organisms are imperative. One promising method is the use of synthetic biology to recode organisms for dependence on the presence of synthetic amino acids.

Q What are some mechanisms for controlling a genetically modified organism other than making it dependent on synthetic amino acids?

Application Questions

- Suppose you have just determined the DNA base sequence for an especially strong promoter in *Escherichia coli* and you are interested in incorporating this sequence into an expression vector. Describe the steps you would use. What precautions are necessary to be sure that this promoter actually works as expected in its new location?
- Many genetic systems use the *lacZ* gene encoding β -galactosidase as a reporter. What advantages or problems would there be if (a) luciferase or (b) green fluorescent protein were used instead of β -galactosidase as reporters?
- You have just discovered a protein in mice that may be an effective cure for cancer, but it is present only in tiny amounts. Describe the steps you would use to produce this protein in therapeutic amounts. Which host would you want to clone the gene into and why? Which host would you use to express the protein in and why?
- Describe how you could recode *Escherichia coli* to produce novel proteins containing more than the standard 22 amino acids.

Chapter Glossary

Bacterial artificial chromosome (BAC) a circular artificial chromosome with a bacterial origin of replication

Biotechnology the use of organisms, typically genetically altered, in industrial, medical, or agricultural applications

Cassette mutagenesis creating mutations by the insertion of a DNA cassette

Complementary DNA (cDNA) DNA made from an RNA template during the reverse transcription PCR (RT-PCR) procedure

DNA cassette an artificially designed segment of DNA that usually carries a gene for resistance to an antibiotic or some other convenient marker and is flanked by convenient restriction sites

Expression vector a cloning vector that contains the necessary regulatory sequences to allow transcription and translation of cloned genes

Gel electrophoresis a technique for separation of nucleic acid molecules by passing an electric current through a gel made of agarose or polyacrylamide

Gene disruption (also called gene knockout) the inactivation of a gene by insertion of a DNA fragment that interrupts the coding sequence

Gene fusion a structure created by joining together segments of two separate genes, in particular when the regulatory region of one gene is joined to the coding region of a reporter gene

Genetically modified organism (GMO) an organism whose genome has been altered using genetic engineering; the abbreviation GM is also used in terms such as GM crops and GM foods

Genetic engineering the use of in vitro techniques in the isolation, alteration, and expression of DNA or RNA and in the development of genetically modified organisms

Green fluorescent protein (GFP) a protein that glows green and is widely used in genetic analysis

Heterologous expression transcription and translation of a gene or genes from one organism in a different organism

Hybridization the formation of a double helix by the base pairing of single strands of DNA or RNA from two different sources

Molecular cloning the isolation and incorporation of a fragment of DNA into a vector where it can be replicated

Northern blot a hybridization procedure where RNA is the target and DNA or RNA is the probe

Nucleic acid probe a strand of nucleic acid that can be labeled and used to hybridize to a complementary molecule from a mixture of other nucleic acids

Operon fusion a gene fusion in which a coding sequence that retains its own translational signals is fused to the transcriptional signals of another gene

Pathway engineering the assembly of a new or improved biochemical pathway using genes from one or more organisms

Polymerase chain reaction (PCR) the artificial amplification of a DNA sequence by repeated cycles of strand separation and replication

Polyvalent vaccine a vaccine that immunizes against more than one disease

Protein fusion a gene fusion in which two coding sequences are fused so that they share the same transcriptional and translational start sites

Recombinant DNA a DNA molecule containing DNA originating from two or more sources

Reporter gene a gene used in genetic analysis because the product it encodes is easy to detect

Restriction enzyme an enzyme that recognizes a specific DNA sequence and then cuts the DNA; also known as a restriction endonuclease

Site-directed mutagenesis construction in vitro of a gene with a specific mutation

Southern blot a hybridization procedure where DNA is the target and RNA or DNA is the probe

Subunit vaccine vaccine that contains only a specific protein or two from a pathogen

T-DNA the segment of the *Agrobacterium tumefaciens* Ti plasmid that is transferred into plant cells

Ti plasmid a plasmid in *Agrobacterium tumefaciens* capable of transferring genes from bacteria to plants

Transgenic organism a plant or an animal with foreign DNA inserted into its genome

Vector (as in cloning vector) a self-replicating DNA molecule that is used to carry cloned genes or other DNA segments for genetic engineering

Vector vaccine a vaccine made by inserting genes from a pathogenic virus into a relatively harmless carrier virus

Yeast artificial chromosome (YAC) an artificial chromosome with a yeast origin of replication and a centromere sequence

Microbial Evolution and Systematics

13

microbiologynow

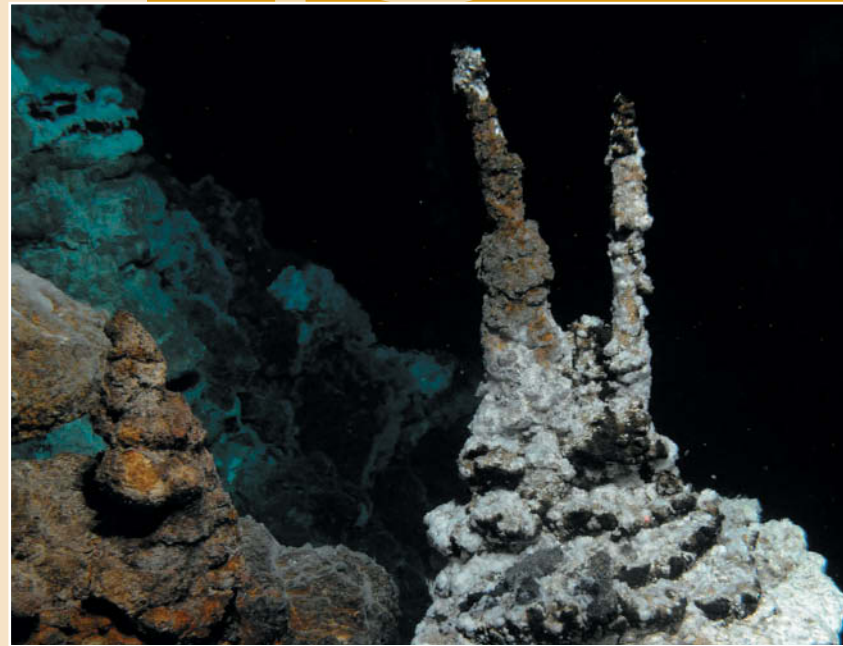
Lokiarchaeota and the Origin of *Eukarya*

The domain *Eukarya* includes plants, animals, fungi, and a tremendous diversity of microorganisms. The plants, animals, and fungi are relative newcomers on the scene, as their evolutionary origins occurred some 400–600 million years ago. In contrast, the first eukaryotic microbes originated well over a billion years ago. The evolutionary origin of the eukaryotic cell remains enigmatic and we still do not know when or how the domain *Eukarya* was formed.

Genomic analyses clearly reveal *Eukarya* to be genetic chimeras. Eukaryotic genomes contain a mixture of genes that originated either within the *Bacteria* or within the *Archaea*, as well as many genes that are unique to *Eukarya*. Most evidence suggests that *Eukarya* share an ancestor with the domain *Archaea*, but *Eukarya* contain numerous “signature genes” not found in the *Archaea*. These unique eukaryotic genes encode proteins associated with the distinctive cell biology of *Eukarya* and were likely essential for the origins of multicellularity, a property widespread in the eukaryotic world.

The recent discovery of *Lokiarchaeota*—a new phylum of *Archaea*—has provided fresh insights into the origin of *Eukarya*. *Lokiarchaeota* were discovered through metagenomic analyses of microbial communities that inhabit deep marine sediments near a hydrothermal vent system known as Loki’s Castle (see photo), located along the Mid-Atlantic Ridge between Greenland and Norway. Remarkably, the genomes of *Lokiarchaeota* contain a number of eukaryotic signature genes, and in particular, genes associated with membrane remodeling and the development of a cytoskeleton. The presence of a cytoskeleton and the ability to remodel intracellular membranes would have facilitated membrane invagination in primitive eukaryotic cells, and this would have allowed bacterial endosymbionts to be acquired and provided new nutritional strategies, such as phagocytosis.

These results suggest that features uniquely associated with the eukaryotic cell may actually have their origins in the domain *Archaea*. The discovery of the *Lokiarchaeota* also indicates that, rather than emerging as a sister group to the *Archaea*, the earliest eukaryotic cells emerged from within the *Archaea* following the endosymbiotic acquisition of the bacteria that gave rise to the eukaryotic cell’s respiratory organelle, the mitochondrion. Hence, the first steps toward the origins of cellular complexity may have occurred within the domain *Archaea*.



- I Early Earth and the Origin and Diversification of Life 400
- II Microbial Evolution 408
- III Microbial Phylogeny and Systematics 412