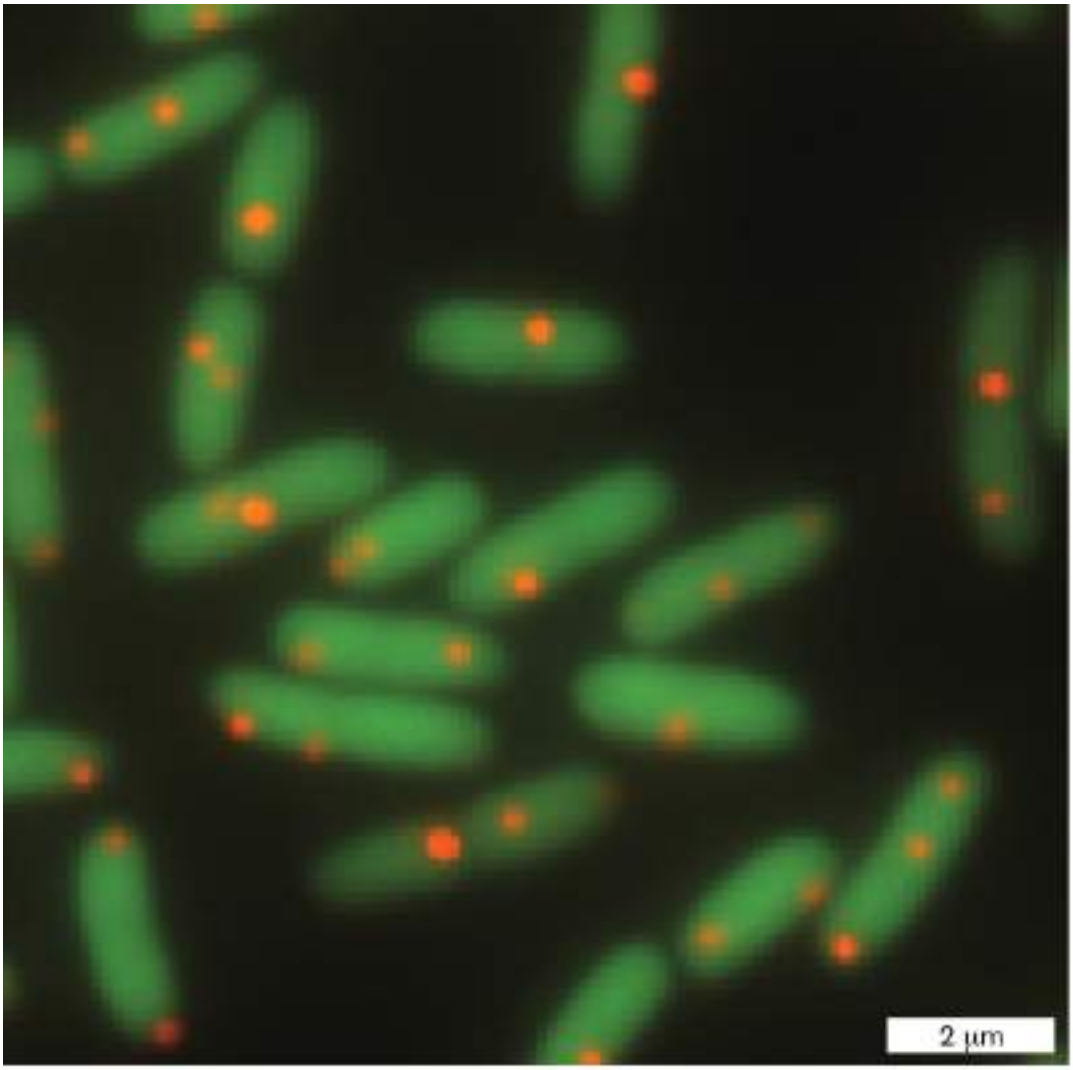
CHAPTER 7

Genomes and Chromosomes



Fluorescence microscopy reveals location and abundance of plasmids within host cells. Locations of

engineered plasmids in cells of *Escherichia coli* are revealed by fluorescent proteins that bind to specific regions of the plasmid DNA (orange). These plasmids carry the gene encoding green fluorescent protein (GFP), which can diffuse freely throughout the cell. The green fluorescence allows researchers to quantify the plasmid copy number for each *E. coli* cell: The more copies of the plasmid an *E. coli* cell carries, the more GFP proteins made, and the brighter green the cytoplasm fluoresces.

Source: Bin Shao et al. 2021. *Nat. Commun.* **12** :1475.

B. SHAO ET AL. 2021. *NAT COMMUN* . **12** :1475

Chapter Sections		Learning Objectives
7.1 DNA: The Genetic Material	»	Discuss the evidence that led researchers to identify DNA as the genetic material of cellular organisms.

Chapter Sections		Learning Objectives
7.2 Genome Organization	»	Describe the structure and function of DNA and RNA and the measures that bacteria take to compact their chromosomes into condensed nucleoid structures.
	»	Describe what genetic elements are present in a prokaryotic genome.
Special Topic 7 DNA as Digital Storage		
7.3 DNA Replication	»	Outline the key steps in chromosome replication, including initiation, elongation, and termination.
7.4 Plasmids and Secondary Chromosomes	»	Describe the features, functions, and replication mechanisms of plasmids and secondary chromosomes that distinguish them from primary chromosomes.

Chapter Sections		Learning Objectives
7.5 Eukaryotic and Archaeal Chromosomes	»	Compare and contrast the structure, composition, and replication of bacterial, archaeal, and eukaryotic genomes.
7.6 Microbiomes and Metagenomes	»	Describe how metagenomes and single-cell genomes are prepared and their utility in capturing the genetic potential of uncultured microbes.
eResearch Activity 7 Can Single-Cell Genomics Reveal Prebiotic Consumers of the Gut Microbiome?		

A genome is all of the genetic information that defines an organism. For a bacterial species, this can include one or more chromosomes, plasmids, and viruses. In Chapter 7 we explore the physical structure of DNA and how it is replicated, and we discuss how bacterial genomes replicate, organize, and migrate through the dividing bacterial cell. We compare these bacterial genome properties with those of archaea and eukaryotic microbes. We also describe how innovations in molecular biology help researchers explore the genomes in natural microbial communities.

In the subsequent chapters of Part 2 of this textbook, we describe how genes are encoded and get "read" to make products (for example, RNA, protein) for the cell (Chapter 8), how genes

and genomes can change through mutation (Chapter 9), and how gene expression is regulated by internal and external stimuli (Chapter 10). The remaining chapters of Part 2 discuss the specialized genetic mechanisms of viruses (Chapter 11) and how cells coordinate molecular processes through space and time to facilitate chemotaxis, biofilm development, defenses and counterdefenses against phage infection, and engineering of novel microbes via synthetic biology (Chapter 12).

7.1 DNA: The Genetic Material

Genes are the units of heredity, and most genes are located on chromosomes. Early in the twentieth century, chromosomes were known to contain protein as well as nucleic acid, but which molecule served as the genetic material was uncertain. For many years scientists believed erroneously that the protein component is what served as the genetic material, while the DNA might serve simply as a scaffold for the proteins. In support of the protein hypothesis was the fact that there are far more amino acid variants (20) than nucleic acid variants (5, if RNA's uracil and DNA's thymine are considered). The larger array of protein-building components suggested the possibility of a more diverse set of genes because the genetic alphabet would have more letters.

Studies of bacteria and their viruses in the 1920s–1950s finally overturned the protein hypothesis and revealed DNA as the genetic material. In 1928, Frederick Griffith (1879–1941) discovered that he could kill mice with live but seemingly harmless (avirulent) *Streptococcus pneumoniae* , but only if he coinjected the mice with dead cells from a virulent strain of the bacteria. Something from the dead bacteria transformed the innocuous live bacteria into killers. Through careful biochemical analysis in 1944, Oswald Avery, Colin MacLeod, and Maclyn McCarty demonstrated that this transforming agent was DNA, and they coincidentally discovered the first of several mechanisms of horizontal gene transfer between microbes: transformation .

Overturning a long-standing model often requires multiple lines of evidence, and in the case of DNA, the evidence came from the 1952 study by Alfred Hershey and Martha Chase on the *Escherichia coli* –infecting virus T2. Using radioisotopes of sulfur and phosphorus to track the protein and DNA components of the virus, respectively, Hershey and Chase showed that during viral adsorption and replication, the viral protein coat remains on the cell exterior,

while the DNA penetrates into the interior and serves as the code to produce new viral DNA and protein.

While all known cellular organisms on Earth use DNA for their genetic material, some viruses such as influenza and SARS-CoV-2 use RNA (discussed in Chapter 6). Additionally, the first lifeforms on Earth were thought to be encoded by RNA, at a time in evolution referred to as the RNA world (described in Chapter 17).

Thought Question

7.1 Before the studies by Avery, Hershey, and others, some scientists believed that, unlike plants and animals, bacteria lacked genes. Considering what little was known about the modes of reproduction and the recombination of alleles in bacteria, why was this a reasonable, albeit incorrect, assumption?

To Summarize

- **Genes** are units of heredity.
- **Studies in bacteria** provided the critical evidence that the genetic material is composed of DNA, not protein.
- **RNA** serves as the genetic material for some viruses, including SARS-CoV-2.

Glossary

gene

A sequence of nucleotides that has a distinct function (regulatory) or whose encoded product (protein or RNA) has a distinct function. The functional unit of heredity.

transformation

The internalization of free DNA from the environment into bacterial cells.

7.2 Genome Organization

The genome is the total of all genes in an organism, whether they are all located on a single chromosome or distributed among multiple nucleic acid molecules that replicate independently in the cell. With the exception of RNA viruses, microbial genomes are encoded by DNA. DNA is an ideal molecule to encode the genome because (1) it is stable and not readily degraded within the lifetime of the cell; (2) it is mutable and thus can allow the organism to evolve through heritable genetic change; and (3) as a double-stranded polymer it is readily replicated by using the original strands as the template for making copies during cell replication.

By 1977, researchers had completed only one genome sequence: that of the *Escherichia coli* –infecting phage phiX174 (φX174). Over 40 years later, we have over 100,000 well-curated genomes, and many more on the way (introduced in Chapter 1). We now know that genome organization varies widely in the microbial world. While most of the prokaryotic genomes sequenced have a single chromosome, about 10% have more than one. Many genomes also include extrachromosomal elements such as plasmids and viruses, some of which can insert into the host chromosome. Researchers continue to ask questions about microbial genomes. Why do some organisms divide their genome into multiple chromosomes, while others use just a single chromosome? How does a microbe acquire a second chromosome? What is the advantage of gene placement on plasmids instead of chromosomes? How does a bacterium replicate its chromosomes and plasmids, and how is this replication coordinated with cell division? And finally, do the answers to these questions differ between bacteria and archaea?

Microbes are tremendously diverse (see Chapters 18–20 for examples). They are found in every imaginable habitat (for example, ocean sediments, the human colon), carrying out all kinds of processes (photosynthesis and anaerobic respiration, for instance) and performing an extraordinary variety of activities (such as

swimming, fruiting-body formation, or invasion of animal or plant cells). These lifestyle differences are often reflected in the size and content of microbial genomes. In addition, the physical arrangement of genes within a genome can vary from microbe to microbe. As we will see, certain genome arrangements provide clear advantages for the cell, whereas in other cases the advantages are not yet known.

Genomes Vary in Size

Bacterial and archaeal genomes range in size from approximately 106 to 16,000 kilobase pairs (kb). For comparison, eukaryotic genomes range from 2,900 kb (Microsporidia) to more than 100,000,000 kb (flowering plants). The human genome is over 3,000,000 kb.

Note: The designation "kb" can refer to the length of either a double-stranded or a single-stranded DNA molecule. A bacterial genome is, by definition, double-stranded. Some viral genomes are single-stranded.

One of the smallest cellular genomes sequenced thus far is that of the candidate bacterial species *Tremblaya princeps* (**Table 7.1**). The complete genome of *T. princeps* consists of only 139 kb and encodes a mere 120 proteins. It lacks the genes required for many biosynthetic functions, including several genes for the translation of messenger RNA (mRNA) into protein (see Chapter 8). With the absence of so many genes, how does this organism survive? Survival depends on the peculiar symbiotic lifestyle of *T. princeps* . It lives as an endosymbiont inside mealybug insects, and in turn, *T. princeps* plays host to smaller microbes living inside its own cytoplasm (**Fig. 7.1**). Evolution likely permitted the dramatic reduction in the *T. princeps* genome because its symbiotic partners could provide many of those lost gene functions. For additional examples of genomic reduction of host-associated microbes, see the discussions of the

evolution of mitochondria and chloroplasts (Section 17.6) and of obligate intracellular pathogens (Chapter 23).

TABLE 7.1 Genomes of Representative Bacteria and Archaea

Species (strain)	Chromosome(s)* (kilobase pairs, kb) Circular and linear	Plasmid(s)* (kb) Circular	Total (kb)
Bacteria			
Tremblaya princeps Endosymbiont of mealybugs	139		139
Mycoplasma genitalium Normal flora, human skin	580		580
***Vibrio cholerae* (El Tor N16961)** Cholera	2,961 + 1,072		4,033
Mycobacterium tuberculosis Tuberculosis	4,400		4,400
***Escherichia coli* K-12 (W3110)** Model strain for *E. coli* research	4,600		4,600
Agrobacterium tumefaciens Tumors in plants; genetic engineering vector	2,840 + 2,070	214 + 542	5,666
***Anabaena* species (PCC 7120)** Cyanobacteria; major photosynthetic producer of carbon source for freshwater ecosystems	6,370	110 + 190 + 410	7,080
Borrelia burgdorferi Lyme disease	911	21; sizes 9 to 58	>1,250
Archaea			
Methanocaldococcus jannaschii Methanogen from thermal vent	1,660	16 + 58	1,734
Haloarcula marismortui Halophile from volcanic vent	3,130 + 288	33 + 33 + 39 + 50 + 155 + 132 + 410	4,270

1,000 kb

500 kb

*Purple circles and lines indicate relative sizes of genomic elements and whether these are circular or linear. Size bars are provided under each column.

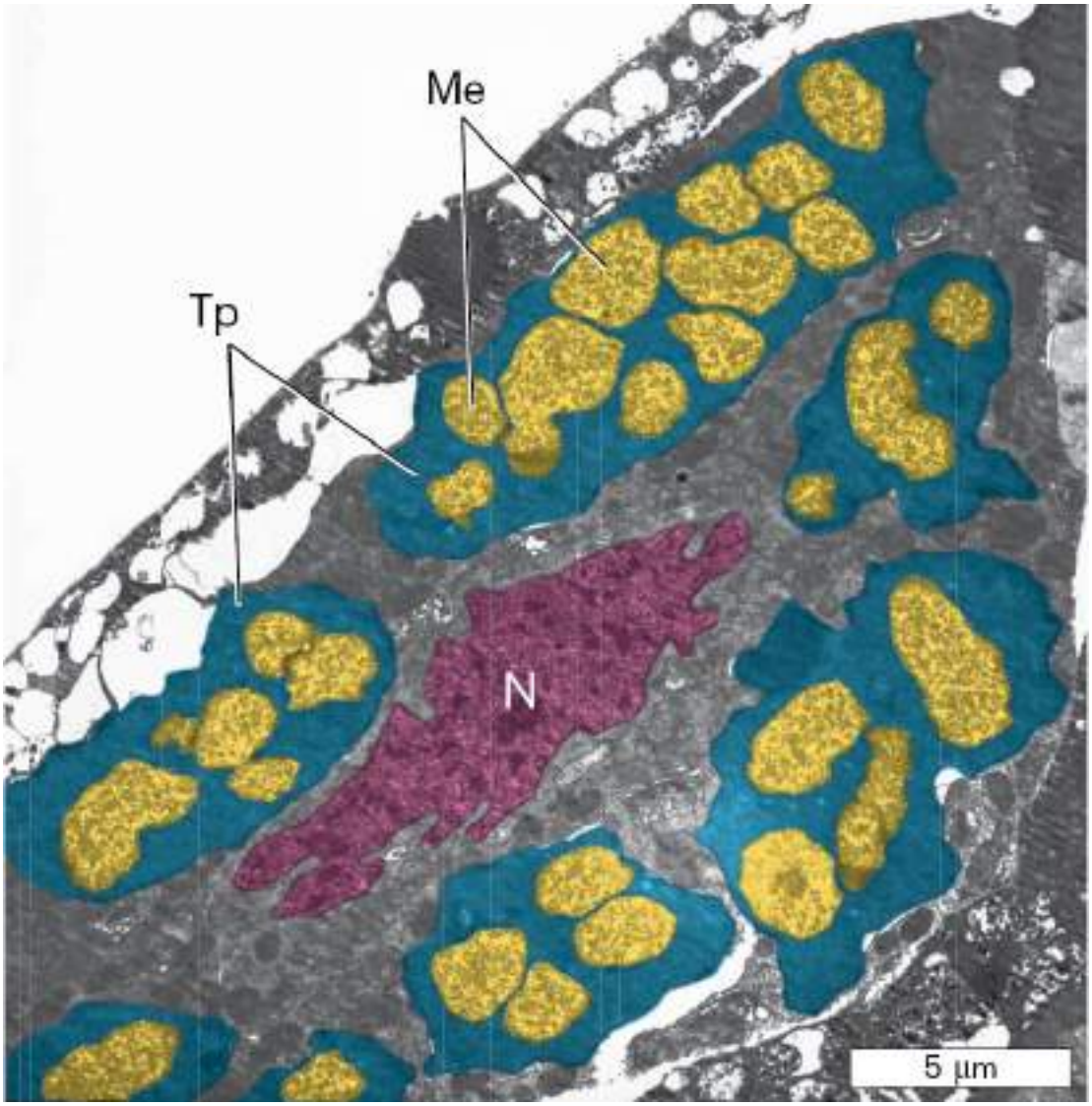


FIGURE 7.1 ■ The mealybug endosymbiont *Tremblaya princeps*. *T. princeps* cells (Tp; the leader lines point to cytoplasm of two teal *T. princeps* cells), living inside mealybug *Planococcus citri* cells (N = nucleus), are themselves host to other endosymbiotic bacteria, the candidate species *Moranella endobia* (Me; yellow cells). Colors are added for clarity in this TEM micrograph.

COURTESY OF CAROL VON DOHIEN

In contrast, many bacteria that grow as free-living cells have larger genomes and dedicate many genes to the synthesis or acquisition of amino acids or tricarboxylic acid (TCA) cycle

intermediates. One example is *Vibrio cholerae* , the causative agent of cholera. Its 4,033-kb genome is divided into two chromosomes, of 2,961 and 1,072 kb (**Fig. 7.2**). Are genomes like that of *V. cholerae* divided into two chromosomes because there is a size restriction per chromosome, perhaps because of difficulties associated with replication or packaging? This does not appear to be the case, as genomes of much larger size have been found on single chromosomes; for example, the single chromosome of the bacterium *Myxococcus xanthus* is over 9,000 kb. We will consider other possible explanations later in the chapter.

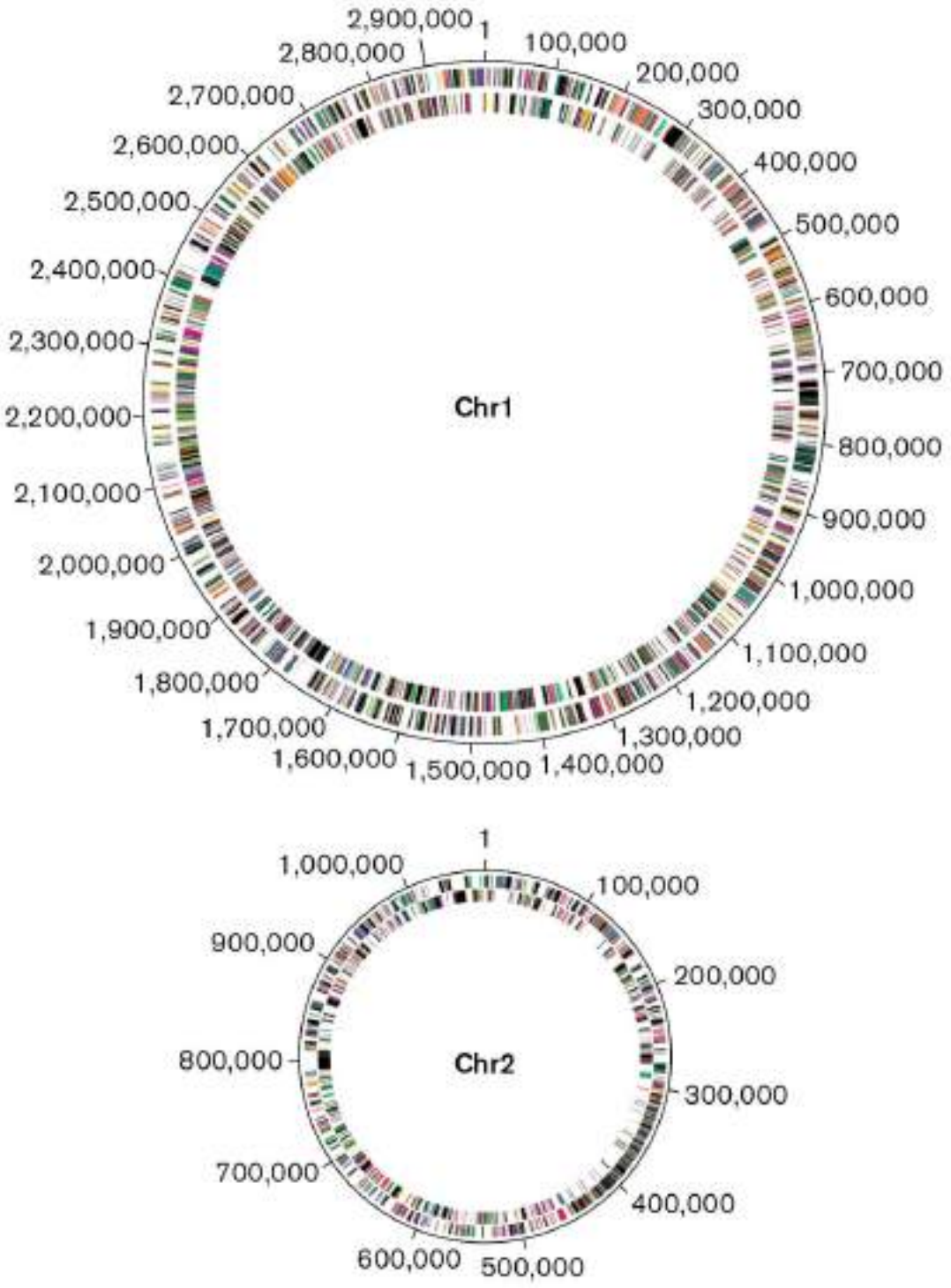


FIGURE 7.2 ■ The two chromosomes of the *Vibrio cholerae* genome. In each chromosome, the numbers refer to the position in base pairs, starting at the origin of replication and extending clockwise. Genes encoded on the plus and minus

strands of the chromosome are depicted on the outer and inner rings, respectively. Chr1 = primary chromosome; Chr2 = secondary chromosome.

SOURCE : MODIFIED FROM J. F. HEIDELBERG ET AL. 2000. *NATURE* **406** :477–483, FIG. 2.

Because eukaryotic chromosomes are linear, scientists initially expected that bacterial chromosomes would be linear too. However, the early genetic maps for bacteria such as *Escherichia coli* just would not fit together in a manner consistent with a linear model. We now know that most bacteria (including *E. coli*) and archaea have circular chromosomes. Some species, however, do have linear chromosomes or even a mix of linear and circular chromosomes. Examples include the Lyme disease agent *Borrelia burgdorferi* and the plant tumor agent *Agrobacterium tumefaciens* (**Table 7.1**). We will revisit the structure of chromosomes later in this chapter and consider the advantages of linear versus circular forms.

DNA Function Depends on Its Chemical Structure

Chapter 1 describes the discovery of the double-helix structure of DNA and the contributions of Rosalind Franklin, James Watson, and Francis Crick to this landmark achievement. DNA is composed of four different nucleotides linked by a phosphodiester backbone (**Fig. 7.3A**). Each nucleotide consists of a nucleobase (also called a nitrogenous base) attached through a ring nitrogen to carbon 1 of 2-deoxyribose in the phosphodiester backbone. The 2-deoxy position that distinguishes DNA from RNA (**Fig. 7.3B**) is highlighted in the figure. A phosphodiester link (marked in **Fig. 7.3A**) joins adjacent deoxyribose molecules in DNA to form the phosphodiester backbone. Phosphodiester links connect the 3′ carbon of one deoxyribose to the 5′ carbon of the next deoxyribose. The two backbones are antiparallel so that at either end of a linear DNA molecule, one DNA strand ends with a 3′ hydroxyl group and the complementary strand ends with a 5′ phosphate. This antiparallel arrangement is necessary

so that complementary bases protruding from the two strands can pair properly via hydrogen bonding.

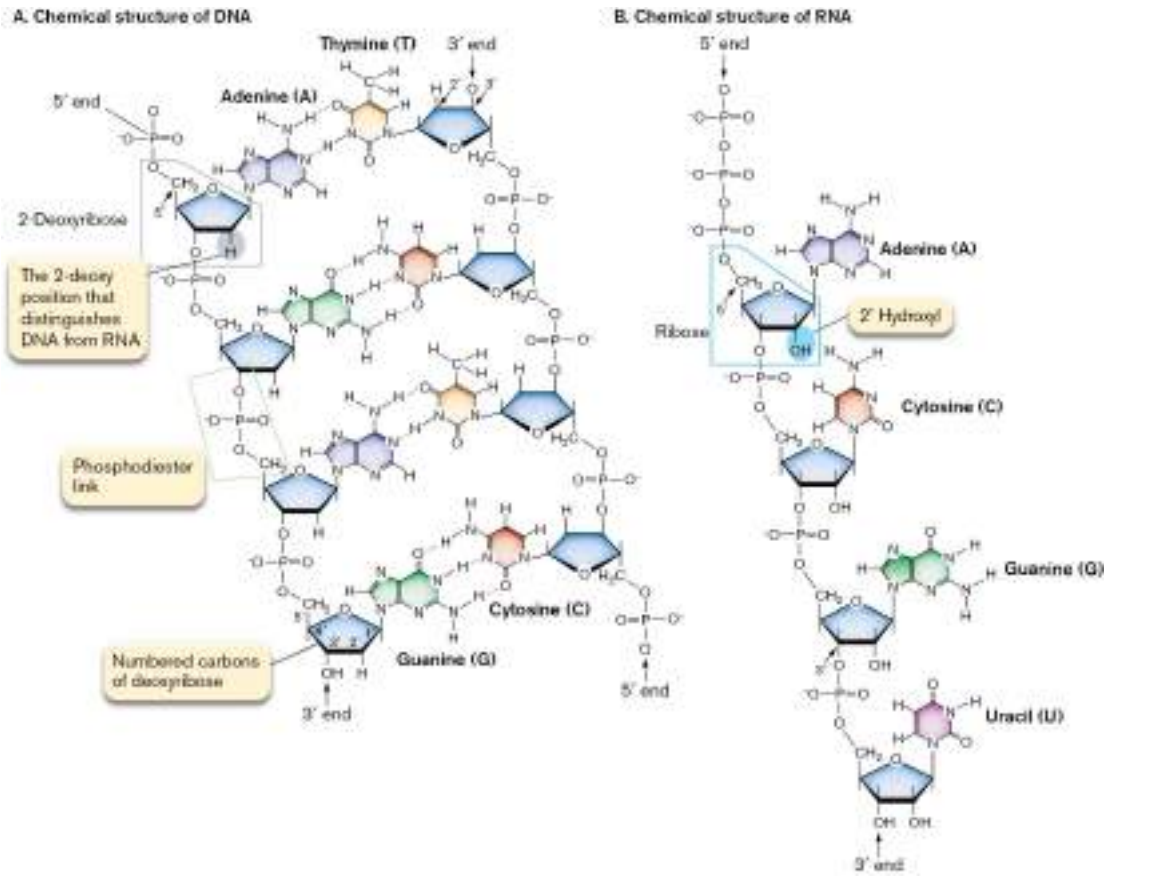


FIGURE 7.3 ■ Structures of DNA and RNA. A. In the cell, DNA bases are added only to a preexisting 3′ OH of a nucleoside monophosphate, so the 5′ ends in this figure are drawn as nucleoside monophosphates. (Dotted lines indicate hydrogen bonds between bases.) **B.** Cellular RNA molecules, however, begin with a 5′ triphosphate.

The nucleobases in DNA are planar heteroaromatic structures (aromatic rings that include nitrogen) arranged perpendicular to the phosphodiester backbone and parallel to each other like a stack of coins. Purines (adenine or guanine) pair with pyrimidines (thymine or cytosine). Under physiological conditions of salt (about 0.85% NaCl) and pH (pH 7.8), the hydrogen bonding of the bases permits adenine to pair only with thymine (via two hydrogen bonds) and likewise

guanine with cytosine (via three hydrogen bonds). These complementary base interactions enable the two phosphodiester backbones to wrap around each other to form the classic double helix, or duplex.

The H-bonds that form between purines and pyrimidines along the interior of a DNA duplex (**Fig. 7.3**) make the bonding of the two complementary strands of DNA highly specific. Although H-bonds govern the specificity of strand pairing, the thermal stability of the helix is due predominantly to the stacking of the hydrophobic base pairs. Stacking of base pairs excludes water from the hydrophobic helix interior while allowing water and ions to interact with the hydrophilic, negatively charged phosphate backbone of DNA.

In the space-filling model of DNA in **Figure 7.4A** and the contour map in **Figure 7.4B** , notice that the DNA double helix has grooves: a wide major groove and a narrower minor groove. The two grooves are generated by the angles at which the paired bases meet each other. These grooves provide DNA-binding proteins access to base sequences buried in the center of the molecule, so that the proteins can interact with the bases without the strands being separated.

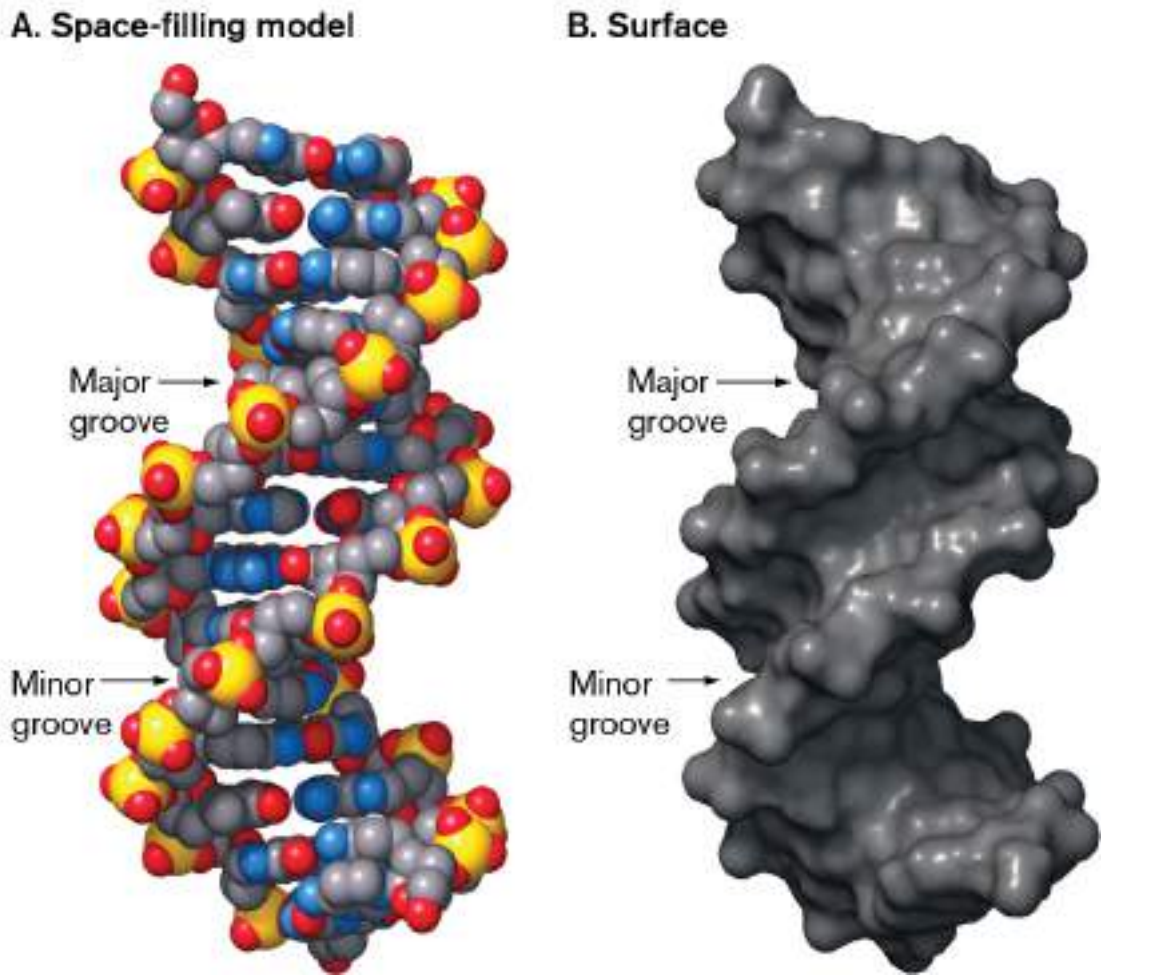


FIGURE 7.4 ■ Models of DNA. A. Space-filling model of DNA. **B.** DNA surface, modeled using nuclear magnetic resonance. (PDB code: 1K8J)

Figure 7.5 shows an example of an important DNA-binding protein, DtxR of *Corynebacterium diphtheriae* , the cause of diphtheria. The DtxR dimer (pair of subunits) binds the major groove of DNA at the promoter for the diphtheria toxin gene (*dtxA*). Toxin expression is repressed in the presence of a high iron concentration, which signals to the bacteria that they are in an environment outside a human host. Gene regulation is discussed further in Chapter 10 .

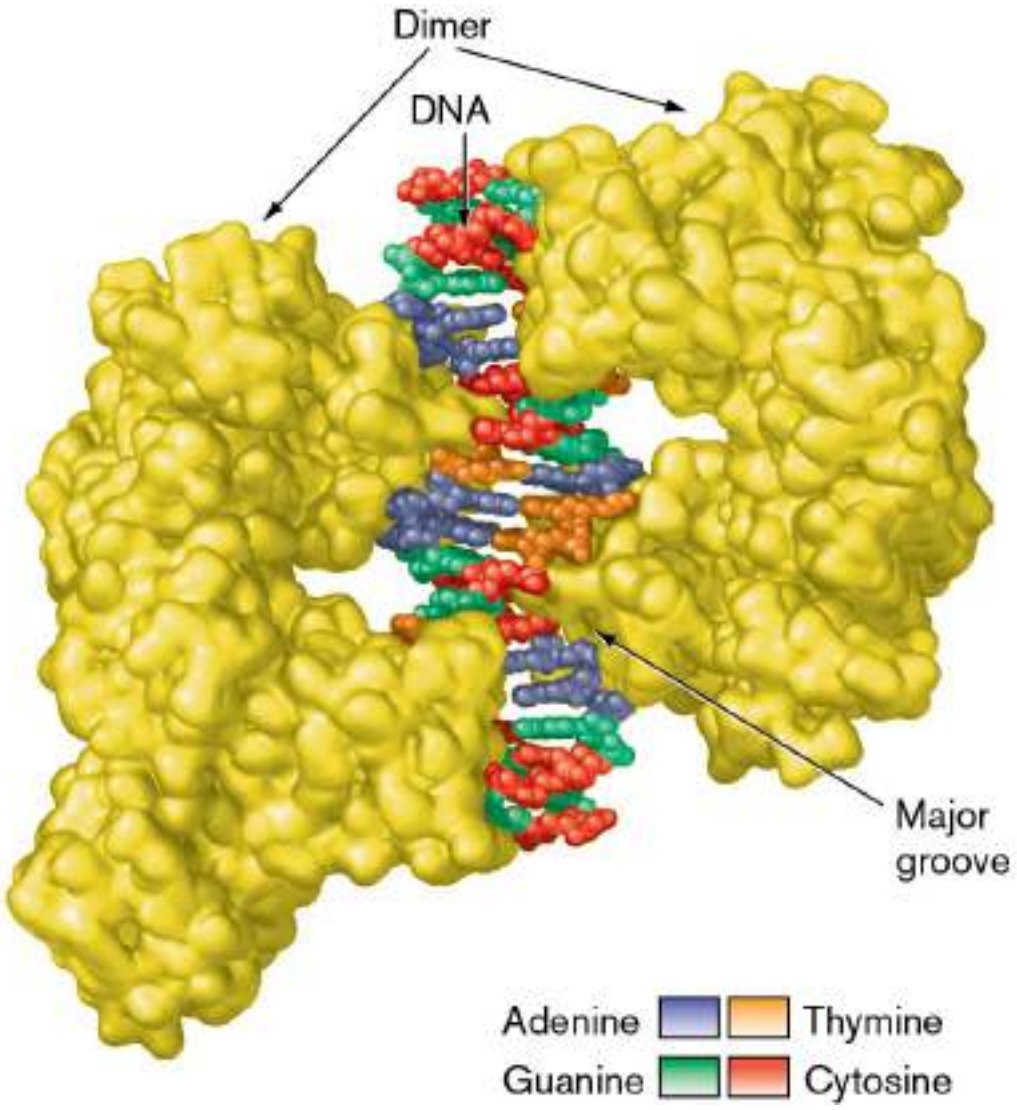


FIGURE 7.5 ■ Protein that recognizes DNA. DtxR repressor protein dimer binds in the major groove of DNA. (PDB code: 1C0W)

At high temperatures (50°C–90°C), the hydrogen bonds in DNA break and the duplex falls apart, or denatures , into two single strands. The temperature required to denature a DNA molecule depends on the GC/AT ratio of a sequence. More energy is required to break the three H-bonds of a GC base pair than the two H-bonds of an AT base pair. Thus, DNA with a high GC content requires a higher temperature to denature than does similar-sized DNA with a lower GC content.

When DNA has been heated to the point of strand separation, lowering the temperature permits the two single strands to find each other and reanneal into a stable double helix. The kinetics of DNA renaturation is much slower than that of denaturation, because renaturation is a random, hit-or-miss process of complementary sequences finding each other. This melting/reannealing property of DNA is exploited in a number of molecular techniques (see the discussions of gene cloning and the polymerase chain reaction in eAppendix 3). Note, however, that bacteria and archaea growing at extreme pH or temperature protect their DNA from denaturation through the use of remarkable DNA-binding proteins, such as the archaeal histone proteins Hmf and Htz. The role of DNA-binding proteins in microbial survival is the subject of considerable research.

Thought Questions

7.2 What do you think happens to two single-stranded DNA molecules isolated from different genes when they are mixed together at very high concentrations of salt? *Hint:* High salt concentrations favor bonding between hydrophobic groups.

7.3 How do the kinetics of denaturation and renaturation depend on DNA concentration?

RNA Differs from DNA

Why do cells make both RNA and DNA? As we learned in Chapter 3 , DNA and RNA have different missions. The cell uses DNA to stably archive the information needed to make a functional cell. In fact, the high stability of DNA makes it possible for human engineers to record and store digital information on this biomolecule, much like a silicon-based hard drive. This novel use of DNA to archive data is discussed in Special Topic 7 .

The growing cell continually accesses the information stored in DNA by making temporary copies of its genes in the form of RNA molecules that direct the synthesis of proteins. The cell also makes RNA molecules that behave as enzymes or can modulate expression of genes. To keep their roles separate, DNA and RNA must have slightly different structures.

DNA in a cell usually consists of two complementary strands, whereas RNA usually consists of a single strand. DNA and RNA are chemically similar, except that in RNA the sugar ribose replaces deoxyribose and the pyrimidine base uracil replaces thymine (see **Fig. 7.3B**). Functionally, these two differences prevent enzymes meant to work on DNA, such as DNA polymerases, from acting on RNA. They also prevent RNA nucleases (RNases) from degrading DNA. However, uracil can base-pair with adenine, which means that hybrid RNA-DNA double-stranded molecules can form (hybridize) when base sequences are complementary. In fact, this hybridization is a necessary step in the decoding of genes to make proteins.

Although RNA molecules are commonly thought of as single-stranded, each single-stranded RNA molecule has regions that loop back to form "hairpins." Hairpin structures form when complementary nucleotide sequences within the primary RNA sequence bend back and hybridize. These double-stranded hairpins have a variety of biological functions.

Bacterial Chromosomes Are Compacted into a Nucleoid

The chromosome of *E. coli* has over 4.6 million bases in one strand, or over 9 million, counting both strands. This is a huge molecule. At the normal pH of the cell (7.8), all the phosphates in the backbone (all 9 million of them) are unprotonated and negatively charged, so this one molecule contributes greatly to the overall negative charge of the cytoplasm.

Figure 7.6 shows DNA spewing out of a damaged bacterial cell. Laid out, the chromosome is 1,500 times longer than the cell. It

is obvious from this photomicrograph that an intact, healthy cell must compact a huge bundle of DNA into a very small volume. DNA is the second-largest molecule in the cell (only peptidoglycan is larger) and constitutes a large portion of a bacterial cell's dry mass, about 3%–4%. Although packaging 3% of a cell's dry weight may not seem like a challenge, note that DNA is further confined only to ribosome-free areas of the cell, so the chromosome-packing density reaches about 15 mg/ml. In a test tube, DNA at 15 mg/ml is almost a gel, so how can anything move inside a cell? And how does all of this DNA keep from getting hopelessly tangled?

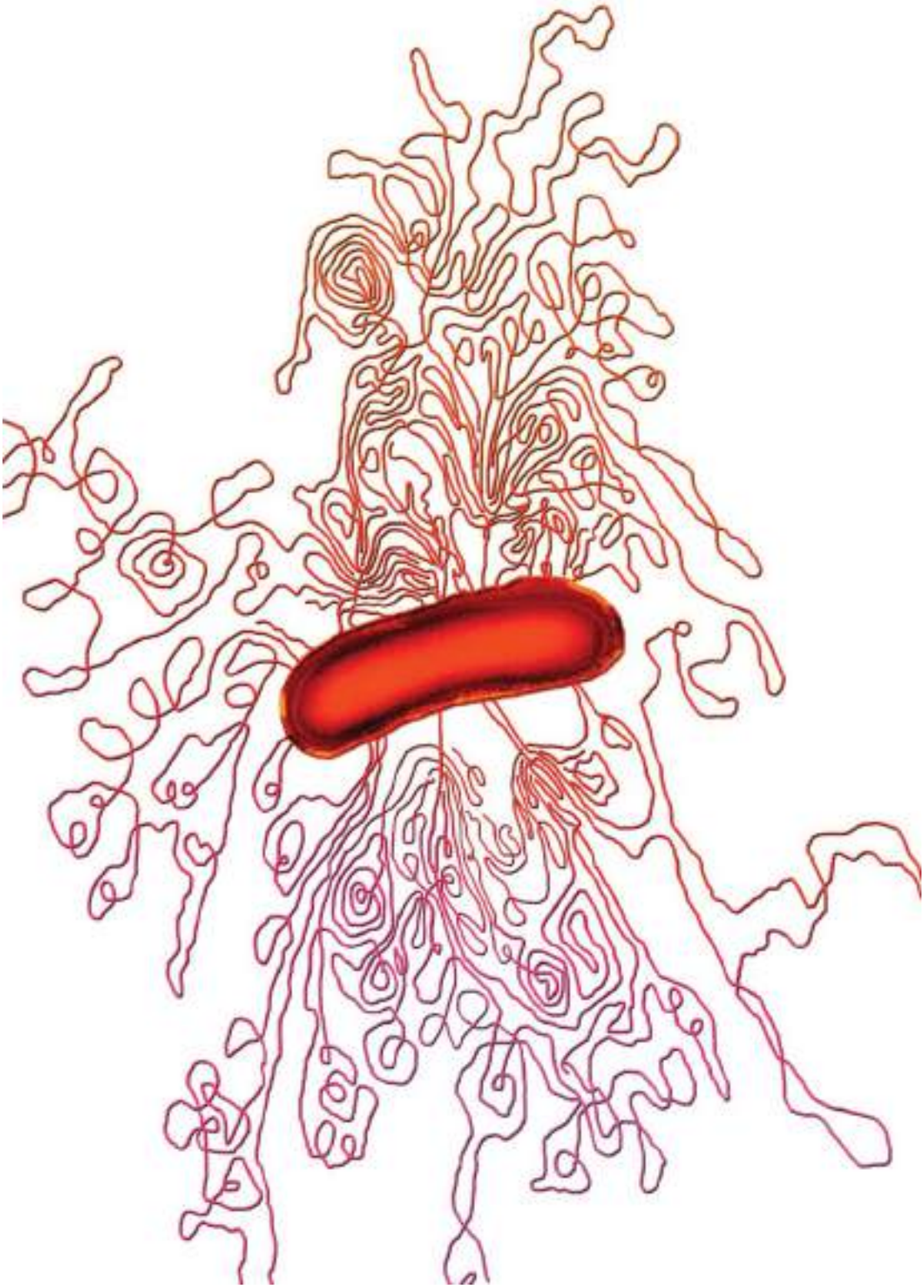


FIGURE 7.6 ■ Osmotically disrupted bacterial cell with its DNA released. In this colorized transmission electron micrograph, the length of the bacterium is approximately 2 μm.

DR. GOPAL MURTI/VISUALS UNLIMITED, INC.

As introduced in Chapter 3 , cells pack their DNA into a manageable form that still allows ready access to DNA-binding proteins. Although bacteria lack a nuclear membrane, they pack their DNA into a series of protein-attached domains collectively called the nucleoid (see Section 3.4).

DNA Supercoiling Compacts the Chromosome

A nucleoid gently released from *E. coli* appears as 30–100 tightly wound loops, or "domains" (**Fig. 7.7**). The boundaries of each loop are defined by anchoring proteins called histone-like proteins for their similarity to histones, the DNA-binding proteins of eukaryotes (discussed in Section 7.5). The DNA double helix within each domain is itself helical, or supercoiled . Supercoiled DNA is quite compact, taking up much less space than a relaxed molecule.

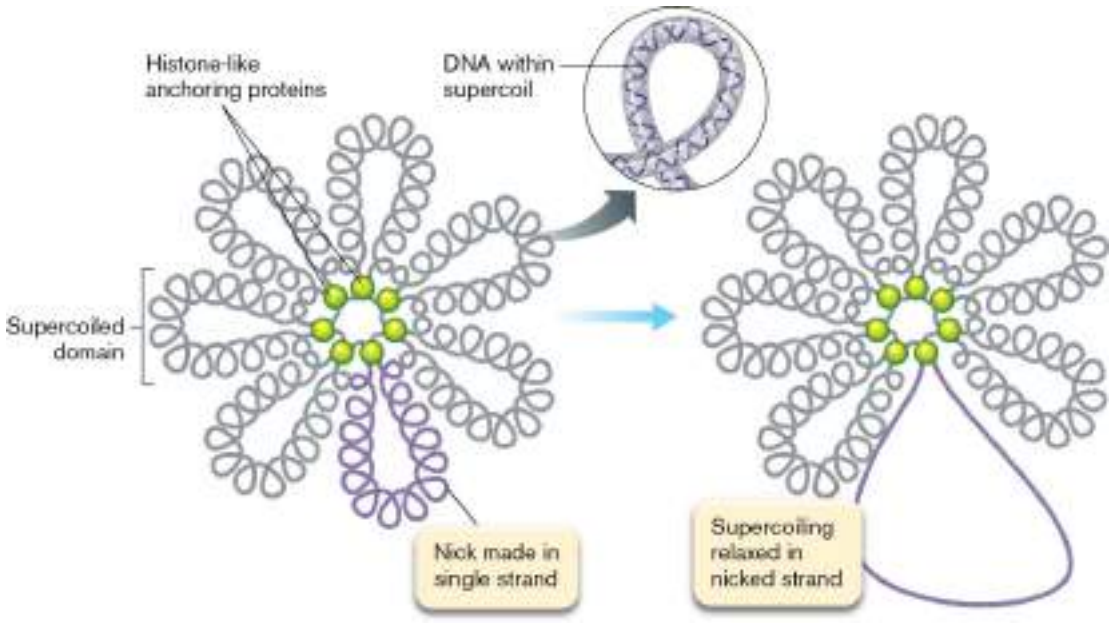


FIGURE 7.7 ■ Bacterial nucleoid. A nucleoid, showing domain loops after gentle release from cells. The single-strand nick unwinds (relaxes) only one loop.

See above for *Supercoiling and Topoisomerases* animation

SPECIAL TOPIC 7 DNA as Digital Storage

Currently, this textbook is available in paper and silicon-based digital formats; in the near future, you may be able to access a version that is stored as DNA. This DNA copy would occupy a space not much larger than the head of a pin, and it could survive in legible condition for thousands of years or longer.

Just as the sequence of the four bases is used to code genes in living organisms, those bases can be exploited as digits for information storage. Each base can theoretically serve as 2 bits of digital information; for example, using the binary code of 0's and 1's: A = 00, C = 01, G = 10, and T = 11. To store data as DNA, the digital code is translated into a code involving the four nucleotides. The nucleotide sequence is then synthesized by machine to become the storage molecule. It can be stored as an isolated molecule or introduced into the genome of a microbial host cell. The DNA code is then read, usually with the polymerase chain reaction (PCR) to amplify the DNA, followed by sequencing of the DNA amplicons (see eAppendix 3 for descriptions of these techniques). Finally, the sequences are decoded back to the digital code.

DNA has been hailed as a possible improvement to silicon-based storage for several key reasons. DNA is a very stable molecule and, under ideal conditions, can be preserved for centuries or longer, as evidenced by retrieval of DNA from ancient prokaryotes and Neanderthals in permafrost. This is a significant improvement over the durability of silicon-based storage, which is limited to decades or less. DNA has greater potential information density as well. A maximum of 455 exabytes (455×10^{18}) per gram has been estimated, which is orders of magnitude higher than silicon, tape, or optical storage

capacities of terabytes to petabytes ($10^{12} - 10^{15}$ bytes). Finally, the decrease in costs for DNA synthesis and sequencing is outpacing the cost reduction for silicon-based digital storage, suggesting that it may soon be more cost-effective to store information on DNA than on silicon.

As a proof-of-concept study, Seth Shipman and George Church (**Fig. ST 7.1**), together with their colleagues, recorded a movie into the DNA of a population of *E. coli* , and then played it back. For the image pixels, the tones within the gray scale were encoded by a triplet nucleotide: The 64 possible triplets generated a redundant set to encode 21 different tones (**Fig. ST 7.2A**). The pixel locations within the image were determined by the position of the triplet along an oligonucleotide (**Fig. ST 7.2B**). Multiple oligonucleotides were needed to encode the entire image, and these different DNA molecules were tagged with unique bar codes (pixets; **Fig. ST 7.2B**). In total, five sets of oligonucleotides were synthesized, each encoding one frame of a five-frame movie of a galloping mare from Eadweard Muybridge's *Human and Animal Locomotion* .



FIGURE ST 7.1 ■ Seth Shipman (left) and George Church.

WYSS INSTITUTE AT HARVARD UNIVERSITY

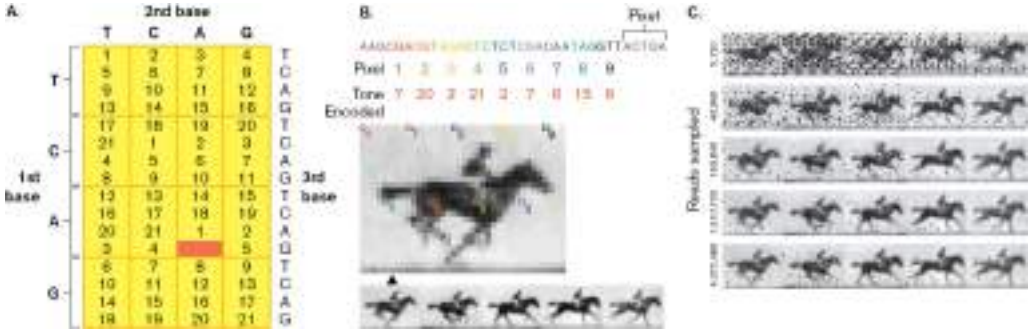


FIGURE ST 7.2 ■ Recording a movie as DNA in the *E. coli* genome. A. The triplet codes used to convert the 21 pixel tones (numbers 1–21) into a DNA sequence. **B.** A representative oligonucleotide, with its own unique identification tag (pixet), and the locations of the nine coded pixels on the encoded image. **C.** Playback of the DNA-encoded movie by increasing amounts of sequencing leads to greater recovery of the five movie frames.

Source: Part A modified from S. L. Shipman et al. 2017. *Nature* **547** :345–349, fig. 1b.

S. L. SHIPMAN ET AL. 2017. *NATURE* **547** :345–349, FIG. 3A.
S. L. SHIPMAN ET AL. 2017. *NATURE* **547** :345–349, FIG. 3E.

The CRISPR-Cas system (see Chapters 6 and 12) was used to incorporate the oligonucleotides into the chromosome of *Escherichia coli* . Because new oligonucleotides are almost always added adjacent to the leader sequence of the CRISPR array, the researchers were able to load the frames sequentially into the *E. coli* genome. Thus, just as a camera records and stores images over time, they were able to "record" and store the images of a movie in DNA. To play back the video, the researchers used PCR to amplify the oligonucleotides within the population of *E. coli* cells and high-throughput sequencing to

obtain the nucleotide sequence, which was then back-translated into pixel information (**Fig. ST 7.2A** and **B**) to obtain the images (**Fig. ST 7.2C**). Increasing the number of sequence reads led to more accurate recall of the images, though perfect recall was not achieved in this initial study.

For DNA to become a viable option for data storage, several practical limitations need to be dealt with. As indicated in the study just described, perfect data recall is a major challenge because of technical aspects of DNA sequencing and the potential for DNA mutation when encoded in a live organism. Recently, other groups have shown that adaptation of error-correcting codes, such as the one used to prevent video dropouts during streaming, could provide error-free data recovery from DNA. The long recording time (weeks) and retrieval time (days) of current technologies limits the application of DNA as a rapid-retrieval storage system for daily use.

The high capacity and durability, but relatively long access time, currently make DNA storage most applicable to long-term archiving of information. This capability is highly relevant to the medical field, where secure, long-term storage of patient records is critical. There is even consideration of DNA as a means of safeguarding human knowledge via deep, ultracold storage in outer space, away from any future catastrophes on Earth.

RESEARCH QUESTION

The first genome to be fully sequenced was that of the *E. coli* –infecting virus phiX174. Shortly after the genome sequence was published in 1977, some scientists searched for possible messages from extraterrestrial beings embedded within the genetic code. This effort, while ultimately futile, inspired the concept that DNA could be used to communicate to alien life if placed on our space probes the way the Golden Records were on *Voyager 1* and *2* . What are the possible limitations of such

forms of biochemical communication, given the potentially different chemistries and biological origins on different planets?

Shipman, Seth L., J. Nivala, J. D. Macklis, and George M. Church. 2017. CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria. *Nature* 547:345–349.

Note that to form supercoils, DNA must have its ends tethered. In a circular chromosome, the DNA ends are tethered to each other. Introducing an extra twist by breaking one or both strands, twisting one end, and then resealing the strands means that the increased, or decreased, torsional (twisting) stress is trapped in the final circular molecule. Supercoiled DNA cannot spontaneously unwind but, as we will discuss, can be unwound by special enzymes during replication.

Remarkably, the nucleoid can maintain different superhelical densities for its 30–100 domains. The independence of supercoiled domains was demonstrated by experimental introduction of a single-strand nick in the phosphodiester backbone of one domain (**Fig. 7.7**). Adding very small amounts of a particular nuclease (an enzyme that cleaves a nucleic acid) can ensure that only a single strand within a single domain is nicked. The ends of the nicked strand, driven by the energy inherent in the supercoil, rotate about the unbroken complementary strand of the duplex and relax the supercoil of the affected domain. However, the other domains in the nucleoid remain supercoiled. How is this possible if the chromosome is one circular molecule? The unaffected chromosomal domains remain supercoiled because they are constrained by the anchoring proteins from which these domains extend (for nucleoid organization, see **Fig. 3.27**).

How do cells supercoil their chromosome? The bacterial cell produces enzymes that can twist DNA into supercoils and other enzymes that relieve supercoils. A single twist introduced into a small (300-bp) circular DNA molecule forms a single supercoil as shown in **Figure 7.8** . To introduce the DNA twist, a supercoiling enzyme

makes a double-strand break at one point in the circle, passes another part of the DNA through the break, and reseals it. The result is the same as if one end of the broken circle were twisted one full turn.

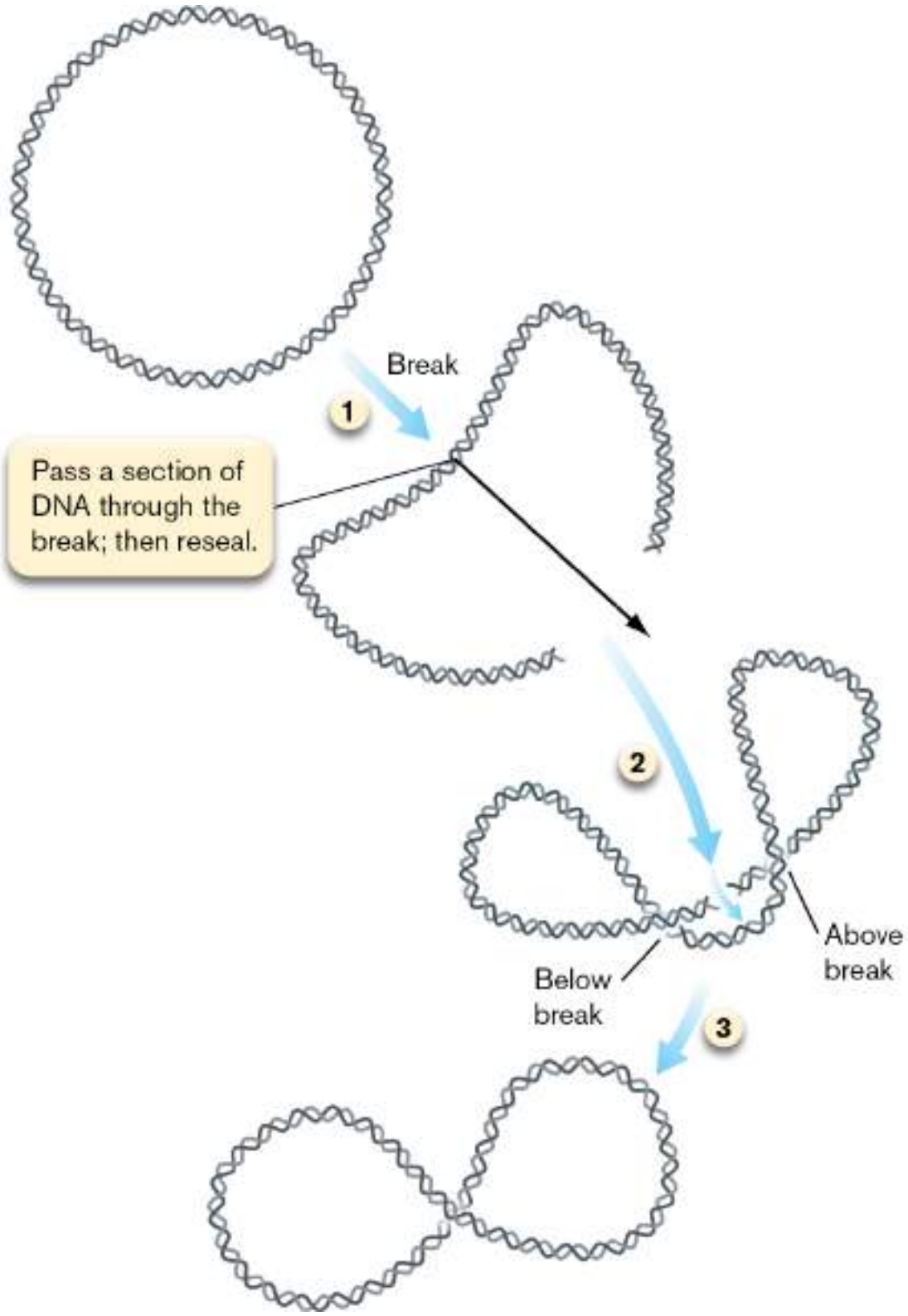


FIGURE 7.8 ■ Supercoiling of 300-bp circular DNA. To introduce supercoils into a double-stranded, circular DNA

molecule, both strands are cleaved at one site in the molecule (step 1), an intact part of the molecule is passed between ends of the cut site (step 2), and the free ends are reconnected (step 3).

See above for *Supercoiling and Topoisomerases* animation

The nucleoids of bacteria and most archaea, as well as the nuclear DNA of eukaryotes, are kept negatively supercoiled. This is established by underwinding the helix; that is, reducing the number of helical twists relative to a relaxed state. To understand how underwinding generates negative supercoils, refer to the model of DNA in **Figure 7.4** . The helix is said to be in a right-handed conformation, which means the helix turns clockwise when we look down the length of the double strand. When DNA is underwound, the number of clockwise twists is lower, causing torsional stress that tends to open the helix. This stress is relieved by the formation of compensatory clockwise supercoils. **Figure 7.9** illustrates negative, clockwise supercoils superimposed on a clockwise helix.

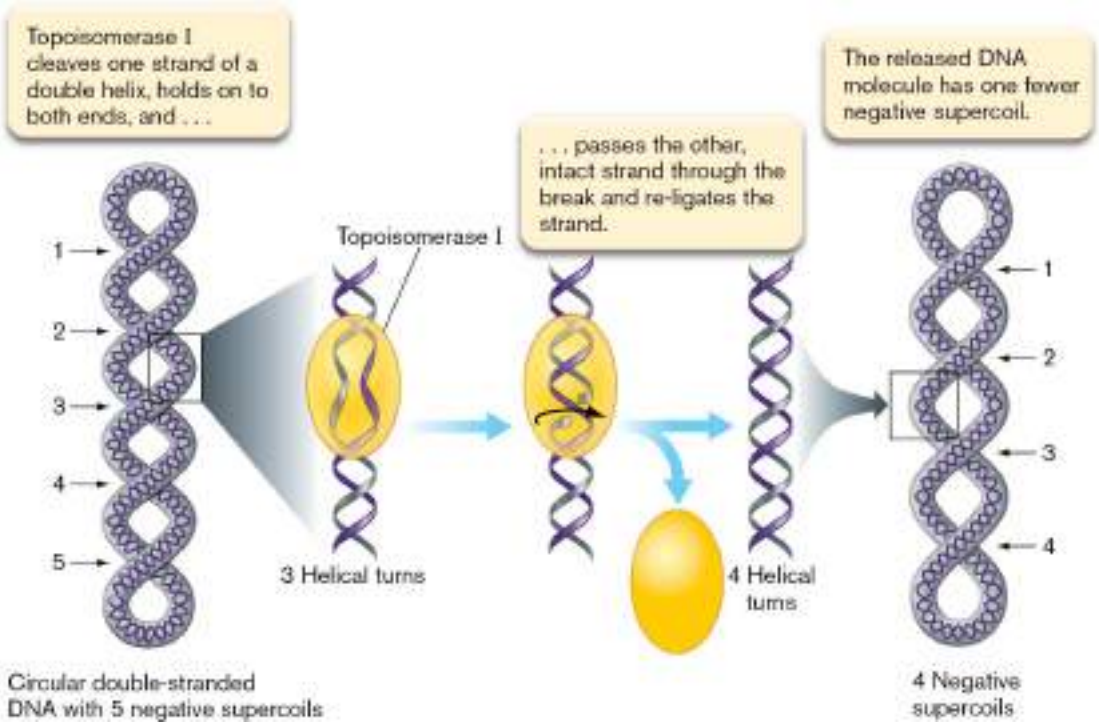


FIGURE 7.9 ■ Mechanism of action for type I topoisomerases (Topo I of *E. coli*). Topoisomerase I relaxes a negatively supercoiled DNA molecule by introducing a single-strand nick.

See above for *Supercoiling and Topoisomerases* animation

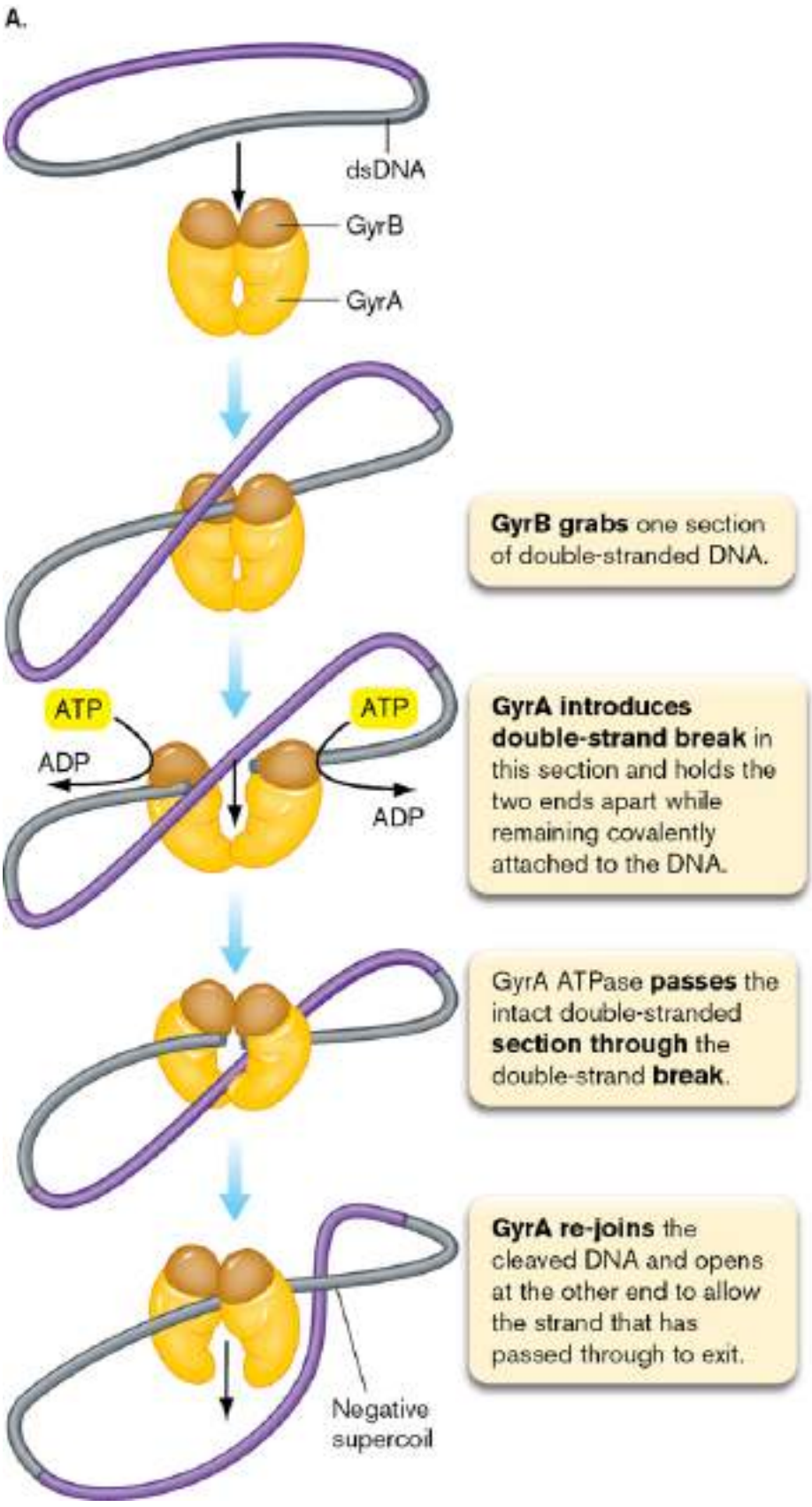
Because the DNA is underwound, the two strands of negatively supercoiled DNA are easier to separate than those of positively supercoiled DNA. This is important for transcription enzymes, such as RNA polymerase, that must separate strands of DNA to make RNA.

Note, however, that some archaeal species living in acid at high temperature have nucleoids that are positively supercoiled to keep DNA double-stranded in these inhospitable environments (discussed next). Positively supercoiled DNA is harder to denature, because it takes excess energy to separate overwound DNA.

Topoisomerases Supercoil DNA

Supercoiling changes the topology of DNA. Topology is a description of how spatial features of an object are connected to each other. Thus, enzymes that change DNA supercoiling are called topoisomerases . To maintain proper DNA supercoiling levels, a cell must delicately balance the activities of two types of topoisomerases. Type I topoisomerases are typically single proteins that cleave only one strand of a double helix, while type II topoisomerases have multiple subunits that cleave both strands of a DNA molecule. Type I enzymes relieve or unwind supercoils. As shown in **Figure 7.9** , topoisomerase I cleaves one strand of a negatively supercoiled double helix and holds on to both ends of the break. The release of the energy stored in the negative supercoil allows the enzyme to pass the intact strand through the break and re-ligate (reconnect) the strand, thereby reintroducing a helical turn. The molecule is released with one fewer negative supercoil.

Type II topoisomerases such as DNA gyrase introduce negative supercoils in DNA (**Fig. 7.10**), which is important during DNA replication (see Section 7.3). Adding a supercoil requires spending energy, by hydrolysis of ATP. Gyrase is a tetrameric complex composed of two gyrase A (GyrA) and two gyrase B (GyrB) proteins. The GyrB subunits first grab a section of the double helix. Then GyrA, in an ATP-dependent process catalyzed by GyrB, introduces a double-strand break, passes a different part of the double helix through the break, and reseals the break. The other end of GyrA then opens to release the DNA, now with one more negative supercoil. The 3D representation in **Figure 7.10B** shows DNA gyrase in the midst of generating a supercoil.


A.
dsDNA
GyrB
GyrA
GyrB grabs one section of double-stranded DNA.
ATP
ATP
ADP
ADP
GyrA introduces double-strand break in this section and holds the two ends apart while remaining covalently attached to the DNA.
GyrA ATPase passes the intact double-stranded section through the double-strand break.
GyrA re-joins the cleaved DNA and opens at the other end to allow the strand that has passed through to exit.
Negative supercoil

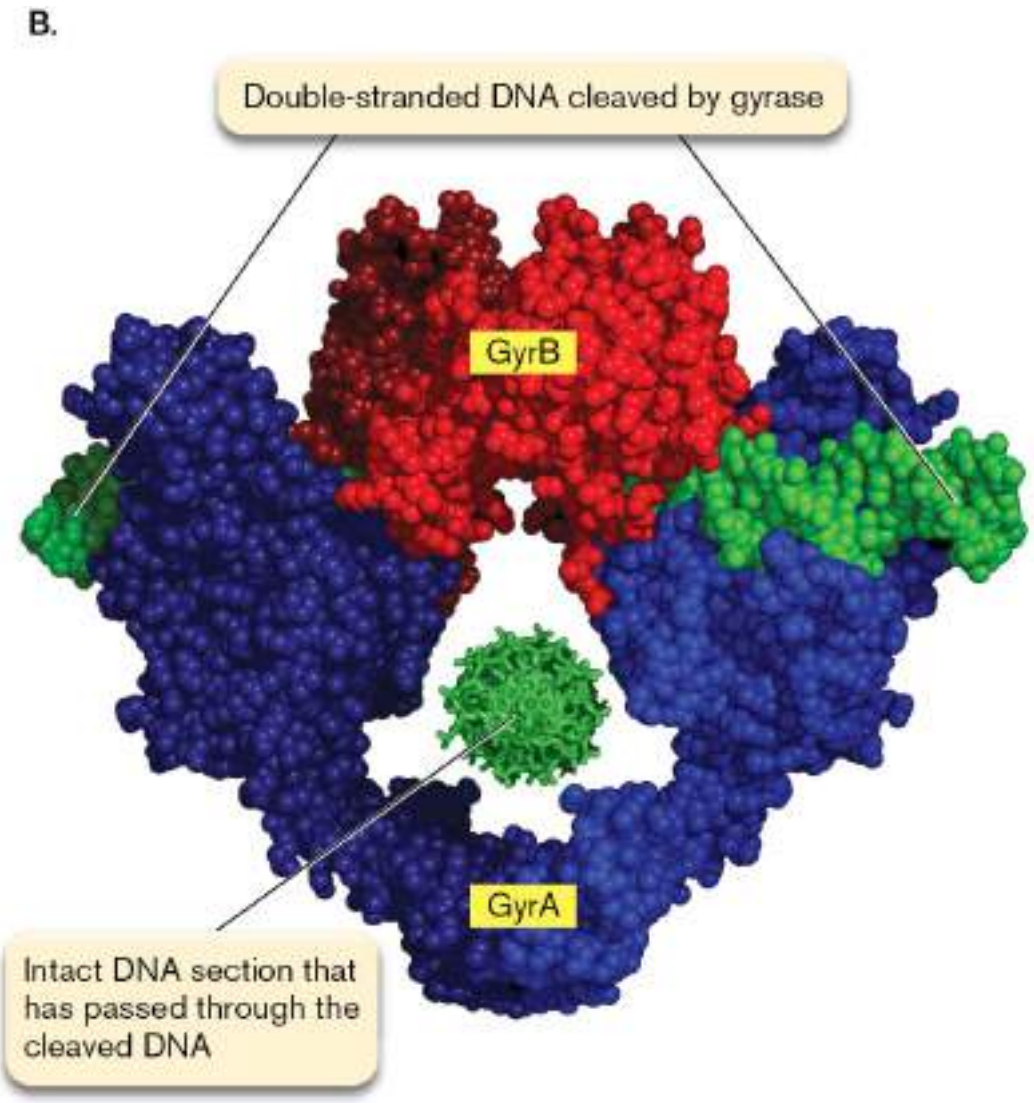


FIGURE 7.10 ■ Mechanism of action for type II topoisomerases. A. Mode of action of DNA gyrase from *E. coli* . **B.** Three-dimensional representation of DNA gyrase. The gyrase complex grips a broken DNA duplex (shown in green) and has transported a second duplex (the multicolored rosette) through the break.

COURTESY OF JAMES BERGER

See above for *Supercoiling and Topoisomerases* animation

Note: The degree of supercoiling depends on the linking number, which is the number of times one strand goes around the other. In relaxed DNA, the linking number is zero. Type I topoisomerases, which nick one DNA strand, change the linking number by increments of one. In contrast, type II topoisomerases such as DNA gyrase, which cleave both strands, change the linking number by two. For simplicity, the reaction of DNA gyrase is shown in Figure 7.10A to generate a single negative supercoil.

Enzymes that make or manage bacterial DNA are common targets for antibiotics. For instance, the quinolone antibiotics specifically target bacterial type II topoisomerases. These antibiotics do not affect eukaryotic topoisomerases. A modern quinolone, ciprofloxacin, was the treatment of choice for anthrax pneumonia during the 2001 domestic terrorism attacks that used letters containing anthrax spores to kill 5 people and infect 17 others in the United States. The quinolones nalidixic acid and oxolinic acid inhibit DNA gyrase. Long before the genome of *E. coli* was sequenced, the *gyrA* gene encoding the GyrA component was identified through genetic analysis of mutants with newfound resistance to nalidixic acid and oxolinic acid. The modern successors of these drugs, the fluoroquinolones, are among the most widely used antimicrobials in the world. These drugs stabilize the complex in which DNA gyrase is covalently attached to DNA (see **Fig. 7.10**). The stuck complex forms a physical barrier in front of the DNA replication complex, and the bacterial cell dies.

Extreme thermophiles (hyperthermophilic archaea) possess an unusual gyrase called reverse DNA gyrase. In contrast to the DNA gyrase from mesophiles, reverse gyrase introduces positive supercoils into the chromosome. It is proposed that tightening the coil helps protect the chromosome against thermal denaturation. Because the DNA has extra twists, it takes more energy (heat) to separate the strands.

Thought Questions

7.4 DNA gyrase is essential to cell viability. Why, then, are nalidixic acid–resistant cells that contain mutations in *gyrA* still viable?

7.5 Bacterial cells contain many enzymes that can degrade linear DNA. How, then, do linear chromosomes in organisms like *Borrelia burgdorferi* (the causative agent in Lyme disease) avoid degradation?

To Summarize

- **A genome** is all of the genetic information that defines an organism.
- **Genomes** of bacteria and archaea are made up of chromosomes and plasmids consisting of DNA.
- **DNA is composed of two antiparallel chains** of purine and pyrimidine nucleotides in which phosphate links the 5′ carbon of one nucleotide with the 3′ carbon of its neighbor. The result is a double helix containing a deep major groove and a more shallow minor groove.
- **Hydrogen bonding and interactions between the stacked bases** hold together complementary strands of DNA.
- **Supercoiling** by topoisomerases compacts DNA into an organized nucleoid.
- Bacteria, eukaryotes, and most archaea possess **negatively supercoiled DNA** . Archaea living in extreme environments have **positively supercoiled genomes** .
- **Type I topoisomerases** cleave one strand of a DNA molecule and relieve negative supercoiling; **type II topoisomerases** such as DNA gyrase cleave both strands of DNA and use ATP to introduce negative supercoils.

Glossary

genome
: The complete genetic content of an organism. The sequence of all the nucleotides in a haploid set of chromosomes.

nucleobase
: Also called *nitrogenous base* . A planar, heteroaromatic, nitrogen-containing base that forms a nucleotide of nucleic acids; nucleobases determine the information content of DNA and RNA. There are five nucleobases: adenine, cytosine, guanine, thymine, and uracil.

nitrogenous base
: See **nucleobase** .

phosphodiester link
: The linkage between two adjacent nucleotides in a nucleic acid. A phosphate forms ester bonds with the 5′ carbon of one (deoxy-)ribose and the 3′ carbon of the other (deoxy-)ribose.

antiparallel
: Oriented such that the two strands are in opposite directions. Commonly refers to a nucleic acid double helix with one strand in the 5′-to-3′ orientation and the other strand in the 3′-to-5′ orientation.

purine
: A nitrogenous base with fused rings (that is, a bicyclic nucleobase) found in nucleotides; examples are adenine and guanine.

pyrimidine
: A single-ring nitrogenous base (that is, a monocyclic nucleobase) found in nucleotides; examples are cytosine, thymine, and uracil.

denature
: To lose secondary and tertiary structure in a protein or nucleic acid because of high temperature or chemical treatment.

hybridization
: The annealing of a nucleic acid strand with another nucleic acid strand containing a complementary sequence of bases. The

binding of one nucleic acid strand with a complementary strand.

nucleoid

The looped coils of a bacterial chromosome.

supercoiled DNA

A physical state of circular DNA where the number of twists of the two strands around each other is either less than or greater than the number found in the relaxed state of the double-stranded helix. The resulting tension generates a higher-order helix that compacts the DNA double-stranded helix.

topoisomerase

An enzyme that can change the supercoiling of DNA.

quinolone

A type of antibiotic drug that inhibits DNA synthesis by targeting bacterial topoisomerases such as DNA gyrase.

Fig. 3.27

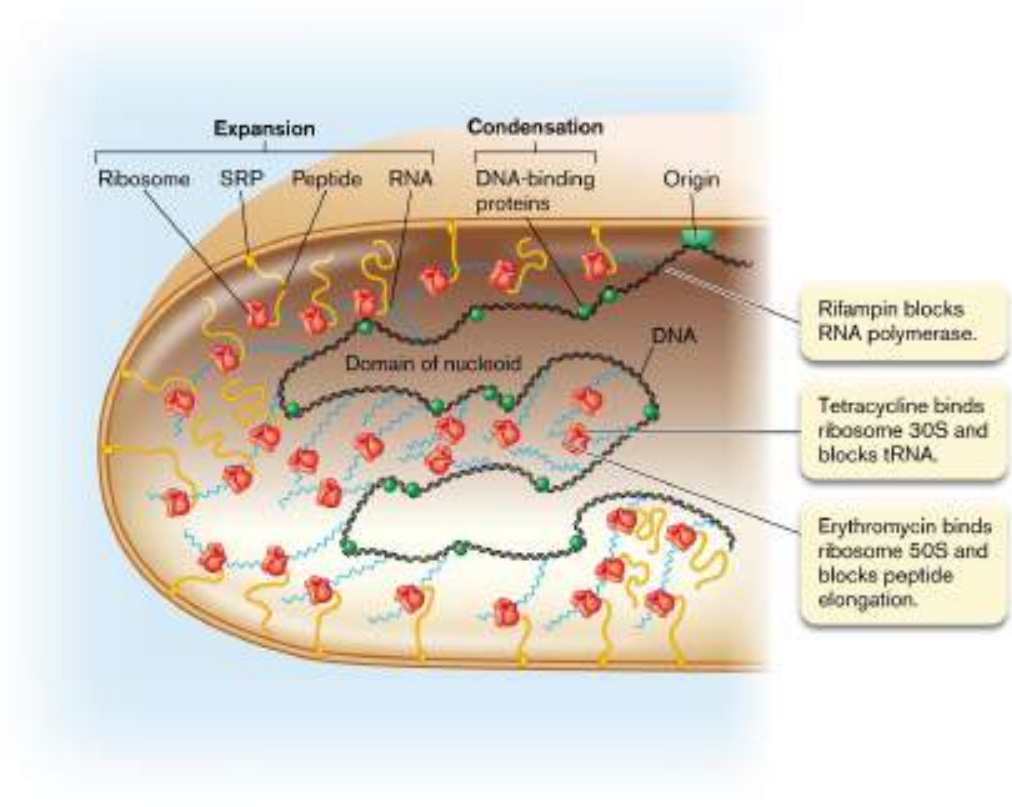


FIGURE 3.27 ■ DNA transcription and RNA translation to peptides. The nucleoid forms chromosome loops called domains, which loop out from the origin of

attachment to the cell envelope. Bacterial transcription of DNA to RNA is coordinated with translation of RNA to make proteins. Growing peptide chains destined for the membrane bind the signal recognition particle (SRP) for membrane insertion.

7.3 DNA Replication

Quick and accurate replication of DNA can help a microorganism grow rapidly and compete with other species. Replication efficiency is one reason why bacterial pathogens such as *Salmonella* can cause disease so quickly after ingestion. In this respect, unicellular bacteria differ from multicellular organisms, which need to regulate cell division carefully within their tissues because unregulated growth within tissues leads to cancer. The process of bacterial replication involves over 20 proteins coming together in a complex machine. Operation of the replication complex is all the more remarkable, considering that some bacteria, such as thermophilic *Bacillus* species that live in hot springs, can double their population in less than 15 minutes.

The molecular details of bacterial DNA replication are important for us to understand because they provide targets for new antibiotics and tools for biotechnology, such as the polymerase chain reaction (PCR; see eAppendix 3). In addition, the bacterial proteins of DNA repair have homologs in the human genome, defects in which produce inherited human diseases such as xeroderma pigmentosum that predispose the carrier to certain cancers.

Overview of Bacterial DNA Replication

To replicate a molecule containing millions of base pairs poses formidable challenges. How does replication begin and end? How is accuracy checked and maintained without causing a major drop in replication rate that could slow reproduction?

Semiconservative replication. Replication of cellular DNA is semiconservative , meaning that each daughter cell receives one parental strand and one newly synthesized strand (**Fig. 7.11**). At the replication fork , the advancing DNA synthesis machine separates the parental strands while extending the new, growing strands. The

semiconservative mechanism provides a means for each daughter duplex to be checked for accuracy against its parental strand.

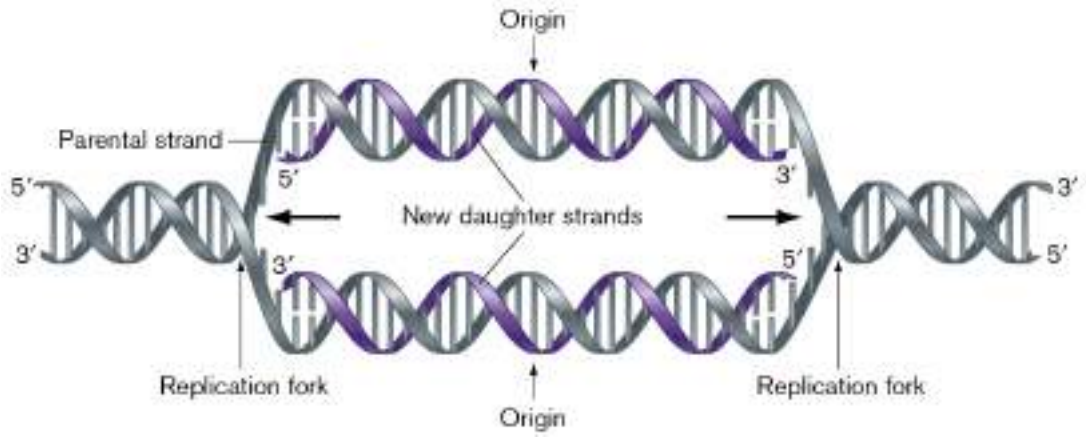
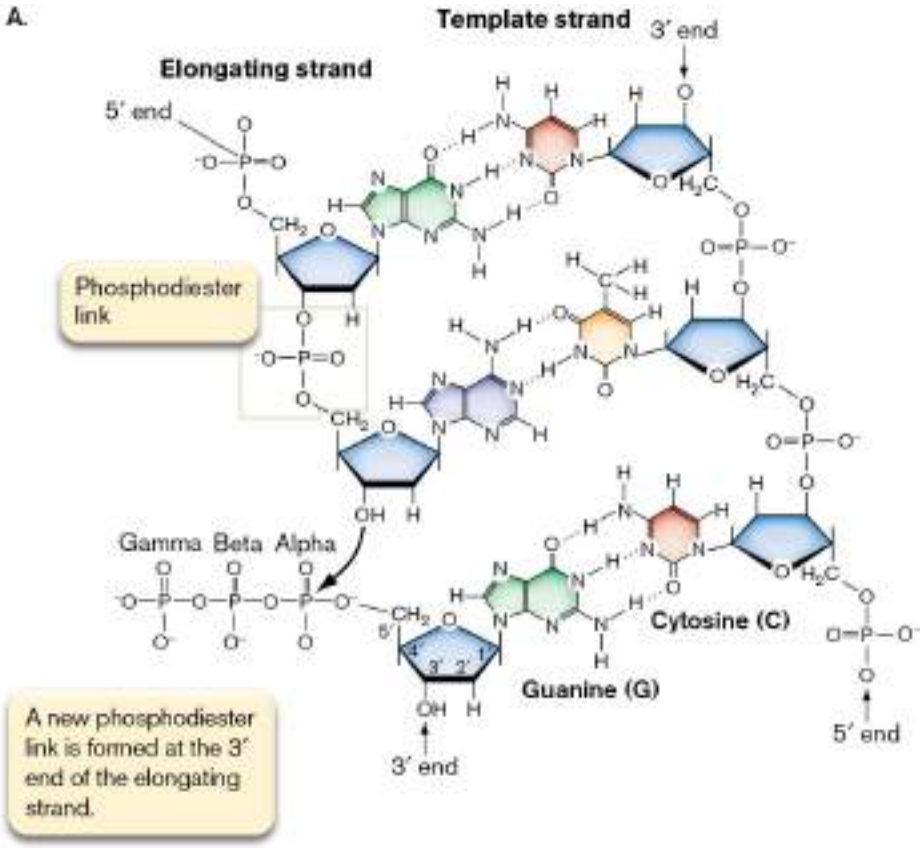


FIGURE 7.11 ■ Semiconservative replication. A replication bubble with two replication forks. DNA replication is termed "semiconservative" because each of the resulting double-stranded chromosomes contains one of the original parental strands and one newly synthesized strand. It is called "bidirectional" because it begins at a fixed origin and progresses in opposite directions.

Enzymes that synthesize DNA or RNA can connect nucleotides only in a 5′-to-3′ direction (**Fig. 7.12**). That is, the nucleic acid polymer elongates only at the 3′ end. A polymerase (a chain-lengthening enzyme complex) forms a phosphodiester link between the 3′ end of the growing chain and the alpha-phosphate located at the 5′ end of an incoming nucleoside triphosphate. (The alpha-phosphate is the phosphate closest to the sugar.) This reaction releases the beta-phosphate and gamma-phosphate of the incoming nucleoside as a diphosphate molecule called pyrophosphate. Pyrophosphate is subsequently cleaved by pyrophosphatase; removal of this product of the polymerization reaction prevents that reaction from working in the reverse direction. Polymerization thus results in the eventual breaking of both phosphoryl bonds within the triphosphate moiety of the incoming nucleoside. Given the high cost for the formation of these phosphoryl bonds, DNA chain elongation is expensive for the cell, but essential.


A.
Template strand
3′ end
Elongating strand
5′ end
Phosphodiester link
Gamma Beta Alpha
Cytosine (C)
Guanine (G)
5′ end
3′ end
A new phosphodiester link is formed at the 3′ end of the elongating strand.

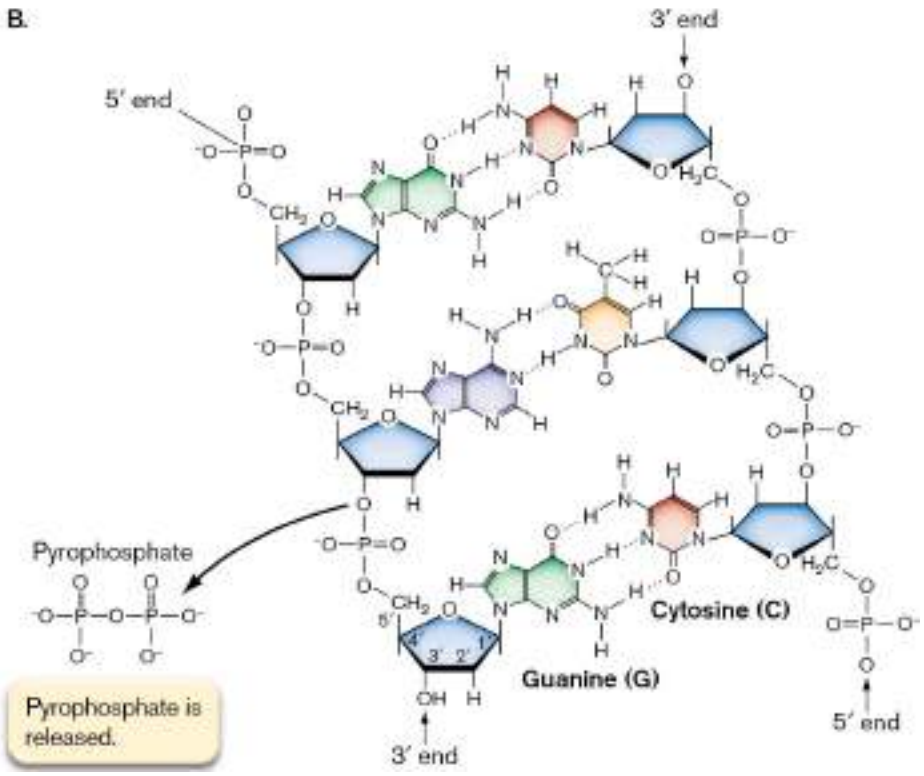

B.
3′ end
5′ end
Pyrophosphate
Cytosine (C)
Guanine (G)
5′ end
3′ end
Pyrophosphate is released.

FIGURE 7.12 ■ DNA synthesis by chain elongation. A. A new phosphodiester link is formed between the 3′ OH of the elongating strand and the alpha-phosphate of the incoming nucleoside. **B.** The incorporated nucleotide becomes the new 3′ end of the elongating chain and forms hydrogen bonds with the complementary nucleotide on the template strand. Pyrophosphate composed of the beta-phosphate and gamma-phosphate is released during this reaction.

The 5′-to-3′ enzymatic constraint of polymerization produces an interesting mechanistic puzzle: If polymerases can synthesize DNA only in a 5′-to-3′ direction and the two phosphodiester backbones of the double helix are antiparallel, then how are both strands of a moving replication fork synthesized concurrently? One strand presents no problem, because it is synthesized in a 5′-to-3′ direction toward the fork, but synthesizing the other, new strand in a 5′-to-3′ direction would seem to dictate that it move away from the fork (**Fig. 7.11**). How then is the cell able to copy both parental strands during semiconservative replication?

Note: A nucleo s ide (nucleobase plus ribose or deoxyribose) also condensed with phosphoryl group(s) is referred to as a nucleo t ide.

DNA replication proceeds in three phases: (1) initiation, which is the melting (unwinding) of the helix and the loading of the DNA polymerase enzyme complex; (2) elongation, which is the sequential addition of deoxyribonucleotides to a growing DNA chain, followed by proofreading; and finally (3) termination, in which the DNA duplex is completely duplicated, the negative supercoils are restored, and key sequences of new DNA are methylated. The basic process of chromosome replication is outlined in **Figure 7.13** .

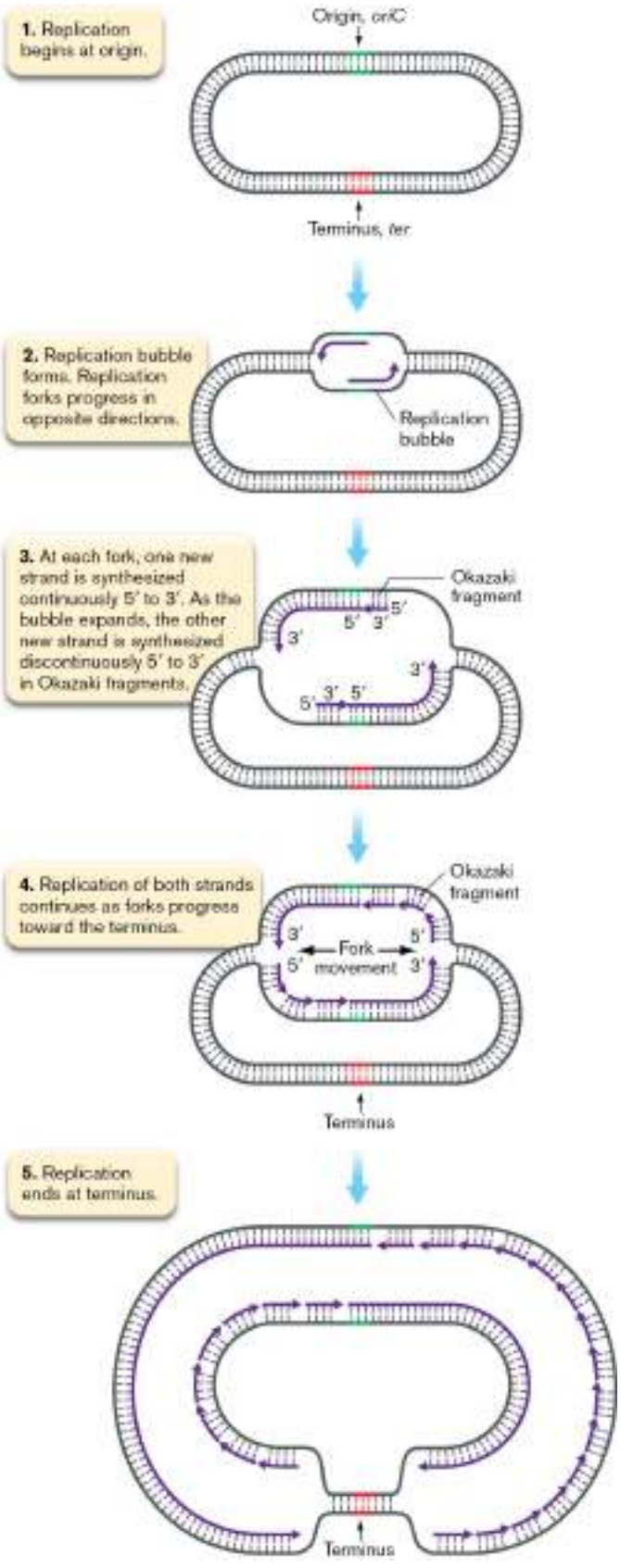

1. Replication begins at origin.
Origin, oriC
Terminus, ter
2. Replication bubble forms. Replication forks progress in opposite directions.
Replication bubble
3. At each fork, one new strand is synthesized continuously 5′ to 3′. As the bubble expands, the other new strand is synthesized discontinuously 5′ to 3′ in Okazaki fragments.
Okazaki fragment
4. Replication of both strands continues as forks progress toward the terminus.
Okazaki fragment
Fork movement
Terminus
5. Replication ends at terminus.
Terminus

FIGURE 7.13 ■ Comparing direction of fork movement with direction of DNA synthesis.

Replication from a single origin. Replication in bacteria begins at a single defined DNA sequence called the origin (***oriC ;*** **Fig. 7.13** , step 1). Following initiation, a circular bacterial chromosome replicates bidirectionally (in both directions away from the origin) until it terminates at defined termination (*ter*) sites located on the opposite side of the molecule.

What happens to new replication origins after replication begins? Recall from Chapter 3 that newly replicated origins move away from midcell toward opposite cell poles (see **Figure 3.28**). This movement is part of the partitioning mechanism that places chromosomes out of harm's way before the division septum forms at midcell.

Fundamentals of DNA replication. After initiation of replication, a replication bubble forms at the origin. The bubble contains two replication forks that move in opposite directions around the chromosome (**Fig. 7.13** , step 2). DNA polymerases synthesize DNA in a 5′-to-3′ direction. At each fork, therefore, one new DNA strand can extend continuously until the terminus region (step 3). However, because the two DNA strands are antiparallel and the DNA polymerases synthesize only in the 5′-to-3′ direction, the other daughter strand has to be synthesized discontinuously, in stages—seemingly backward relative to the moving fork (step 4). The fragments of DNA formed on this discontinuously synthesized strand are called Okazaki fragments , after the Japanese scientists Reiji and Tsuneko Okazaki, a married couple, who discovered them. As we will discuss later, the Okazaki fragments are progressively stitched together to make a continuous, unbroken strand. Ultimately, the two replicating forks meet at the terminal sequence (**Fig. 7.13** , step 5), and the two daughter chromosomes separate.

Now let's examine each step in molecular detail, to answer some important questions about this mechanism.

Initiating Replication

Initiation is controlled by the binding of a specific initiator protein to the origin sequence. Subsequent molecular events load the DNA polymerase complex and generate the first RNA primer for the new DNA strand. Once the replication process has begun, the cell is committed to completing a full round of DNA synthesis. As a result, the decision of when to start copying the genome is critical. If that process starts too soon, the cell accumulates unneeded chromosome copies; if it starts too late, the dividing cell's septum severs the chromosome, killing both daughter cells. Consequently, elaborate fail-safe mechanisms link the initiation of DNA replication with cell mass, generation time, and cellular health, making the timing of initiation remarkably precise.

The replication initiator protein, DnaA. Timing of initiation is determined by the concentration of the replication initiator protein DnaA complexed with ATP (DnaA-ATP) (**Fig. 7.14**).

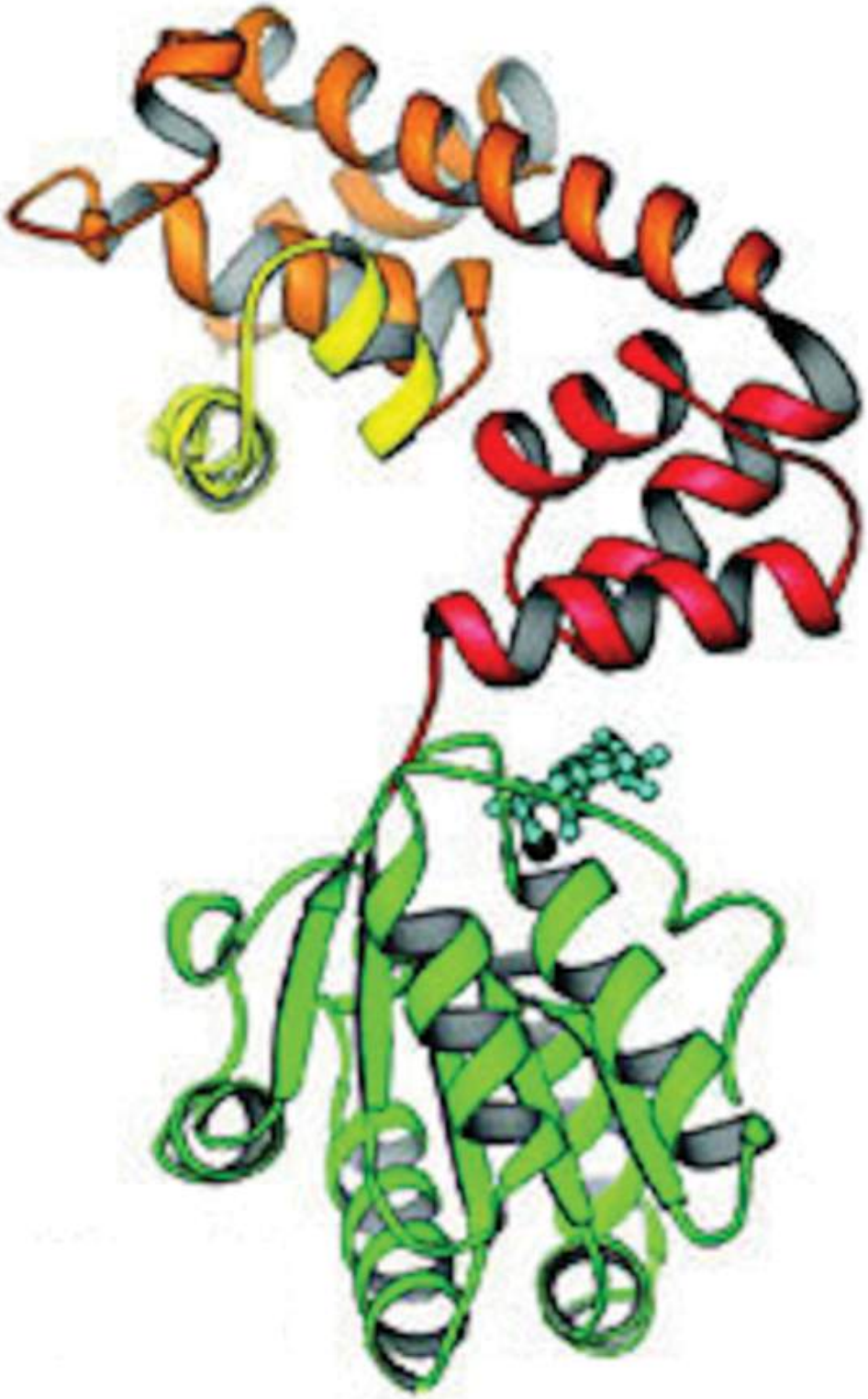


FIGURE 7.14 ■ DnaA monomer of *Aquifex aeolicus*. The helix-turn-helix DNA binding motif is in yellow. The ATP-binding domain (green) is bound to an ATP molecule (light blue).

J. P. ERZBERGER ET AL. 2002. *EMBO J*. **21** :4763–4773

In *E. coli,* DnaA-ATP recognizes specific 9-bp repeats within the 245-bp origin of replication (*oriC)* . As the cell grows, the level of active DnaA-ATP rises until it is sufficient to bind to these repeats (**Fig. 7.15** , step 1). Binding of DnaA to the origin facilitates melting of DNA and initiates the assembly of a membrane-attached replication machine called the replisome , a complex of numerous proteins that come together and bind at *oriC* .

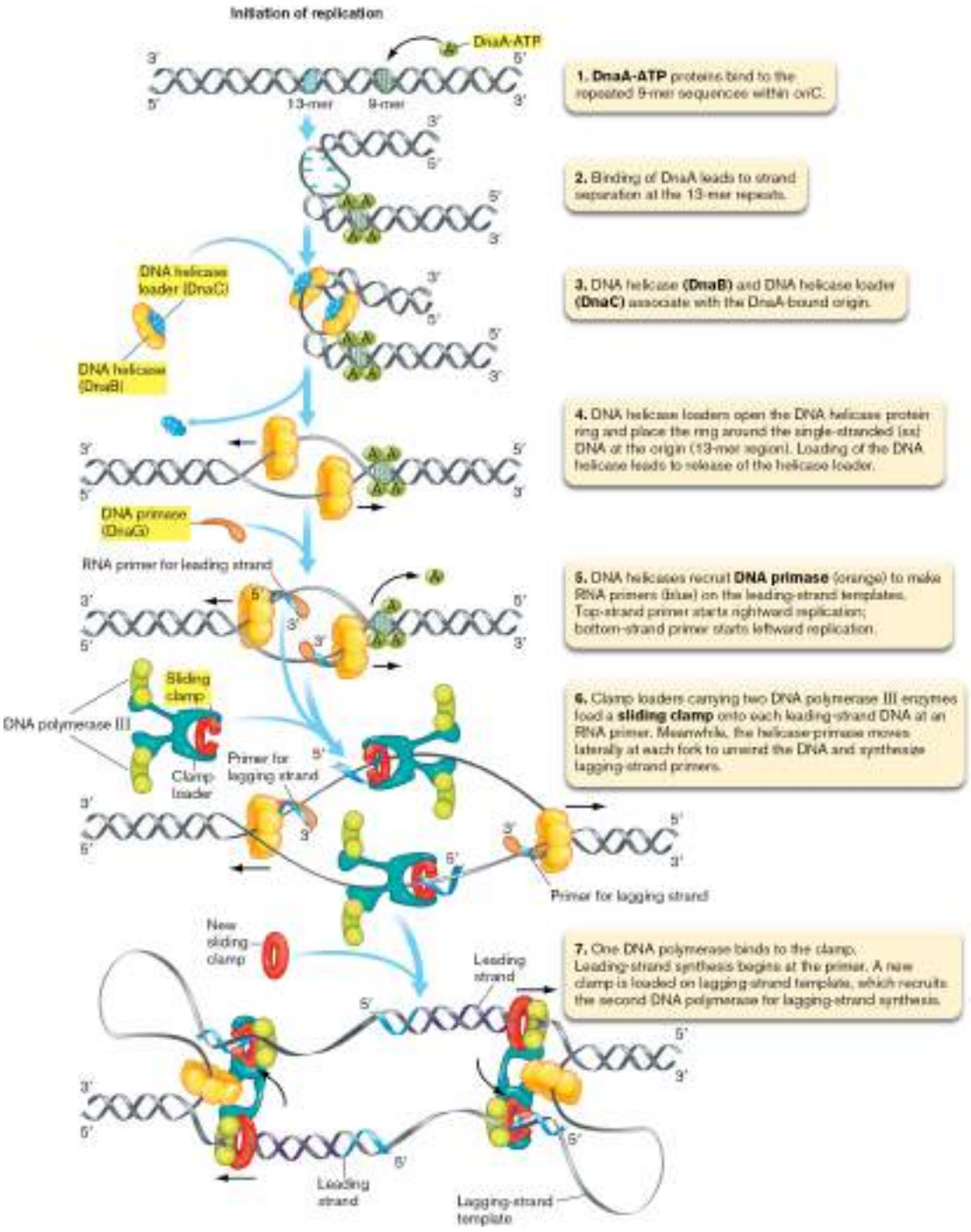


FIGURE 7.15 ■ Initiation of DNA replication.

After it is replicated, the origin cannot immediately trigger another round of replication for two reasons: decreasing levels of unbound DnaA-ATP and inhibition by binding of a second protein, SeqA, to the

origin. Another round of replication can begin only after SeqA dissociates and the DnaA-ATP concentration rises.

DNA methylation controls timing. How does SeqA bind just after the origin has replicated? The key is DNA methylation. *E. coli* uses the enzyme DNA adenine methyltransferase (Dam) to attach a methyl group to the N-6 position of adenine in the sequence GATC (in Fig. 7.3A , see the N of the NH_2 group attached to the six-membered ring of adenine). GATC sequences are scattered along the chromosome on both strands. Just after the origin has replicated, there is a short lag before the newly synthesized strand is methylated by Dam. As a result, the origin is temporarily hemimethylated—a situation in which only one of the two complementary strands is methylated. Because SeqA has a high affinity for hemimethylated origins, this inhibitor will bind most tightly immediately after the origin has replicated. Thus bound, SeqA will prevent another initiation event. Eventually, Dam will methylate the new strand and decrease SeqA binding.

Initiation requires RNA polymerases. An unexpected feature of DNA replication is that its initiation actually requires two RNA polymerases. The first is the housekeeping RNA polymerase used to make most of the RNA in the cell (discussed in Chapter 8). The second is the DNA primase (discussed shortly).

The housekeeping RNA polymerase transcribes DNA at *oriC* , which helps separate the two DNA strands (**Fig. 7.15** , step 2; this RNA polymerase is not shown). Strand separation at *oriC* allows a special DNA helicase (DnaB), in association with a DNA helicase loader (DnaC), to bind the two replication forks formed during initiation (step 3). DnaC facilitates proper placement of DNA helicase at the fork, and then it disengages and leaves.

DnaB uses energy from ATP hydrolysis to unwind the DNA helix before DNA moves into the DNA polymerase replicating complex. The ringlike DnaB is assembled around one DNA strand at each replication fork. After loading DnaB at the origin, DnaC is released (**Fig. 7.15** , step 4). As the DNA unwinds, small single-stranded DNA-binding proteins (SSBs; seen in **Fig. 7.17**) coat the exposed single-stranded DNA, protecting it from nucleases patrolling the cell and preventing

re-formation of double-stranded DNA. The origin is almost ready to receive DNA polymerase.

DNA-dependent DNA polymerases possess the unique ability to "read" the nucleotide sequence of a DNA template and synthesize a complementary DNA strand. The discovery of this activity earned Arthur Kornberg (1918–2007; **Fig. 7.16**) the 1959 Nobel Prize in Physiology or Medicine. However, as remarkable as these enzymes are, no DNA polymerase can start synthesizing DNA unless there is a preexisting DNA or RNA fragment to extend—that is, a primer fragment. The primer fragment possesses a 3′ OH end that receives incoming deoxyribonucleotides. Consequently, once the DnaB is bound to DNA, the next step is to make RNA primers at each fork (**Fig. 7.15** , step 5). In contrast to DNA polymerases, RNA polymerases can synthesize RNA without a primer. The RNA polymerase required for DNA replication is called DNA primase (DnaG). Primase synthesizes short RNA primers (10–12 nucleotides) at the origin that can launch DNA replication. One primase is loaded at each of the two replication forks.



FIGURE 7.16 ■ Arthur Kornberg and Sylvy Kornberg, circa 1960. A biochemist in her own right, Sylvy Kornberg shared in the discoveries of DNA replication.

BETTMANN/GETTY IMAGES

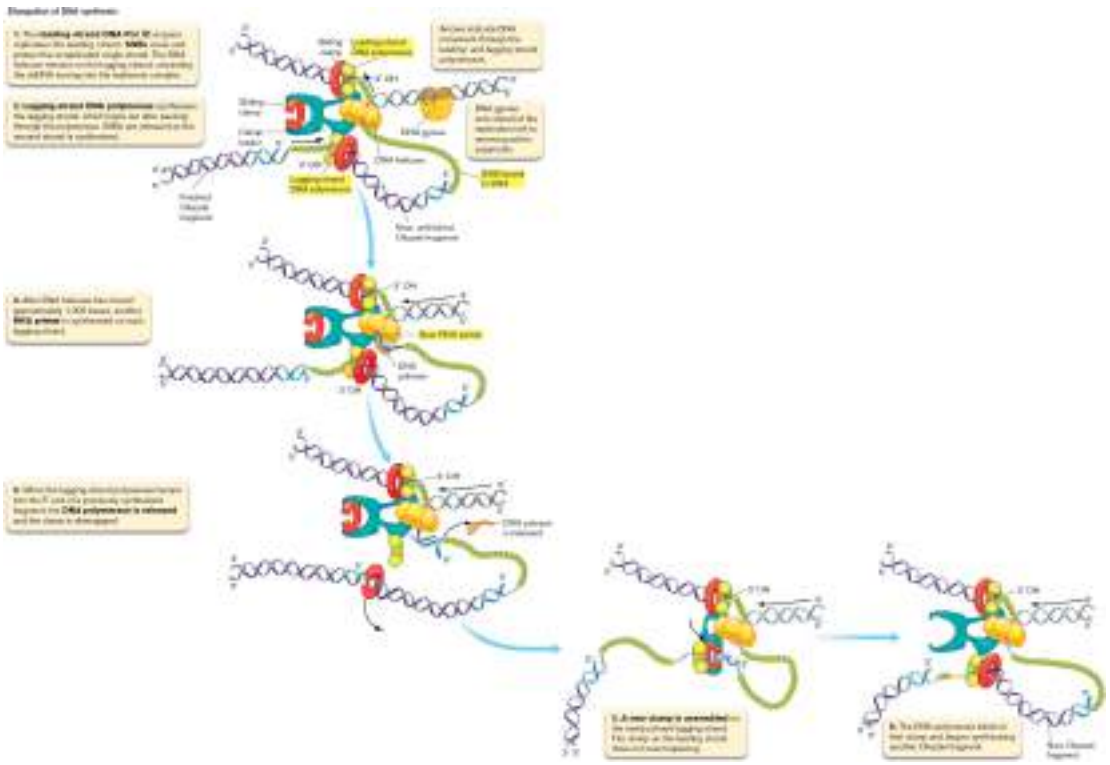


FIGURE 7.17 ■ The DNA polymerase dimer acting at a replication fork. The leading and lagging strands are synthesized simultaneously in the 5′-to-3′ direction. For clarity, the beta clamp on the lagging strand is shown on the opposite side of Pol III as compared to its position on the leading strand.

Why do DNA polymerases require RNA primers? RNA primers may be a holdover from the "RNA world" (an exciting model of molecular evolution, discussed in Chapter 17), when RNA served as the genetic material for life. In addition, the initial stages of nucleotide condensation involved in generating a primer are relatively inaccurate, and distinguishing these early polymers as RNA rather than DNA enables their targeted removal and replacement with DNA synthesized by a high-fidelity DNA polymerase. The mechanism of primer removal and replacement is described later in the chapter.

A sliding clamp tethers DNA polymerase to DNA. At this point, the DNA is almost ready for DNA polymerase. But first a sliding clamp protein (the beta subunit of DNA polymerase III) must be loaded to

keep the DNA polymerase affixed to the DNA (**Fig. 7.15** , step 6). Without this clamp, DNA polymerase would frequently "fall off" the DNA molecule. A multisubunit complex (called the clamp-loading complex) places the sliding clamp, along with an attached pair of DNA polymerase molecules, onto DNA. DNA polymerase (specifically DNA Pol III, discussed next) then binds to the 3′ OH terminus of the primer RNA molecule and begins to synthesize new DNA (step 7).

Elongation of Replicating DNA

Escherichia coli contains five different DNA polymerase proteins, designated Pol I through Pol V. All DNA polymerases catalyze the synthesis of DNA in the 5′-to-3′ direction. However, only Pol III and Pol I participate directly in chromosome replication. The other polymerases conduct operations to rescue stalled replication forks and repair DNA damage.

DNA polymerase III. The main replication polymerase, Pol III, is a complex, multicomponent enzyme. Pol III was another discovery by Arthur Kornberg. The DNA synthesis activity of Pol III is held in the alpha subunit of the complex, while other subunits are used for improving fidelity (accuracy of replication) and processivity (a measure of how long the polymerase remains attached to, and replicates, a template). The Pol III epsilon subunit (DnaQ), for example, contains a proofreading activity that corrects mistakes and improves fidelity.

Proofreading activities within DNA polymerases scan for mispaired bases that have been mistakenly added to a growing chain. A mispaired base is more mobile than the correct base in a DNA molecule because a mispaired base does not properly hydrogen-bond to the template base. This motion halts DNA elongation by Pol III because the base is not properly positioned at the enzyme's active site. Stalling of Pol III activity triggers an intrinsic 3′-to-5′ exonuclease activity in the epsilon subunit. Exonucleases degrade DNA starting from either the 5′ end or the 3′ end, depending on the enzyme. The exonuclease activity of Pol III cleaves the phosphodiester link,

releasing the improperly paired base from the growing chain. Once the wayward base has been excised, Pol III can resume elongation. In *E. coli* , proofreading by the Pol III exonuclease is extremely effective, limiting errors to about one in every 10^8 bases replicated (see Section 9.2).

Both DNA strands are elongated concurrently. After initiation, each replication fork contains one elongating 5′-to-3′ strand, called the "leading strand" (look back at **Fig. 7.15** , step 7). But how is the opposite strand at each fork replicated? There are no known DNA polymerases capable of synthesizing DNA in the 3′-to-5′ direction, which would seem to be needed if both strands are to be synthesized concurrently. DNA synthesis of one strand continuing all the way back to the origin is not a solution, because it would leave the unreplicated strand at each fork exposed to possible degradation for too long and would double the time needed to complete DNA replication.

The cell has solved this dilemma by coordinating the activity of two DNA Pol III enzymes in one complex—one for each strand. The two associated Pol III complexes, together with DNA primase and helicase, form the replisome. As the double-stranded DNA (dsDNA) unwinds at the fork, the problem strand loops out, and primase (DnaG) synthesizes a primer. The second Pol III enzyme binds to the primed section of the loop and synthesizes DNA in the 5′-to-3′ direction (imagine the lower template strand in **Figure 7.17** threading from left to right, through the polymerase ring). All the while, the second polymerase moves with the first polymerase (on the leading strand) toward the fork (**Fig. 7.17** , step 1). The two *E. coli* replisomes move along DNA toward opposite poles of the cell, but generally stay within the middle third of the cell. They eventually meet again at the terminator located at midcell.

Note that concurrent extension of the two strands at a single fork requires that synthesis of the looped strand must lag behind synthesis of the leading strand, and also that new RNA primers must be synthesized by DNA primase (DnaG) every 1,000 bases or so. Thus, the lagging strand is synthesized discontinuously in Okazaki fragments, while the leading strand can be synthesized continuously.

As the leading strand moves forward, advancing the fork, there remains a long stretch of lagging-strand complementary to the already replicated leading strand. This lagging strand is single-stranded but protected by single-stranded DNA-binding proteins (SSBs), which are released from the DNA as the second strand is synthesized (**Fig. 7.17** , step 2).

After about 1,000 bases, DNA primase reenters and synthesizes a new RNA primer in anticipation of lagging-strand DNA synthesis (**Fig. 7.17** , step 3). At some point, the lagging-strand polymerase bumps into the 5′ end of the previously synthesized fragment. This interaction causes DNA polymerase to disengage from that strand (step 4), and the clamp loader loads a new clamp near the new RNA primer (step 5). The DNA polymerase binds to that clamp and begins synthesizing another Okazaki fragment (step 6).

The model just presented assumes that the replisome contains two DNA polymerase III molecules. However, the replisome actually consists of three polymerases—one on the leading strand and two on the lagging strand. The second polymerase on the lagging strand comes into play only when a large gap of unreplicated DNA remains on the lagging strand. For simplicity, the second lagging strand polymerase is not included in the model shown in **Figure 7.17** .

DNA polymerase I. Discontinuous DNA synthesis results in a daughter strand containing Okazaki fragments—long stretches of DNA punctuated by tiny patches of RNA primers. The RNA must be replaced with DNA to maintain chromosome integrity. To remove the RNA, cells typically use the 5′-to-3′ exonuclease activity of Pol I or an RNase enzyme specific for RNA-DNA hybrid molecules (called RNase H). A DNA Pol I enzyme then synthesizes a DNA patch using the 3′ OH end of the preexisting DNA fragment as a priming site (**Fig. 7.18**). When DNA Pol I reaches the next fragment, the enzyme removes the 5′ nucleotide and resynthesizes it. This process of replicating DNA increases accuracy and decreases mutations.

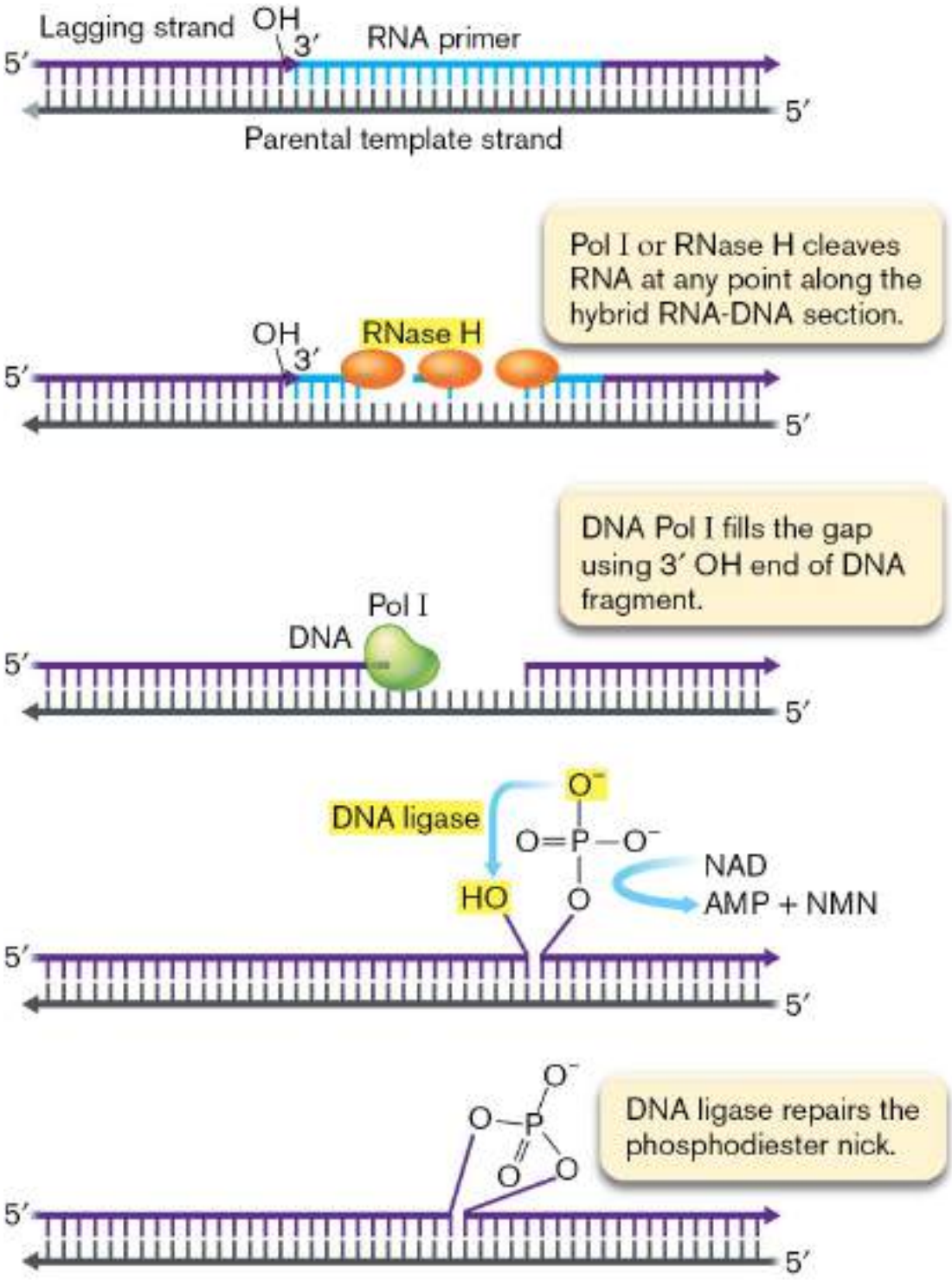


FIGURE 7.18 ■ Removing the RNA primer. The 3′-to-5′ exonuclease activity of RNase H or the 5′-to-3′ exonuclease activity of Pol I cleaves the RNA primer (blue). In either case, Pol I uses the preexisting 3′ OH end of the DNA fragment to fill the gap.

Finally, DNA ligase repairs the phosphodiester nick, using energy derived from the cleavage of NAD. AMP = adenosine monophosphate; NAD = nicotinamide adenine dinucleotide; NMN = nicotinamide monophosphate.

Once DNA Pol I stops synthesizing, it cannot join the 3′ OH of the last added nucleotide with the 5′ phosphate of the abutting fragment. The resulting nick in the phosphodiester backbone is repaired by DNA ligase , which in *E. coli* and many other bacteria uses energy gained by cleaving nicotinamide adenine dinucleotide (NAD) to form the phosphodiester link (**Fig. 7.18**). With this final step, the Okazaki fragments are stitched into a continuous, unbroken strand of replicated DNA.

DNA replication generates positive supercoils. As template DNA is threaded through the replisome, the helicase continually pulls apart the two strands of the DNA helix. As a result, the DNA ahead of the fork twists, introducing positive supercoils. (Try this yourself: Twist two pieces of string together, staple one end of the duplex to a piece of cardboard, and then pull the two strands apart from the free end. Notice the supercoiling that takes place ahead of the moving fork.)

The increasing torsional stress in the chromosome could stop replication by making strand separation more and more difficult. Torsional stress is relieved by DNA gyrase (see Fig. 7.10), positioned ahead of the fork to remove the positive supercoils as they form (see **Fig. 7.17**).

DNA replication is very fast. The elongation phase of DNA replication involves multiple steps—unwinding, priming, leading- and lagging-strand synthesis, proofreading, relaxing of supercoils—that require a complex machinery to accomplish. Perhaps most impressive about this process is its speed, which in bacteria approaches 1,000 nucleotides per second! This is about ten times faster than DNA replication in eukaryotes, and ten times faster than the rate of RNA polymerization during transcription (see Chapter 8).

Note: To put the rate of DNA replication into context, if DNA polymerase III were scaled to the size of an automobile, it would move at a speed of about 375 miles per hour. This is faster than the fastest recorded speed of a 10,000-horsepower dragster at the end of a quarter-mile race (330 miles per hour, as of 2017). The dragster's task is simply to move from point A to point B, whereas the polymerase must make a (near-perfect) copy of DNA at the same time it moves.

Terminating Replication and Segregating Sister Chromosomes

Bidirectional replication of a circular bacterial chromosome results in the two replication forks trying to replicate through the same DNA sequences 180° from the origin; that is, halfway around the chromosome. What tells the polymerases to stop? The *E. coli* chromosome has as many as ten terminator sequences (*ter*) that polymerases enter but rarely, if ever, leave (**Fig. 7.19A**). The *ter* sequences are bound by Tus (t erminus u tilization s ubstance) proteins that allow polymerase movement in only one direction. One set of *ter* -Tus complexes deals solely with the clockwise -replicating polymerase, while the other set halts DNA polymerases replicating counterclockwise relative to the origin. Which terminator site is used depends in part on whether replication of one fork has lagged behind replication of the other. We don't yet know how the two circle ends are joined and the two replisomes detach from the chromosome.

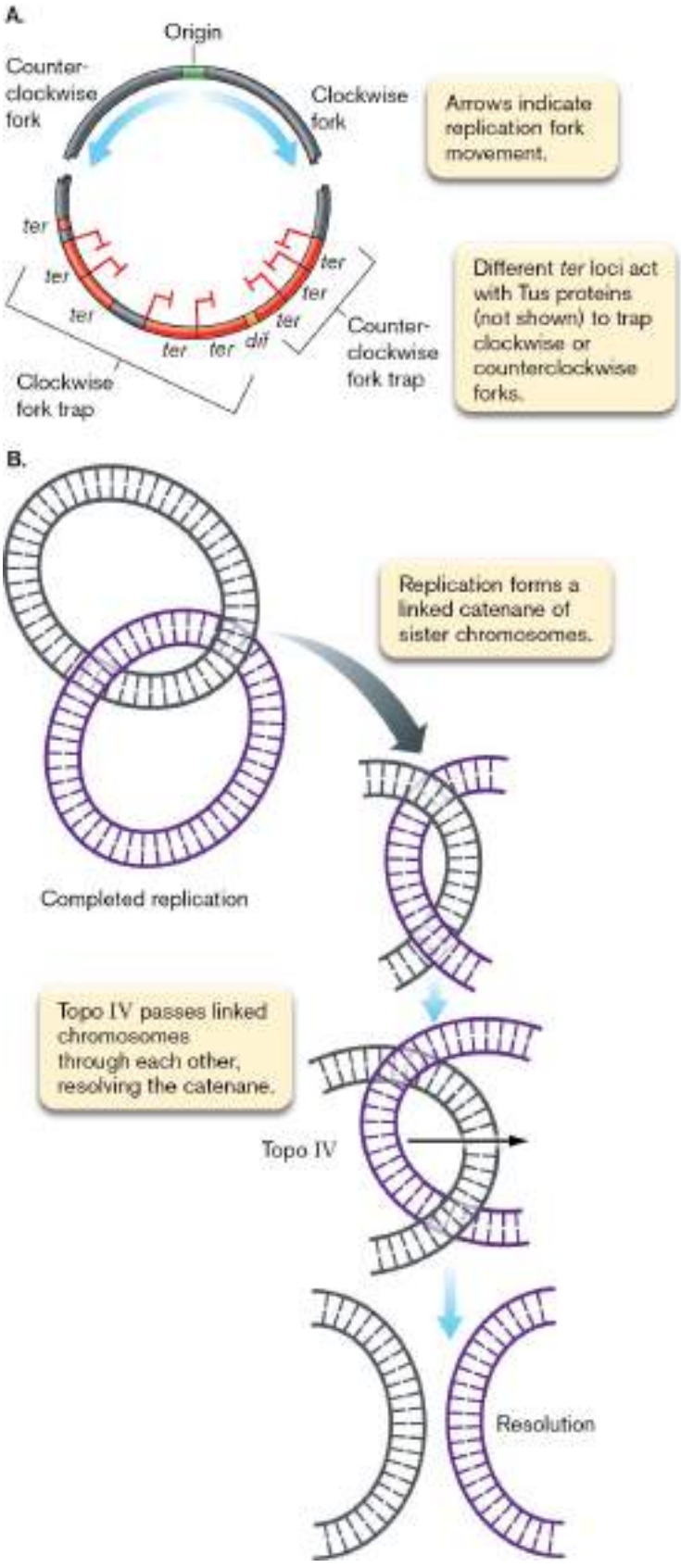

A.
Origin
Counter-clockwise fork
Clockwise fork
Arrows indicate replication fork movement.
ter
ter
ter
ter
ter
ter
ter
ter
ter
dif
Counter-clockwise fork trap
Clockwise fork trap
Different ter loci act with Tus proteins (not shown) to trap clockwise or counterclockwise forks.
B.
Replication forms a linked catenane of sister chromosomes.
Completed replication
Topo IV passes linked chromosomes through each other, resolving the catenane.
Topo IV
Resolution

FIGURE 7.19 ■ Terminating replication of the chromosome. A. Terminator regions for DNA replication on the *E. coli* chromosome. The role of the *dif* site is described in Fig. 7.20. **B.** Resolution of DNA replication catenanes by Topo IV.

The structure of circular chromosomes poses "knotty" problems for their segregation after replication. These problems are solved by special enzymes that cut and re-join the DNA strands. The first problem develops as soon as replication begins. Replicated DNA molecules occasionally form knots between homologous genes on sister chromosomes, holding them together. These knots, called pre-catenanes , must be removed before sister chromosomes can segregate to opposite cell poles. The second "knotty" problem is encountered when a chromosome finishes replicating. Because of the topology of the chromosome, the two daughter molecules will appear as a catenane , a pair of linked rings. The rings must be unlinked so that sister chromosomes can segregate at termination (**Fig. 7.19B**).

Resolving these knots requires the enzyme topoisomerase IV, a type II topoisomerase similar to DNA gyrase. Topo IV activity is stimulated by SeqA, the protein that binds to newly replicated, hemimethylated DNA. Recall that SeqA also prevents reinitiation at newly replicated origins. This protein therefore serves a dual function to prevent the start of a premature round of replication and to remove topological constraints to chromosome segregation.

Yet another challenge for microbes with circular chromosomes is to resolve homologous recombination events that occur along the duplicated regions of the chromosome. Homologous, or general, recombination events (covered in detail in Chapter 9) involve the exchange of strands between two molecules of DNA; exchange occurs within extensive regions of identical or nearly identical sequence. Because duplicated chromosomes are essentially identical, crossovers between these DNA molecules are common.

If replication terminates with an odd number of recombination events between the sister chromosomes, they end up as an end-to-end two-chromosome dimer (**Fig. 7.20** , step 1). To resolve this

dimer, a final recombination event is facilitated by proteins called XerC and XerD, which function at a specific 28-bp site, called *dif* , located in the terminus region of the chromosome. XerC and XerD are activated at the *dif* locus by the DNA translocase FtsK. FtsK monomers assemble as a ring structure around the DNA molecule, and this ring translocates along the DNA to the *dif* locus by following special "arrows," sequences known as KOPS (Fts K - o rienting p olar s equences) that are positioned along the chromosome (**Fig. 7.20** , step 2). The polar orientations of the KOPS are opposite on the two replicating arms of the chromosome, such that the FtsK ring structure is directed to the *dif* site no matter which arm it assembles on.

Once at the *dif* site, FtsK activates XerC and XerD, which catalyze a second homologous recombination event (**Fig. 7.20** , step 3) that resolves the dimer into two separate chromosomes (step 4). Note that the two sister chromosomes have exchanged the DNA between the recombination sites, but unless mutation occurred during replication, these exchanged regions of DNA are identical.

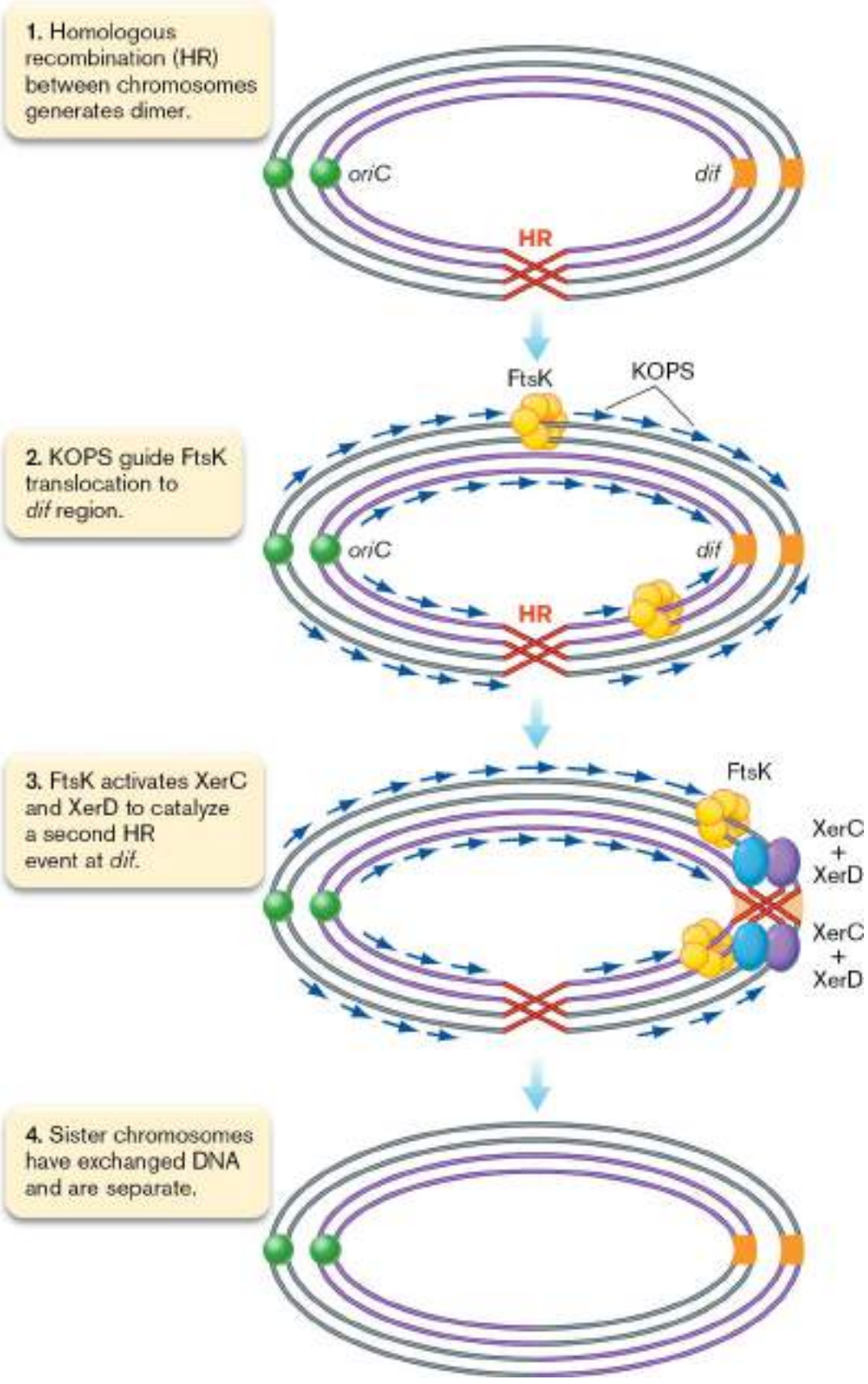

1. Homologous recombination (HR) between chromosomes generates dimer.
oriC
dif
HR
FtsK
KOPS
2. KOPS guide FtsK translocation to dif region.
oriC
dif
HR
3. FtsK activates XerC and XerD to catalyze a second HR event at dif.
FtsK
XerC + XerD
XerC + XerD
4. Sister chromosomes have exchanged DNA and are separate.

FIGURE 7.20 ■ Resolution of chromosome dimers by XerC and XerD at the *dif* locus. The initial homologous recombination (HR) event results in the dimer, and resolution at the *dif* site results in separated chromosomes that have exchanged a segment of their DNA.

FtsK is an important protein for the termination of chromosome replication: It activates not only XerC and XerD to resolve dimers but also Topo IV to resolve catenanes. And using KOPS to act as a pump, FtsK can send DNA from midcell toward both poles at exceptionally high rates (1,700–16,000 base pairs per second!), thus ensuring that the sister chromosomes partition to the two halves of the cell before the septum forms at midcell.

Thought Questions

7.6 Would you expect to find genes encoding Topo IV , XerC, and XerD in prokaryotes with exclusively linear chromosomes? Why or why not?

7.7 Individual cells in a population of *E. coli* typically initiate replication at different times (asynchronous replication). However, depriving the population of a required amino acid can synchronize reproduction of the population. Ongoing rounds of DNA synthesis finish, but new rounds do not begin. Replication stops until the amino acid is once again added to the medium—an action that triggers simultaneous initiation in all cells. Why is this replication synchron

7.8 The antibiotic rifampin inhibits transcription by RNA polymerase, but not by primase (DnaG). What happens to DNA synthesis if rifampin is added to a synchronous culture?

To Summarize

- **DNA replication is semiconservative** , with newly synthesized strands lengthening in a 5′-to-3′ direction. Replication consists of initiation, elongation, and termination steps.
- **Bacterial DNA replication is initiated** from a fixed DNA origin. Initiation depends on the mass and size of the growing cell. It is controlled by the accumulation of initiator and repressor proteins and by methylation at the origin.
- **During elongation** , primase (DnaG) lays down an RNA primer, DNA polymerase III synthesizes a DNA strand extending from the RNA primer, and a sliding clamp keeps DNA Pol III attached to the template DNA molecule.
- **The 3′-to-5′ proofreading** activity of Pol III corrects accidental errors during polymerization.
- **DNA ligase** joins Okazaki fragments in the lagging strand.
- **Two replisomes, each containing three DNA Pol III complexes** , move in opposite directions along the DNA.
- **Termination** involves stopping replication forks halfway around the chromosome at *ter* sites.
- **Ringed catenanes** formed at the completion of replication are separated by topoisomerase IV.
- **Chromosome dimers** are resolved by proteins XerC and XerD at the *dif* site.

Glossary

semiconservative
: Describing the mode of DNA replication whereby each new double helix contains one old, parental strand and one newly synthesized daughter strand.

replication fork

During DNA synthesis, the region of the chromosome that is being unwound.

DNA replication
: The biological process of making an identical copy of double-stranded DNA using existing DNA as a template.

origin (*oriC*)
: The region of a bacterial or archaeal chromosome where DNA replication initiates.

termination (*ter*) site
: A bacterial sequence of DNA that halts replication of DNA by DNA polymerases elongating from both directions around the circular genome.

Okazaki fragments
: Short fragments of DNA that are synthesized on the lagging strand during DNA synthesis.

replisome
: A complex of DNA polymerase and other accessory molecules that performs DNA replication.

primase
: An RNA polymerase that synthesizes short RNA primers complementary to a DNA template to launch DNA replication.

sliding clamp
: A protein that keeps DNA polymerase affixed to DNA during replication.

proofreading
: An enzymatic activity of some nucleic acid polymerases that attempts to correct mispaired bases.

exonuclease
: An enzyme that cleaves DNA from the end.

DNA ligase
: An enzyme that cells use to form a covalent bond at a nick in the phosphodiester backbone. It is also used in molecular biology laboratories to join pieces of DNA.

pre-catenane
: A knot of intertwined DNA that is generated during chromosome replication.

catenane

A pair of linked rings of DNA that occurs as a by-product of replication of circular chromosomes.

homologous recombination

The process by which two DNA molecules exchange arms by cutting and splicing their helix backbones. Exchange occurs between sequences that are identical or nearly identical, as the machinery requires complementary base pairing to exchange the DNA molecules.

Figure 3.28

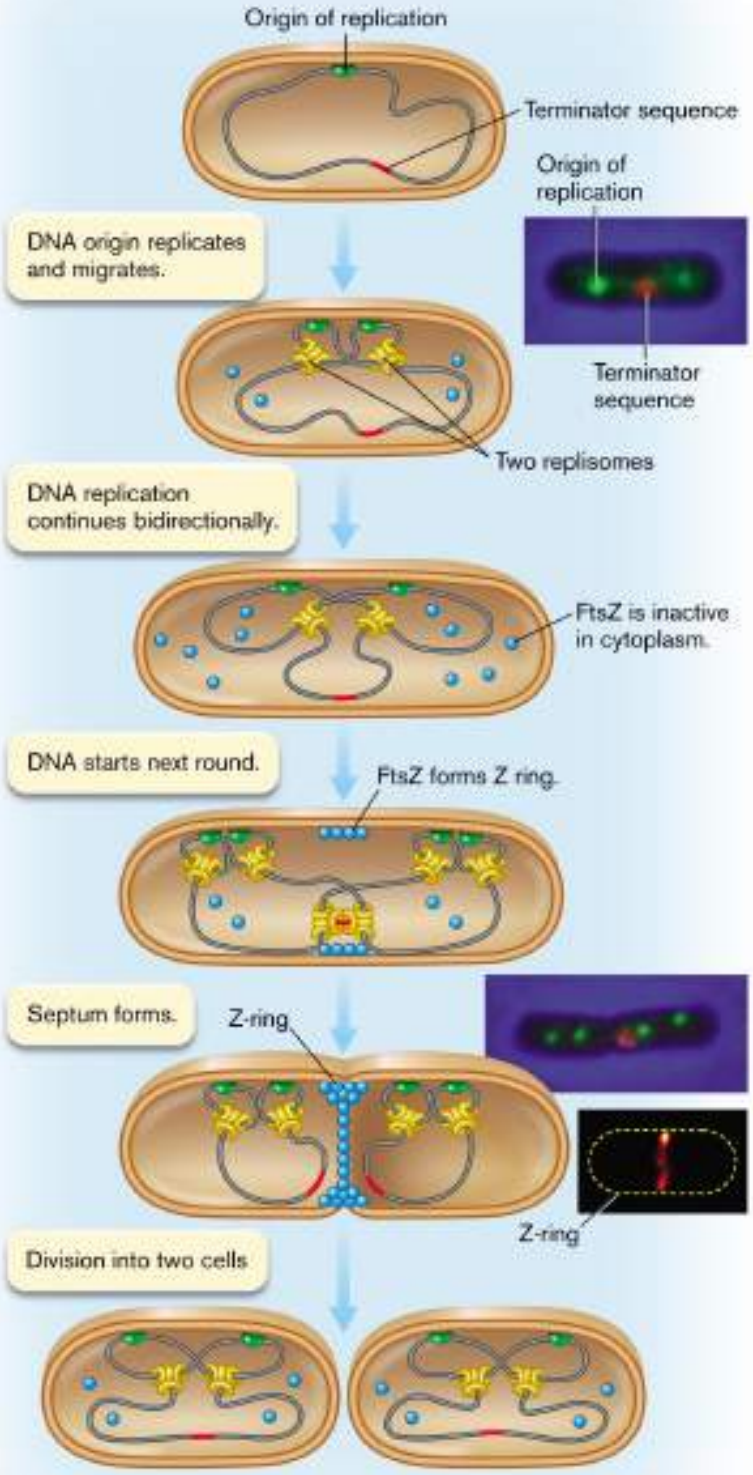

Origin of replication
Terminator sequence
Origin of replication
DNA origin replicates and migrates.
Terminator sequence
Two replisomes
DNA replication continues bidirectionally.
FtsZ is inactive in cytoplasm.
DNA starts next round.
FtsZ forms Z ring.
Septum forms.
Z-ring
Z-ring
Division into two cells

FIGURE 3.28 ■ Replisome movement within a dividing cell. The DNA origin-of-replication sites (green) move apart in the expanding cell as the two replisomes (yellow) stay near the middle, where they replicate around the entire chromosome, completing the terminator sequence last (red). As the terminator sequence nears completion, FtsZ proteins assemble the Z-ring organizing septum formation.

Source : Top 2 insets: Ivy Lau et al. 2003. Mol. Microbiol. **49** :731. Bottom inset: Jackson Buss et al. 2015. *PLoS Genet* . **11** (4).

I. LAU ET AL. 2003. *MOL. MICROBIOL.* **49** :731, FIG. 2A
I. LAU ET AL. 2003. *MOL. MICROBIOL.* **49** :731, FIG. 2A
J. BUSS ET AL. 2015. *PLOS GENET.* **11** (4):E1005128, FIG. 1G

Fig. 7.3A

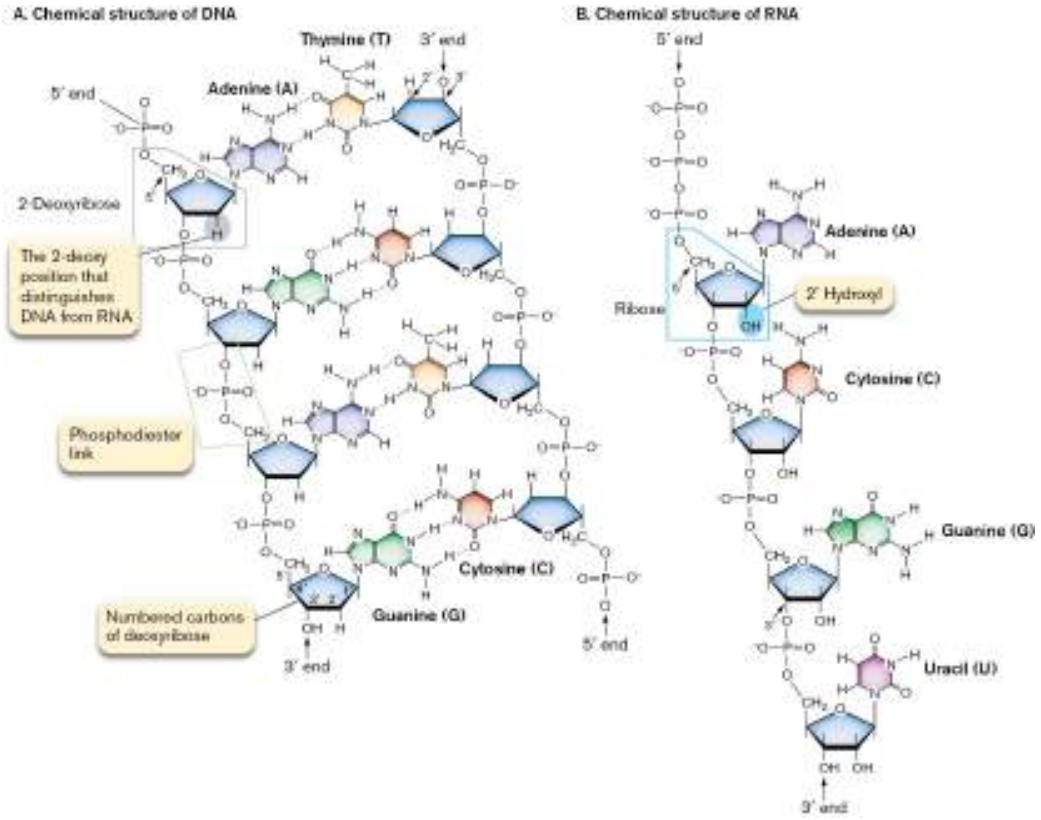
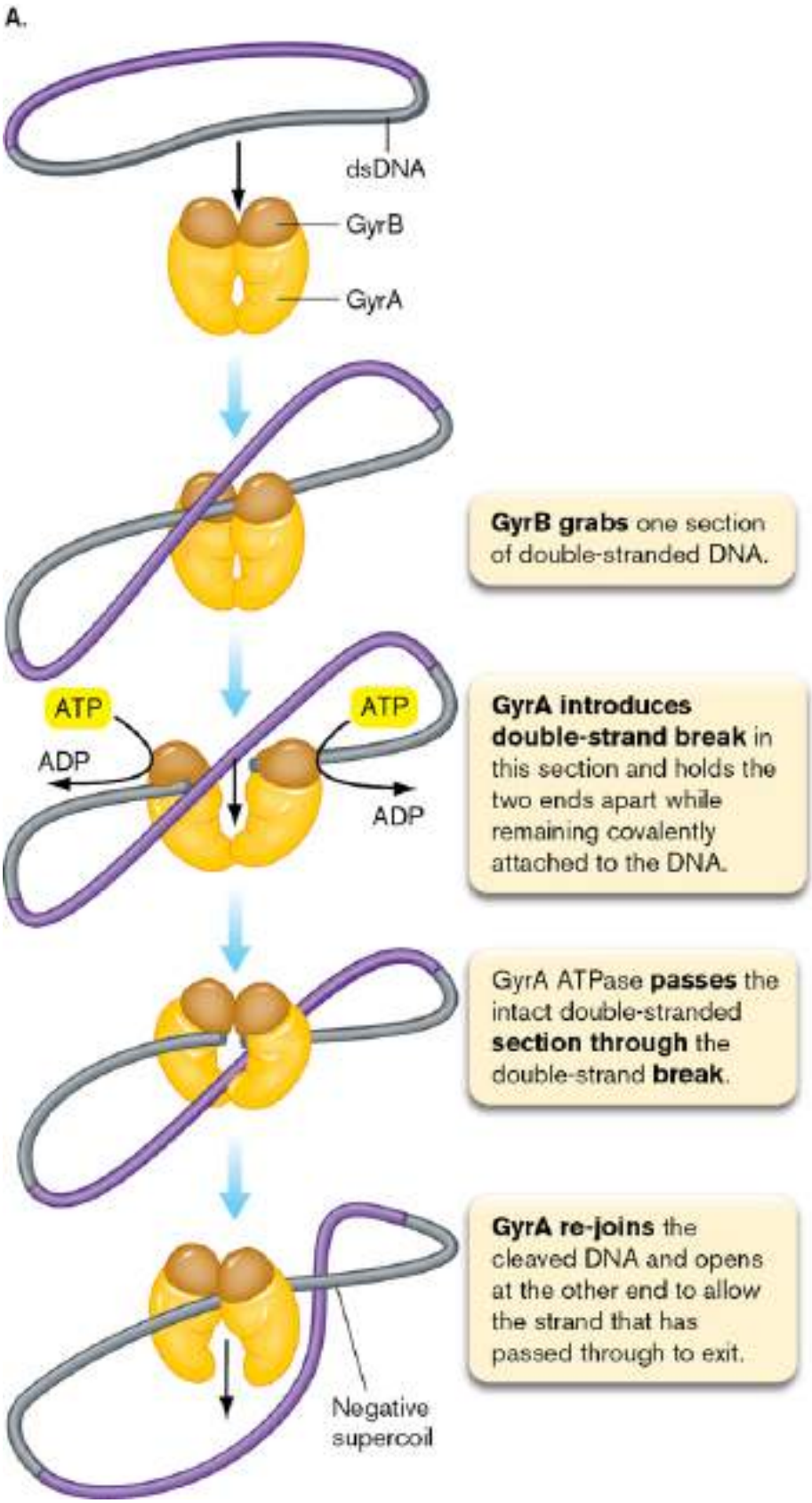


FIGURE 7.3 ■ Structures of DNA and RNA. A. In the cell, DNA bases are added only to a preexisting 3′ OH of a nucleoside monophosphate, so the 5′ ends in this figure are

drawn as nucleoside monophosphates. (Dotted lines indicate hydrogen bonds between bases.) **B.** Cellular RNA molecules, however, begin with a 5′ triphosphate.

Fig. 7.10


A.
dsDNA
GyrB
GyrA
ATP
ADP
ATP
ADP
Negative supercoil
GyrB grabs one section of double-stranded DNA.
GyrA introduces double-strand break in this section and holds the two ends apart while remaining covalently attached to the DNA.
GyrA ATPase **passes** the intact double-stranded **section through** the double-strand **break**.
GyrA re-joins the cleaved DNA and opens at the other end to allow the strand that has passed through to exit.

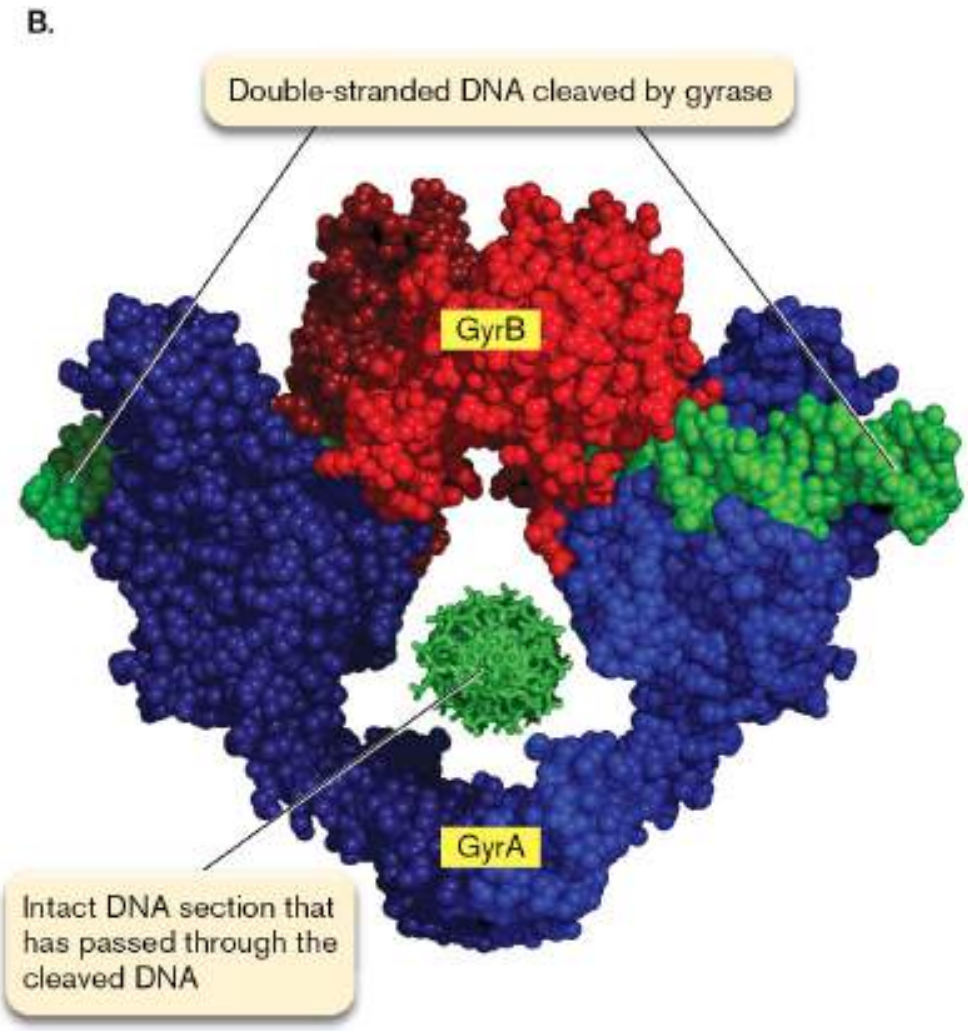


FIGURE 7.10 ■ Mechanism of action for type II topoisomerases. A. Mode of action of DNA gyrase from *E. coli* . **B.** Three-dimensional representation of DNA gyrase. The gyrase complex grips a broken DNA duplex (shown in green) and has transported a second duplex (the multicolored rosette) through the break.

COURTESY OF JAMES BERGER

7.4 Plasmids and Secondary Chromosomes

In this section we discuss genomes that are split into more than one replicating DNA molecule. The additional elements, plasmids and secondary chromosomes, complement the primary chromosome by adding various types of genes to the genome. Some microbes have multiple plasmids or secondary chromosomes that contribute to their genomes (**Table 7.1**). These DNA elements vary in size and genetic composition. While some coordination with the primary chromosome can exist, plasmids and secondary chromosomes control their own replication and copy number. Some plasmids can be lost during cell division, and some can be shared between distantly related species. This dynamic nature of plasmids challenges the notion that a microbial species is defined by a unique genome shared by all members of that species.

Plasmids Vary in Size, Copy Number, and Cargo

Plasmids are extrachromosomal elements found in bacteria, archaea, and eukaryotic microbes. Plasmids are usually circular and negatively supercoiled. They are typically much smaller than chromosomes (several thousand compared to several million base pairs) and may encode only a few genes (**Fig. 7.21**). Copy number varies widely among plasmid types, from a single copy to over 500 per cell. Genes found in high-copy-number plasmids can be dramatically overexpressed relative to genes found on chromosomes and low-copy-number plasmids, and this overexpression can be advantageous if a high concentration of the gene product is important for the gene's function in the cell.

A.



B.

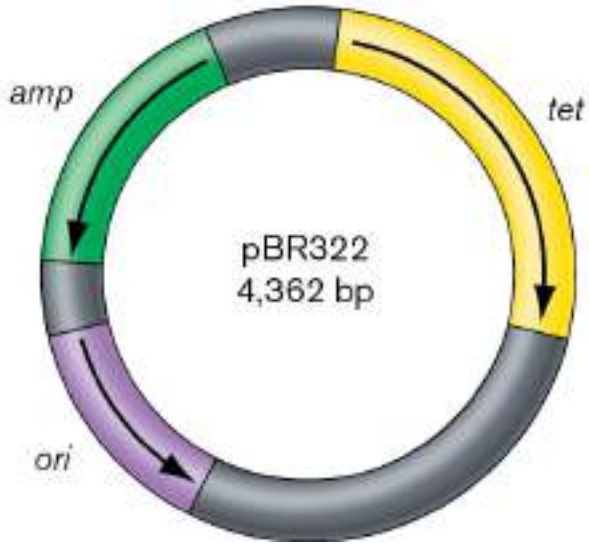


FIGURE 7.21 ■ Plasmid map. A. In this colorized transmission electron micrograph, note the huge difference in size between a circular plasmid DNA molecule (arrow) and chromosomal DNA after both are gently released from a rod-shaped cell approximately 1 μm in length. **B.** Map of plasmid pBR322. This plasmid contains an origin of replication (*ori*) and

genes encoding resistance to ampicillin (*amp*) and to tetracycline (*tet*).

SCIENCE VU/DRS. H. POTTER-D. DRESSLER/VISUALS UNLIMITED, INC.

The genes that plasmids carry are not essential for the basic functions involved in cell growth and metabolism, but they can play critical roles in certain situations. One important class of genes that some plasmids carry confers resistance to antibiotics (**Fig. 7.21B** ; discussed in Chapter 27). Antibiotic resistance plasmids benefit bacteria, but they are a major problem for modern hospitals, where plasmids carrying multiple drug resistance genes are transmitted from harmless bacteria into pathogens. On the other hand, plasmids like pBR322 that contain drug resistance genes are the workhorses of genetic technology and have benefited society tremendously. The use of plasmids like pBR322 for gene cloning is described in eAppendix 3 .

Other kinds of host survival genes carried by plasmids include genes providing resistance to toxic metals, genes encoding toxins that aid pathogenesis, and genes encoding proteins that enable symbiosis. Most of the genes involved in the nitrogen-fixing symbiosis of *Rhizobium* , for example, are plasmid-borne. Genes encoding enzymes for antibiotic synthesis can be found on plasmids in *Streptomyces* . Notably, these plasmids are linear, rather than circular. Far from being freeloaders, plasmids often contribute significantly to the physiology of an organism.

Plasmids Are Transmitted between Cells

Plasmids have played a pivotal role in evolution because they are easily passed among bacterial species. This movement of plasmids is particularly evident today because of the rapid spread of antibiotic resistance, whose genetic basis is often a plasmid-encoded gene.

Some plasmids are self-transferable via conjugation, a process that requires cell-to-cell contact to move the plasmid from a donor cell to a recipient. Other plasmids are incapable of conjugation

(nontransmissible). Thus, they only propagate with the host, when the host genome undergoes replication. A third group can be recognized and transferred by the conjugation machinery produced by a co-occurring "helper" plasmid. Any plasmid released from dead cells can also be taken up by some bacteria in a process called transformation. Finally, plasmids can be transmitted in nature by accidentally being packaged into bacteriophage head coats; in other words, by bacteriophage transduction. (Conjugation, transformation, and transduction are discussed in Section 9.3 .)

Plasmids Regulate Their Replication

Plasmids may "borrow" the replication machinery of the host chromosome, but they have their own origin sequences and initiator proteins to control the frequency of replication. This autonomy enables plasmids to regulate copy number independently of the host chromosome. Plasmids can replicate in two different ways: by rolling-circle or bidirectional replication.

Used by some plasmids, rolling-circle replication (**Fig. 7.22**) is unidirectional, not bidirectional like chromosomal replication. Rolling-circle replication initiates when the initiator protein RepA, encoded by a plasmid gene, binds to the origin of replication and nicks one strand. RepA holds on to one end (5′ PO_4) of the nicked strand, while the other end (3′ OH) serves as a primer for host DNA polymerase to replicate the intact, complementary strand. The RepA initiator protein recruits a helicase that unwinds DNA, which becomes coated by single-stranded DNA-binding proteins.

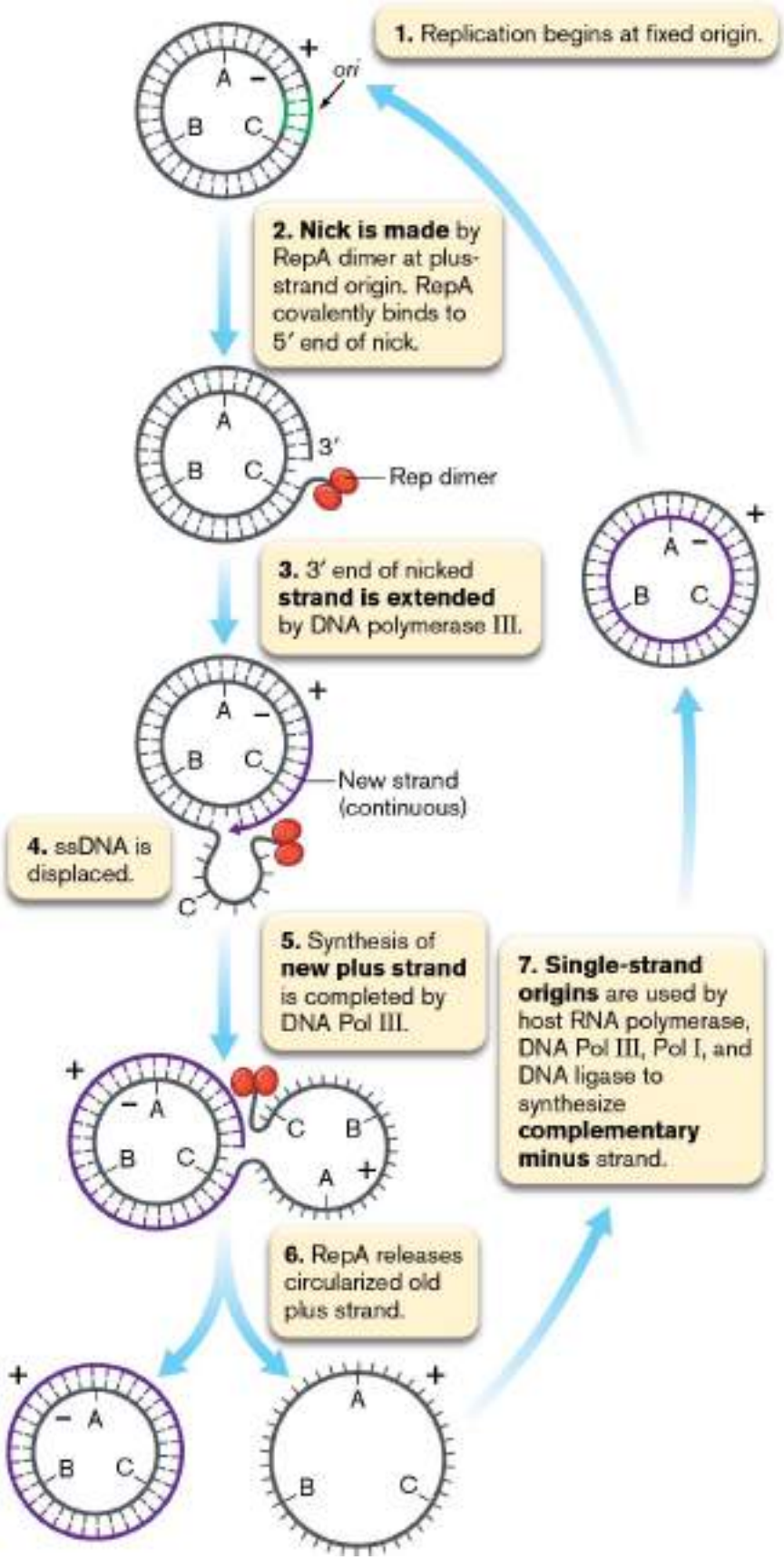

1. Replication begins at fixed origin.
ori
A
B
C
+
−
2. Nick is made by RepA dimer at plus-strand origin. RepA covalently binds to 5′ end of nick.
3′
Rep dimer
3. 3′ end of nicked strand is extended by DNA polymerase III.
New strand (continuous)
4. ssDNA is displaced.
5. Synthesis of new plus strand is completed by DNA Pol III.
6. RepA releases circularized old plus strand.
7. Single-strand origins are used by host RNA polymerase, DNA Pol III, Pol I, and DNA ligase to synthesize complementary minus strand.

FIGURE 7.22 ■ Plasmid replication: rolling-circle model. For simplicity, SSB proteins bound to the single-stranded DNA (ssDNA) are not shown.

As replication proceeds, the nicked strand progressively peels off without replicating, until the strand is completely displaced. Then the two ends of the displaced but nicked single strand are re-joined by the RepA protein and released. The single-stranded circular molecule is protected by host single-stranded DNA-binding proteins until host enzymes replicate a complementary strand and regenerate a double-stranded molecule.

Bidirectional replication of plasmids proceeds from a single origin, and both forks terminate at a single termination site, in a manner similar to initiation and termination of the bacterial chromosome. The molecular details of replication initiation of these plasmids share similarities to chromosomal replication, including the involvement of SeqA and DNA methylation. In most cases, binding of the chromosomal initiator protein DnaA near the origin is also involved; however, DnaA is not the master initiator, and its role in plasmid initiation is not completely understood. Instead, the initiation of replication is controlled by the plasmid-encoded Rep initiator protein, which binds near the origin and functions to melt the double helix at the origin and recruit the helicase to unwind the DNA and initiate replication. Note that the action of these Rep proteins is different from the action of RepA in rolling-circle replication. The sites at which the Rep proteins bind are called iterons (**Fig. 7.23**). Iterons are direct repeats of 17–22 base pairs and vary from two to seven copies in plasmids analyzed to date.

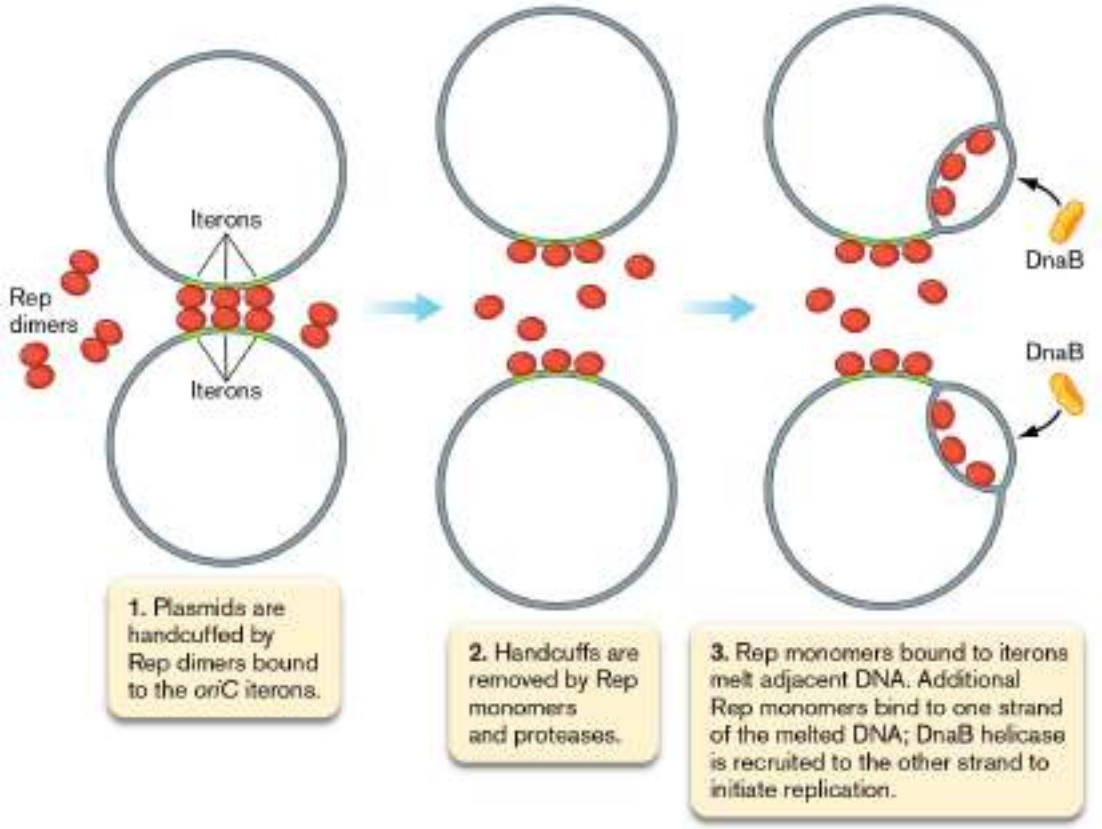


FIGURE 7.23 ■ Regulation of plasmid replication by handcuffing. Dimers of RepA handcuff two plasmids at the iteron loci. Removal of the dimers allows the Rep monomers to initiate replication.

Rep can exist as both monomers and dimers, and this dual state is critical to its function in replication initiation. Both monomers and dimers can bind iteron DNA, but only monomers initiate replication. Rep dimer formation prevents replication both by sequestering monomers and by forming an inactivating bridge between the iterons of two plasmids in a process called "handcuffing" (**Fig. 7.23**). Dimerization and handcuffing are thought to help control the rate of replication initiation to ensure that the plasmid does not overload the host cell with excess copies. Initiation occurs when the handcuffs are disrupted by excess monomers and dimer-targeting proteases. Monomers bound to the iteron then melt the adjacent DNA and recruit DnaB to the plasmid to initiate replication.

Plasmid “Tricks” Ensure Inheritance

Given that plasmids do not carry essential genes, can cells easily lose their plasmids? When a plasmid-containing host cell divides, is it just chance that determines whether both daughter cells inherit the plasmid? In some instances the answer is yes. To maintain themselves in the host cell, plasmids can employ a number of different strategies. Some plasmids ensure their inheritance by carrying genes whose functions benefit the host bacterium under certain conditions. For instance, as discussed earlier, some bacterial plasmids confer resistance to antibiotics. As long as the antibiotic is present in the environment, any cell that loses the plasmid will be killed or stop growing. Still one more method of retention for the plasmid is to integrate into the host chromosome—a process described in detail in Chapter 9 .

One of the surest means of plasmid retention is to flood the cytoplasm of the host with many copies. As the daughter cells split the cytoplasmic inheritance of the mother cells after cell division, there is a very high likelihood that both daughter cells will also inherit at least one copy of the plasmid.

Not all plasmids have a high number of copies, however. Low-copy-number plasmids limit how many copies they make to avoid draining the cell of energy. Cells forced to waste energy making many plasmid copies could be at a growth disadvantage when competing with plasmid-less cells in a natural environment. Low-copy-number plasmids evolved clever partitioning systems that ensure both daughter cells will contain copies of the plasmid.

Figure 7.24 illustrates how partitioning works for plasmid R1, a *Salmonella* plasmid that imparts multidrug antibiotic resistance to host bacteria. The process employs three known plasmid-encoded genes, *parC* , *parM* , and *parR,* to push the plasmids to the cell poles. The *parC* DNA sequence is analogous to the centromere of eukaryotic chromosomes. The DNA-binding protein ParR binds to the *parC* sequence and forms a ParR- *parC* plasmid complex. ParM protein is an actin-like molecule that forms long filaments as it hydrolyzes ATP. The ParM filaments are dynamically unstable,

constantly elongating and shortening. However, each end of a ParM actin-like filament can bind to a ParR- *parC* plasmid complex. When both ends of a ParM filament contact ParR- *parC* plasmid complexes, the filament stabilizes and elongates until the two plasmids hit the opposite poles of the cell. The plasmid is thus dislodged from the filament, and the filament quickly dissociates.

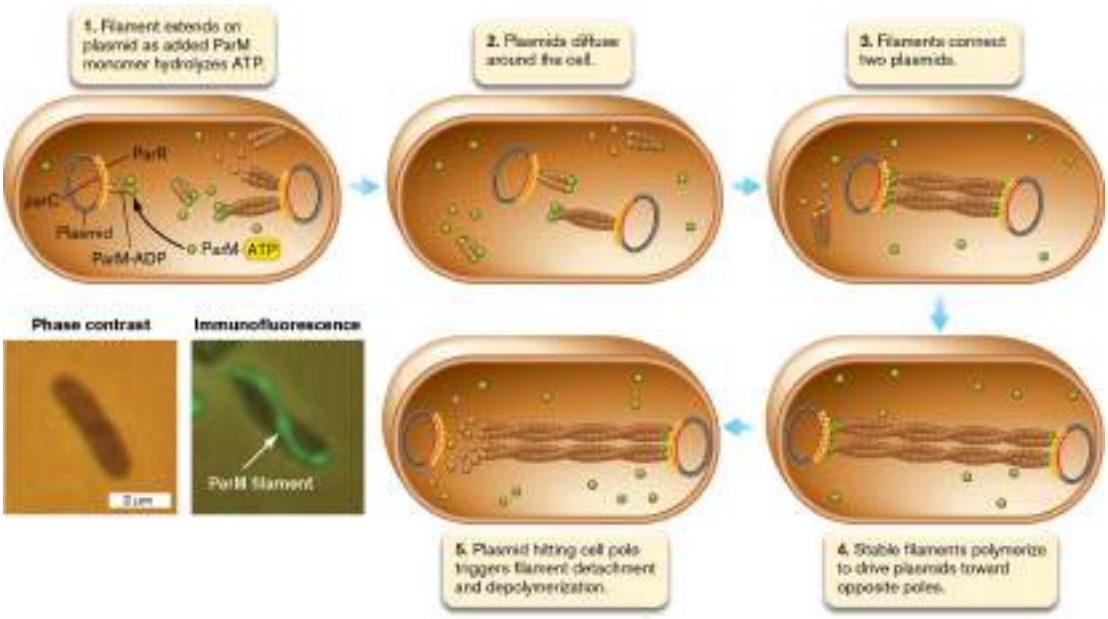


FIGURE 7.24 ■ R1 plasmids segregate via a pushing mechanism. The two micrographs at lower left show a cell with a *parM* mutant plasmid unable to make a ParM filament (right), and a cell with a *parM* $^{+}$ plasmid making a ParM actin-like filament (green; left) to partition daughter plasmids (step 4). ParM filament was seen by combined phase-contrast and immunofluorescence microscopy using rabbit anti-ParM antibodies and fluorescent Alexa 488–conjugated goat antirabbit IgG antibodies.

Source: Modified from Christopher S. Campbell and R. Dyche Mullins. 2007. *J. Cell Biol.* **179** :1059–1066, fig. 4.

J. MØLLER-JENSEN ET AL. 2002. *EMBO J* . **21** :3119–3127

Secondary Chromosomes Carry Essential Genes

In contrast to *E. coli* and most bacteria studied so far, the genome of *Vibrio cholerae* is encoded on two chromosomes (Fig. 7.2). Typically, genomes with multiple chromosomes are composed of one primary chromosome and one or more secondary chromosomes . The secondary chromosomes are smaller than the primary chromosome. A secondary chromosome is distinguished from a plasmid because it carries at least one essential gene . An essential gene is required for cell viability under all environmental conditions. A secondary chromosome typically contains only a few of the essential genes in the genome.

Note: For some strains of *Vibrio cholerae* , the two chromosomes have fused into a single replicating unit. Both origins of replication are present, but it is unclear whether both are operational. The reason for the fusion event is unclear.

Thought Questions

7.9 How would you demonstrate that a gene is essential?

A remarkable feature of the two chromosomes of *Vibrio cholerae* is that their synthesis is synchronized such that replication terminates at the same time. To accomplish this synchrony, the larger Chr1 initiates first, responding to activated DnaA. Replication of a specific locus, *crtS* , downstream of the Chr1 origin triggers the initiation of the smaller Chr2. How this regulation works is not fully understood, but duplication of *crtS* somehow changes the binding affinity of the Chr2 initiator protein RctB, causing it to fall off the

inhibitor region of the Chr2 origin and attach to the activator region within the same origin, initiating replication.

Secondary Chromosomes Evolve from Plasmids

How do secondary chromosomes form during bacterial evolution? One hypothesis is that the secondary chromosome splits off from the primary chromosome to establish its own replicating unit. For this mechanism to work, both chromosomes must have an origin of replication at the time of the split or they would be lost during cell division. The competing hypothesis is that secondary chromosomes evolved from plasmids that captured one or more essential genes from the original chromosome. While the first scenario remains theoretically possible, for all studied secondary chromosomes the evidence supports the plasmid origin hypothesis. The strongest support for the plasmid origin model is the presence of plasmid-type replication initiation and segregation machinery for every secondary chromosome examined to date. No primary chromosome–type replication proteins have been found associated with secondary chromosomes.

Plasmids are thought to evolve into secondary chromosomes via a series of gene transfer events (**Fig. 7.25**). A foreign plasmid is acquired by a cell with a single chromosome, and through horizontal gene transfer the plasmid can acquire additional (nonessential) genes. At some point in its evolution, the plasmid acquires an essential gene, either through translocation of that gene from the chromosome or through acquisition of a second copy of the gene from an outside source. The copy on the primary chromosome can then be lost because it is now redundant with the one on the secondary chromosome. In either scenario, the cell now has two independent DNA molecules containing essential genes. In the case of *Vibrio cholerae* , an additional modification is made to the newly formed secondary chromosome: transfer of the control of initiation from the secondary to the primary chromosome.

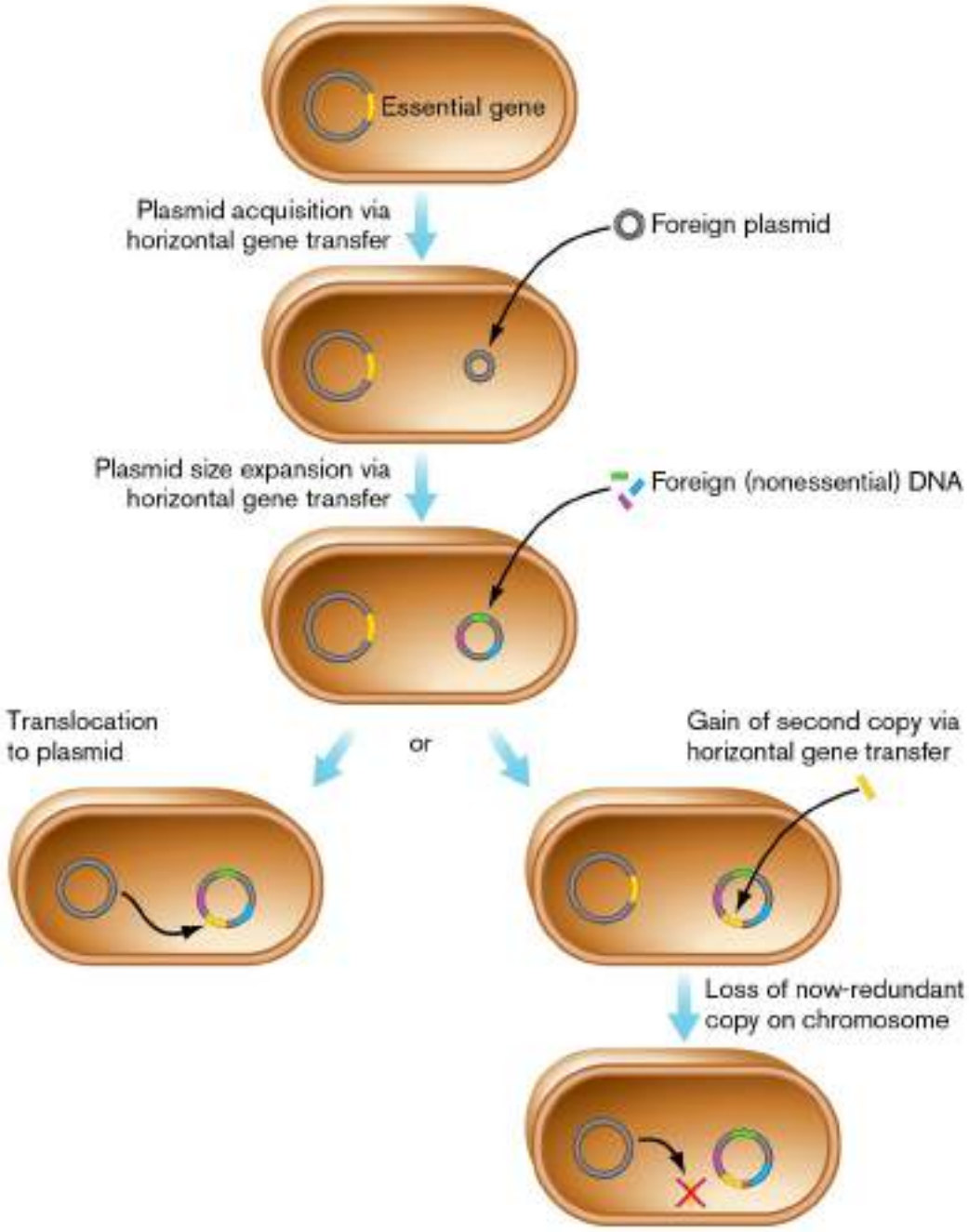


FIGURE 7.25 ■ Model of how organisms acquire a secondary chromosome via evolution of a plasmid. The model highlights the role of horizontal gene transfer in plasmid gene acquisition and shows two possible ways that an essential gene can be found exclusively on the new chromosome.

Regardless of the mechanism, the cells end up with the gain of an essential gene on the secondary chromosome and a corresponding loss of that gene on the primary chromosome. Is there a selective advantage to moving an essential gene onto another DNA element? One thought is that this translocation ensures retention of the second replicon : By marking the replicon with an essential function, it requires that the replicon, and the features it contains, be passed on to progeny. These other features may not be required for growth and reproduction, but they could provide *V. cholerae* with benefits worth maintaining. The nature of those benefits is under active investigation and could provide valuable clues as to why some microbes have multiple chromosomes and some do not.

One popular hypothesis is that the second chromosome is a test bed for evolutionary experimentation and innovation, which involve not only acquisition of new genes via horizontal gene transfer, but optimization of genes via adaptive mutation as well. Chr2 carries the integron island of *V. cholerae* , which is a gene capture system that has acquired several hundred genes, including those implicated in pathogenesis (see Section 9.3). Also, for reasons not well understood, mutations that change the functions of genes are more tolerable on Chr2 than on Chr1.

To Summarize

- **Plasmids are autonomously replicating** circular or linear DNA molecules that are part of a cell's genome.
- **Plasmids can be transferred** between cells.
- **Plasmids replicate** by rolling-circle or bidirectional mechanisms.
- **Secondary chromosomes** are distinguished from plasmids in that they carry essential genes.

- **Secondary chromosomes** evolve from plasmids by acquiring essential genes.

Glossary

plasmid
: An extrachromosomal genetic element that may be present in some cells. Plasmids carry no essential genes.

secondary chromosome
: A plasmid-like small chromosome that carries at least one essential gene.

essential gene
: A gene that is required for cell viability under all environmental conditions.

replicon
: A nucleic acid molecule such as a chromosome or plasmid that is replicated autonomously, with its own origin of replication.

Fig. 7.2

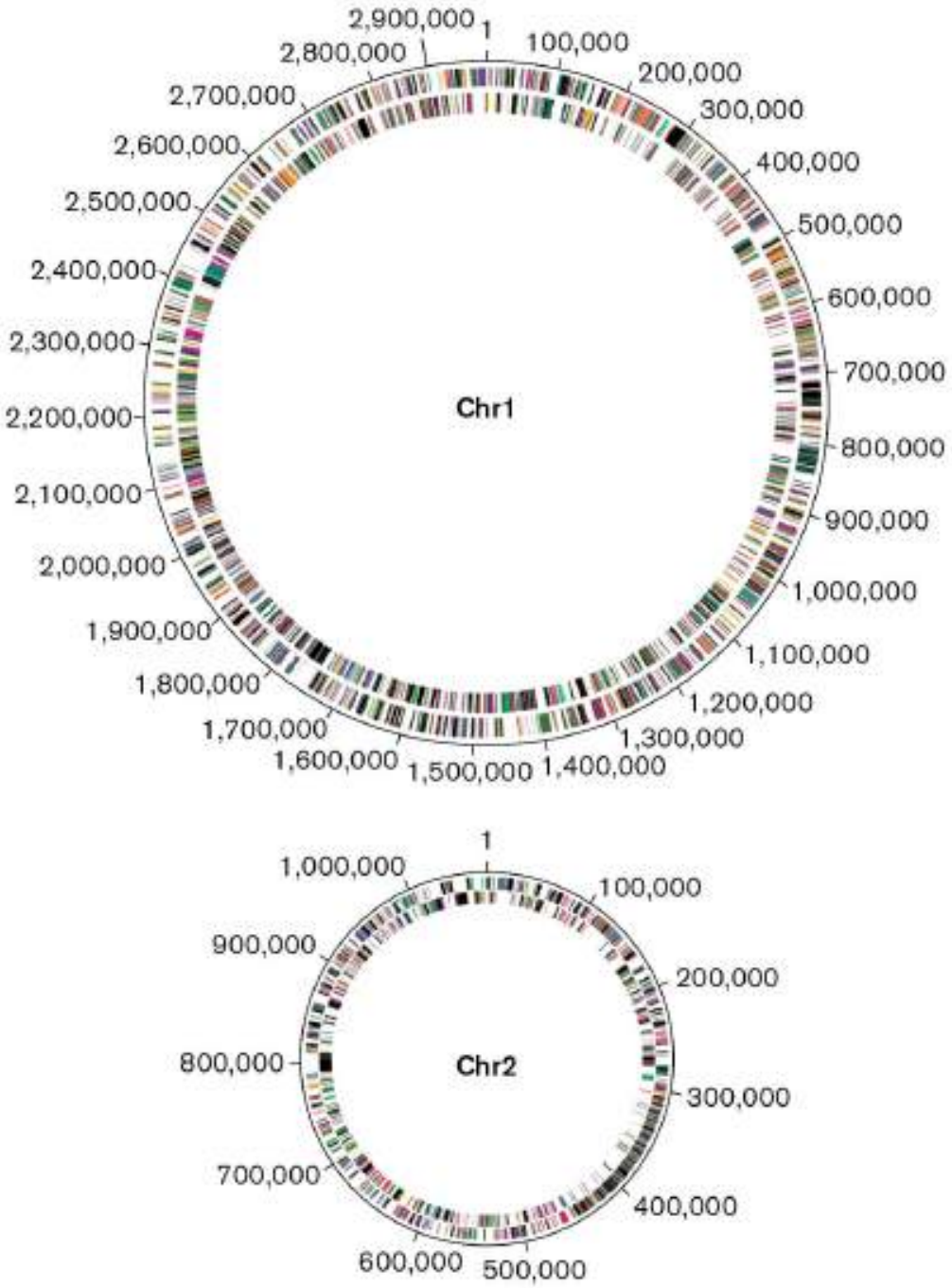


FIGURE 7.2 ■ The two chromosomes of the *Vibrio cholerae* genome. In each chromosome, the numbers refer to the position in base pairs, starting at the origin of replication and extending clockwise. Genes encoded on the plus and minus strands of the chromosome are depicted on

the outer and inner rings, respectively. Chr1 = primary chromosome; Chr2 = secondary chromosome.

SOURCE : MODIFIED FROM J. F. HEIDELBERG ET AL. 2000. *NATURE* **406** :477–483, FIG. 2.

7.5 Eukaryotic and Archaeal Chromosomes

Archaea, Eukarya, and Bacteria represent the three domains of all life (as introduced in Chapter 1). The special features of Archaea and those of eukaryotic microbes are described in Chapters 19 and 20 , respectively. Here in this section, we focus on features of DNA organization and replication that distinguish each of these domains.

The chromosomes of eukaryotic microbes such as protists and algae have much in common with those of bacteria and archaea. All chromosomes consist of double-stranded DNA, for example, and are usually replicated bidirectionally. Nevertheless, important differences exist as well, particularly in genome structure. Eukaryotic chromosomes are linear and contained within a nucleus, and their transmission involves mitosis (reviewed in eAppendix 2). Archaeal chromosomes, similar to bacterial chromosomes, are circular and archaea lack a nuclear membrane, but archaeal proteins involved in transcription and replication appear more eukaryotic than bacterial.

Eukaryotic Genomes Are Large and Linear

Overall, the genomes of eukaryotic nuclei are larger than those of bacteria, sometimes by several orders of magnitude. Eukaryotes typically have duplicate copies of multiple chromosomes. Almost all eukaryotic chromosomes are linear, whereas many, if not most, bacterial chromosomes are circular. Eukaryotes use mitosis to segregate replicated chromosomes to daughter cells. Each eukaryotic chromosome has numerous origins of replication that collectively generate hundreds of replication bubbles, although not all of them "fire" during each replication cycle. Termination zones essential to bacterial chromosomes are not found in eukaryotes.

The ends of eukaryotic chromosomes, called telomeres , have a special problem replicating. The problem is that the normal DNA

replication machinery cannot fully replicate the lagging strands of linear DNA. Recall from Section 7.3 that DNA replication on the lagging strand, which proceeds in a 5′-to-3′ direction, requires an RNA primer to initiate it. At the end of the linear chromosome, an RNA primer could be placed at the 5′ extremity to initiate lagging-strand synthesis. Once that RNA primer was removed, the standard replication machinery would be unable to replace it with DNA. The result would be a replicated chromosome with shortened 5′ ends (and 3′ overhangs). Each round of replication would shorten the chromosome a little more, until the loss of genetic information became catastrophic to the organism.

Eukaryotic microbes (and the germ cells of metazoa like humans) solve the problem of chromosomal shortening during replication with an enzyme complex called telomerase . Telomerase is actually a reverse transcriptase that reads RNA as a template to synthesize DNA (see Sections 6.3 and 11.3 for the role of reverse transcriptase in the replication of retroviruses such as HIV). Telomerase uses an intrinsic RNA (an RNA that is part of the enzyme) as a template to add numerous DNA repeat sequences to the 3′ ends of chromosomes. The purpose of this 3′ extension is to provide a new "upstream" location for the RNA primer to initiate lagging-strand synthesis. This RNA primer could then be used to replicate the 5′ end of the chromosome. In this situation, the only template DNA that is left unreplicated is the DNA placed by the telomerase. Telomere length is thus dynamic in replicating eukaryotic cells, with shrinkage due to RNA priming being countered by the synthesis activity of telomerases.

Telomerase may have evolved from the ancient progenitor cells that contained RNA rather than DNA genomes (see Section 17.2). The reverse transcriptase may also be the evolutionary source of retroviruses (see Section 6.5).

Note: Bacteria with linear chromosomes do not have telomerases. How do they replicate their 3′ ends? Some (for example, *Streptomyces*) cap the ends of their chromosomes with covalently bound terminal proteins that prime DNA replication. Others (for

example, *Borrelia*) form covalently closed hairpin ends (essentially making one long, circular, single-stranded molecule). The ends appear to recombine after being replicated.

Unlike bacterial cells, eukaryotic cells pack their DNA within the confines of a nucleus where a series of proteins called histones compacts the DNA. Histones are rich in arginine and lysine, so they are positively charged, basic proteins that easily bind to the negatively charged DNA. The DNA becomes wrapped around the histones to form units called nucleosomes. Histones also play a regulatory role through methylation and acetylation. Bacteria, too, have DNA-packaging proteins, but they are less essential for function (see Chapter 3 and Section 7.2).

A feature that distinguishes bacterial (and archaeal) genomes from those of eukaryotes is the amount of so-called noncoding DNA (DNA that does not encode proteins). Bacteria and archaea tend to have very little noncoding DNA, typically less than 15% of the genome (**Fig. 7.26A**). Noncoding DNA includes an important class of RNAs, such as ribosomal RNA (rRNA) and transfer RNA (tRNA), that is involved in translation of RNA messages into protein (see Chapter 8). Other instances of noncoding DNA include sequences within mobile genetic elements such as transposons and insertion sequences that can move from one DNA molecule to another (**Fig. 7.26A** , and described further in Section 9.4).

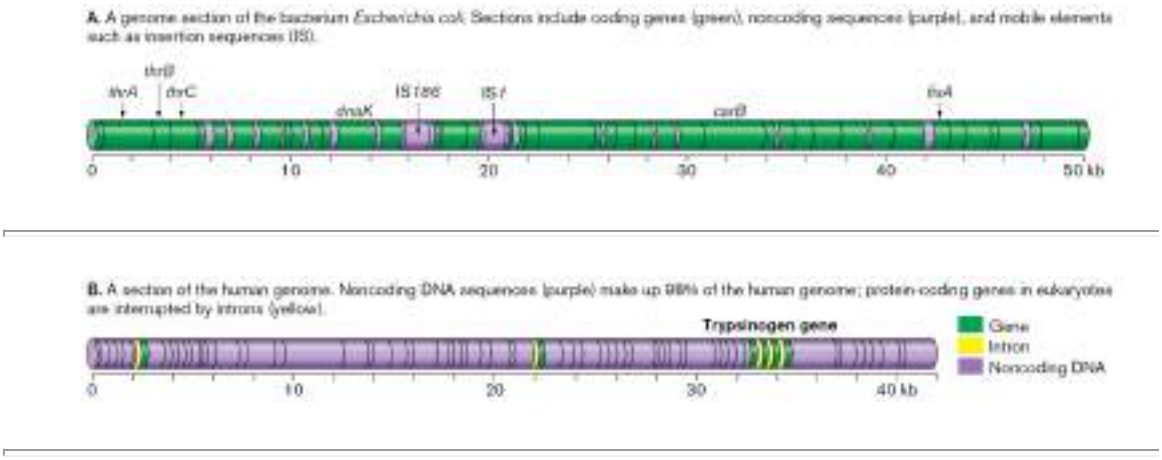


FIGURE 7.26 ■ Genome structure in a bacterium (A) and a eukaryote (B).

In contrast to bacteria and archaea, many eukaryotes contain huge amounts of noncoding DNA that can separate genes by thousands of base pairs (**Fig. 7.26B**). In some species (such as humans), over 90% of the total DNA is noncoding. Noncoding DNA includes sequences that regulate protein-coding genes; genes that encode untranslated, small regulatory RNAs; and the fossil genomes of ancient viruses.

Some noncoding regions include enhancer sequences that initiate transcription of eukaryotic genes (see Section 10.1). Enhancer sequences can function at large distances from the promoters of genes they regulate, and scientists suspect that once enhancers became important, it was necessary to place enough DNA between them to reduce the activation of other, unrelated but adjacent promoters.

Moreover, coding genes in the human genome are interrupted by introns (DNA within a gene that is not part of the coding sequence for a protein, shown as yellow in **Fig. 7.26B**) and ancient gene duplications that have decayed into nonfunctional, vestigial pseudogenes . Bacteria also have pseudogenes, but they are relatively rare, except in intracellular symbionts and pathogens, where gene functions are readily lost because the host cell provides the necessary resources.

Note: Pseudogenes differ from noncoding DNA because at least part of a pseudogene's sequence is similar to that of a gene with a known function. Also note that some pseudogenes express mRNA, but the protein products are typically truncated and nonfunctional.

Archaeal Genomes Combine Features of Bacteria and Eukaryotes

Like bacteria, archaea are haploid and reproduce asexually via cell fission. Archaea are true prokaryotes because their cells lack a nuclear membrane, but the structures of their DNA-packing proteins, RNA polymerase, and ribosomal components more closely resemble those of eukaryotes. How are archaeal genomes organized and replicated?

Archaeal chromosome structure and replication share features with bacteria and eukaryotes. Archaea use several proteins to organize their chromosomes in the nucleoid. Many archaea use Alba proteins, unique to the archaeal domain, to form loops of DNA, and some also employ histones to further organize the chromosome. Like bacteria, most archaea have a single chromosome. Thus far, all known archaeal chromosomes are circular, and replication proceeds bidirectionally from the origin. However, most archaea have multiple origins of replication distributed around the chromosome, as shown in **Figure 7.27** for *Sulfolobus solfataricus* .

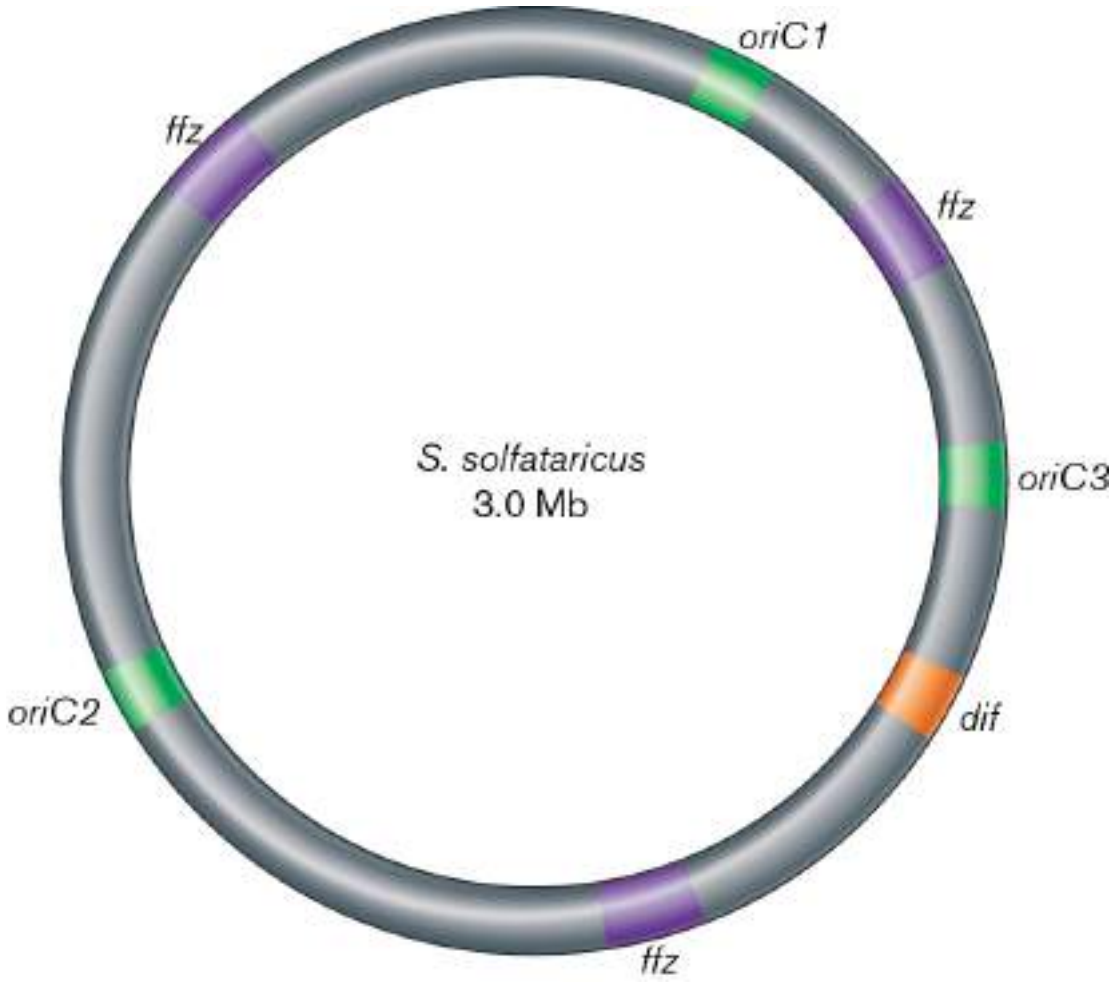


FIGURE 7.27 ■ *Sulfolobus solfataricus* has three origins of replication. Bidirectional replication initiates at three origins of replication (*oriC1* – *oriC3*) and terminates at the fork fusion zones (*ffz*) located between each pair of origins. Chromosome dimers are resolved at a single *dif* site, as in bacteria.

I. G. DUGGIN ET AL. 2011. *EMBO J* . **30** :145–153.

Each origin has its own specific initiator protein, a relative of the eukaryotic ORC (origin recognition complex) protein, sometimes called Cdc6. And even more intriguing, replication initiation at each origin is essentially synchronous: Triggered by an unknown mechanism, all initiators are able to unwind the DNA at their respective origins and recruit DNA helicase and the remainder of the DNA polymerase machinery. *S. solfataricus* lacks defined *ter*

sequences, but replication terminates at locations between each pair of origins called replication fork fusion zones (*ffz*), where the replisomes moving toward each other collide (**Fig. 7.27**). As with bacteria, if a chromosome dimer forms via homologous recombination, it is resolved by Xer proteins at a single *dif* site.

The advantage of possessing multiple functional origins of replication is currently unknown, as genetic mutants with all but one origin removed have minimal to zero changes in growth rate. Notably, the additional origins seem to have been acquired by horizontal gene transfer, rather than through simple duplication of preexisting origin(s) on the chromosome.

Note: The archaeal species *Haloferax volcanii* appears to initiate replication via homologous recombination, an unusual mechanism that may be similar to the initiation of bacteriophage T4 and the non-nuclear chromosomes of eukaryotes: mitochondria, chloroplasts, and kinetoplastids.

Archaea can possess one or two types of DNA polymerases: one that is related to the eukaryotic PolB enzyme, and another, PolD, that appears to be unique to the archaeal domain. Whether both participate in leading- and lagging-strand replication or have alternative roles in the cell is an open question, and the answer may be different depending on the archaeal lineage. The remaining components of the initiation and elongation machinery of archaea, including the helicase, primase, sliding clamp, and DNA ligase, are related to the eukaryotic proteins, rather than to the bacterial versions. Mechanisms of replication termination in archaea with multiple origins are still not well understood, but replication may be terminated as a consequence of replication fork collision. Chromosome dimers appear to be resolved by Xer proteins acting at *dif* sites, as in bacteria.

To Summarize

- **Eukaryotic chromosomes** are linear, double-stranded DNA molecules. After replication, the copies are segregated to daughter cells by mitosis.
- **A reverse transcriptase** called telomerase prevents net loss of DNA at the ends of eukaryotic chromosomes during replication.
- **Histones** (DNA-packing proteins) play a critical role in compacting chromosomes in eukaryotes and some archaea.
- **Noncoding DNA** can constitute a large amount of a eukaryotic genome, while prokaryotes have very little noncoding DNA.
- **Introns and pseudogenes** are noncoding DNA sequences that make up a large portion of eukaryotic chromosomes.
- **Archaeal chromosomes** resemble those of bacteria in size and shape, but archaeal DNA replication machinery is more closely related to eukaryotic enzymes. Archaeal chromosomes often contain multiple origins of replication.

Glossary

telomere
: The DNA segment at either end of a eukaryotic chromosome.

telomerase
: A reverse transcriptase enzyme complex that reads RNA as a template to synthesize DNA.

histone
: A protein that binds eukaryotic DNA and compacts chromosomes in nucleosomes.

enhancer
: A noncoding DNA regulatory region in eukaryotes that can lead to activation of transcription when bound by an appropriate

transcription factor. Its location on the chromosome can be far removed from the regulated gene.

promoter

A noncoding DNA regulatory region immediately upstream of a structural gene that is needed for transcription initiation.

intron

In eukaryotic genes, an intervening sequence that does not code for protein and is spliced out of the mRNA prior to translation.

pseudogene

A nonfunctional gene-like sequence that evolved by degenerative evolution.

7.6 Microbiomes and Metagenomes

Microorganisms in nature do not generally exist as pure cultures. They grow instead as complex consortia containing numerous species, the sum total of which is called the microbiome (or microbiota). Most members of these microbiomes have never been grown in a laboratory and thus cannot be investigated by traditional culture-based methods. However, we can use molecular-based methodologies such as high-throughput DNA sequencing and the polymerase chain reaction (PCR; see eAppendix 3) to investigate the diversity of microbiomes without having to culture them. Like the explorers of old, scientists are finding new species never before seen and are beginning to understand the intricate ways in which bacterial species interact. The study of microbial communities is explored at length in Chapter 21 .

The molecular investigation of the microbiome began with the work of Norman Pace and colleagues at Indiana University, who reasoned that while most microbes may not be readily cultured, their DNA could be isolated directly from an environmental sample and sequenced for analysis. These investigators chose the 16S rRNA gene for sequencing because it is highly conserved in microbes, and phylogenic studies of 16S sequences could be used to determine whether members of the community belonged to known lineages or represented novel microorganisms (see Section 17.3). Since then, environmental sequencing has revolutionized microbiology by shifting the focus even further away from cultivatable organisms toward the estimated 99% of microbial species that cannot be cultivated (discussed in Chapter 21). Originally used to investigate exotic locations like the hot springs of Yellowstone National Park in the US, PCR-based environmental sequencing has been exploited to reveal the microbiomes present in a vast number of environments, including the human gut.

One recent molecular study revealed that human genes can influence the makeup of our gut microbiome. Approximately 100

trillion microbes live in the human body, but we do not know the identity of most of them, because we cannot grow them in the laboratory. How we interact with our microorganisms and they with us is a major question in modern microbiology—one that has led to formation of the Human Microbiome Project, which examines gut, skin, and oral microbiomes. We already know that the composition of intestinal microbiota can influence blood chemistry and may contribute to irritable bowel syndrome (inflammation of the gastrointestinal tract), but microbiota are also suspected of influencing an individual's susceptibility to diabetes and obesity (see Section 23.2).



Ruth E. Ley (Cornell University; **Fig. 7.28A**) and collaborators cataloged the intestinal microbiomes of identical and fraternal human twins to determine whether any part of the microbiome might be heritable; that is, influenced by the genetic makeup of the human host. They used PCR to amplify 16S rRNA sequences from human fecal samples and sequenced the products. The 16S rRNA sequences were matched to a public database to identify bacterial species. Host genotype had the largest influence on *Christensenella minuta* , a recently discovered firmicute. Even more striking, this microbe was found more often in lean than in obese subjects, and it prevented weight gain when transplanted into germ-free mice (**Fig. 7.28B**). Knowing the identities of these microbes and their collective metabolic potential will undoubtedly lead to new advances in the treatment and prevention of human diseases.



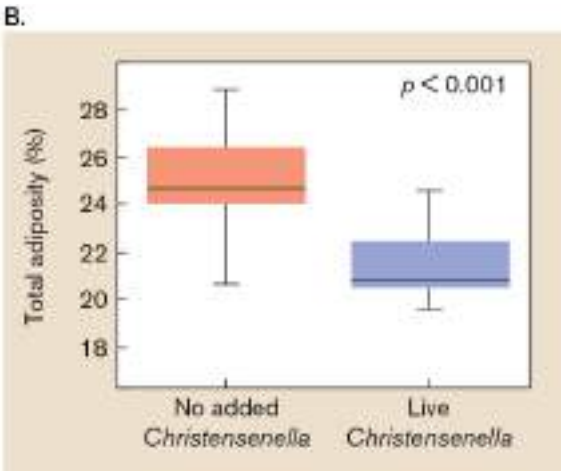


FIGURE 7.28 ■ Heritable member of the human microbiome. A. Ruth Ley led a team that used metagenomics to discover that the abundance of some members of the intestinal microbiome are influenced by host genetics. **B.** One such organism, *Christensenella minuta* , also influenced weight gain and adiposity when orally "transplanted" into mouse intestines. The graph shows that adiposity 21 days after a fecal transplant was significantly lower in germ-free mice transplanted with *C. minuta* –containing fecal preparations (blue) than in germ-free mice transplanted with preparations lacking this organism (orange).

Source: J. K. Goodrich et al. 2014. *Cell* **159** :789–799, fig. 7B.

CORNELL BRAND COMMUNICATIONS

Analysis of an environmental sample using 16S rRNA sequences can provide valuable information on the composition of the microbiome but cannot reliably predict the activity of the community members. This is because, while the 16S rRNA gene is highly conserved, other genes in the genome can be lost and gained though horizontal gene transfer (see Chapters 9 and 17). To gain a better understanding of the metabolic and physiological potential of a microbe, it is thus imperative to sequence every gene in every genome, even if we do not know which genes belong to which species. The result is called the metagenome . The objective of metagenomics , a term coined in 1998 by microbiologist Jo Handelsman (Yale University), is thus not only to identify the members of a microbial community, but also to reveal the potential activities and contributions to the ecosystem of these members. Metagenomic studies have radically accelerated with the development of high-throughput DNA sequencing technology (see Chapter 21). Critically, the sheer scale of information in metagenomic studies can also identify patterns in microbial community function that would have been otherwise difficult or impossible to see in smaller data sets, as the next example will demonstrate.

In a comprehensive metagenomic study, Shinichi Sunagawa and colleagues sequenced 7.2 terabases (7.2×10^{12} bases) of data from 68 locations spanning the world's ocean (**Fig. 7.29** . The researchers first examined all of the 16S rRNA genes within the metagenome to determine how many microbes were in the microbiome. The level of sequence information could not resolve species versus genus classification, so they used a conservative term for each genotype called an operational taxonomic unit (OTU) . The researchers identified 37,000 OTUs in their collection of ocean samples. When the individual OTUs were grouped into higher taxonomic ranks, such as the phylum Cyanobacteria, these taxa were found to contribute unequally to total taxonomic composition, and importantly, their contribution varied depending on the sampling location (**Fig. 7.29B**).



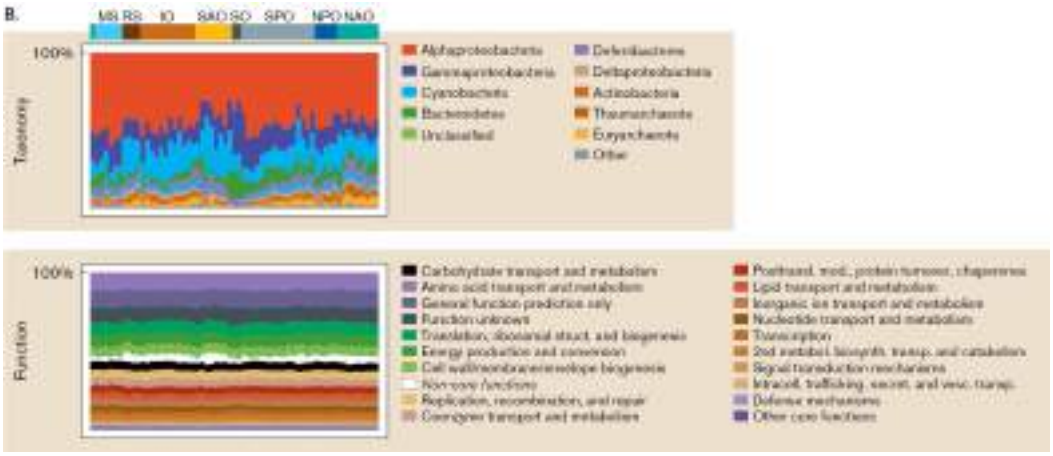


FIGURE 7.29 ■ Metagenome of the ocean's surface. A. The sampling locations in Shinichi Sunagawa's study. **B.** Fractional composition of taxa and functional gene categories for each sample (each column is one sample). Sampling locations are indicated by the color bar above the plots; refer to panel (A) for general location. *Source:* Modified from Shinichi Sunagawa et al. 2015. *Science* **348** :6237, figs. 1A (part A) and 8A (part B).

After identifying the microbiome members, Sunagawa and colleagues analyzed the rest of the metagenome. They found that a large fraction of the metagenome, 40%, consists of genes with no known function. That unknown 40% could be involved in processes critical for global nutrient cycling and might include genes that can synthesize products of human interest, such as novel therapeutics and antibiotics. The remaining 60% of the metagenome consists of genes whose functions could be predicted by their sequence, and these could be categorized into gene families on the basis of predicted function. Over 70% of these gene families were found at every station and were classified as "core." The researchers noted that the distributions of these core gene families did not vary much at all between stations, in sharp contrast to the OTUs (**Fig. 7.29B**). This surprising finding suggests that certain core functions are important for the community, but which microbes perform those functions can vary from location to location. Studies like these have led microbial ecologists to consider that communities assemble as

collections of gene functions, rather than as collections of taxonomic groups.

The Sunagawa study also compared the metagenomes from the ocean ecosystem to that of the human gut (**Fig. 7.30**). The researchers found that the ocean microbiome had a higher fraction of genes for photosynthesis and the uptake of amino acids, lipids, nucleotides, and secondary metabolites, while the gut microbiome had a higher fraction of genes for defense, signal transduction, and carbohydrate transport. Thus, at the community level the microbes behave quite differently in the human gut than they do in the ocean. Comparisons such as these provide new hypotheses and research directions for how microbial communities and functions are dictated by their environment and also provide clues as to which environmental factors may be most critical.

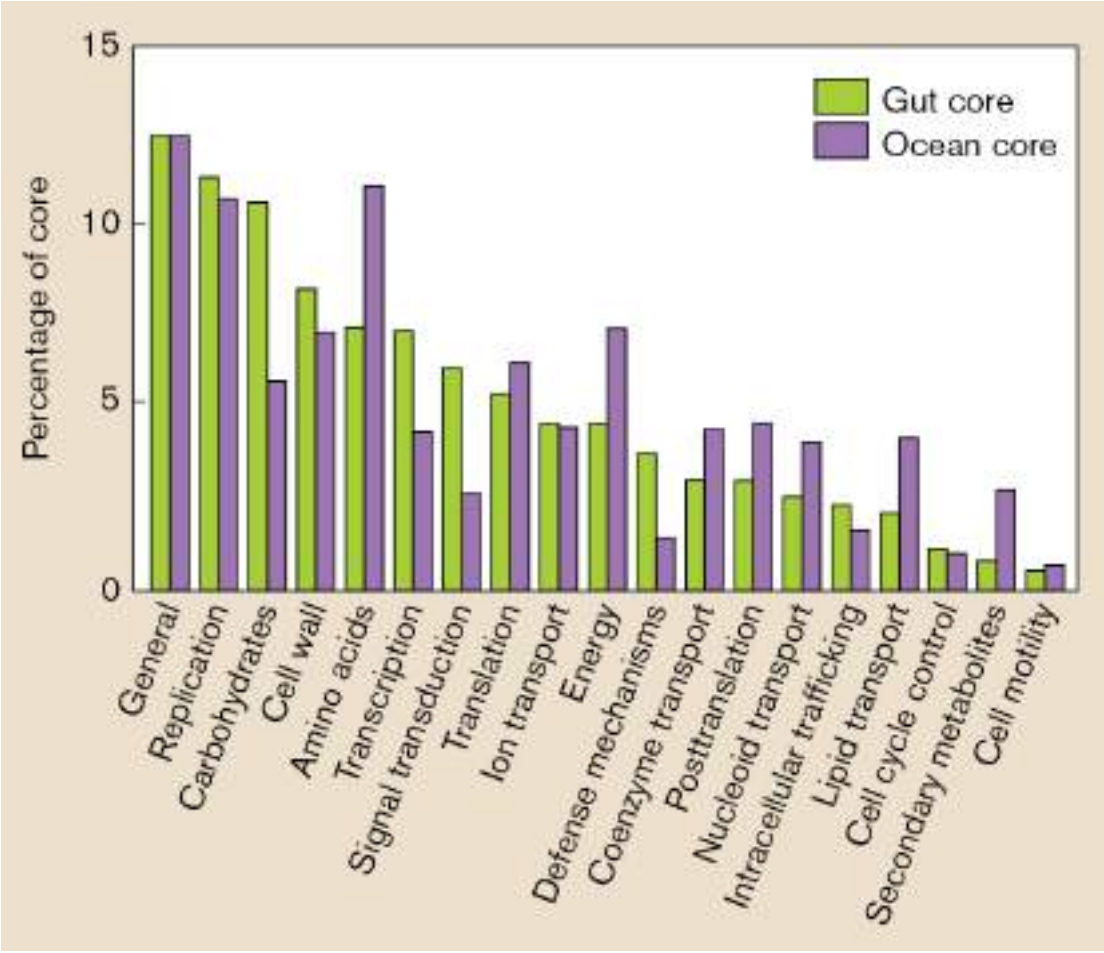


FIGURE 7.30 ■ Comparison of the core set of genes found in the ocean and human gut microbiomes, arranged by functional category.

SOURCE: MODIFIED FROM SHINICHI SUNAGAWA ET AL. 2015. *SCIENCE* **348** :6237, FIG. 7C.

Thought Question

7.10 How might you interpret the discovery of genes for photosynthesis in the metagenome of the human gut? Could it indicate a possible error in the analysis or could something else be going on?

Sequencing and Assembly of Metagenomes from Microbiomes

How are metagenomes such as the ones just described produced from microbiomes? Metagenome sequencing poses challenges far beyond those of sequencing a single intact genome. To sequence a metagenome requires a series of steps, each of which presents important choices (**Fig. 7.31**).

Sampling the target community. The first decision is to define a target community from which to obtain DNA. The cells of the target community must be separated from their surroundings without loss of DNA (**Fig. 7.31** , step 1). For example, sampling a soil community requires removal of humic acids (wood breakdown products), which inhibit DNA polymerases. Suppose the target community inhabits a host plant or animal; what additional separation is required? Before DNA extraction, we need to dislodge host-associated microbes from

their host. Otherwise, the host DNA could contaminate the microbial DNA pool.

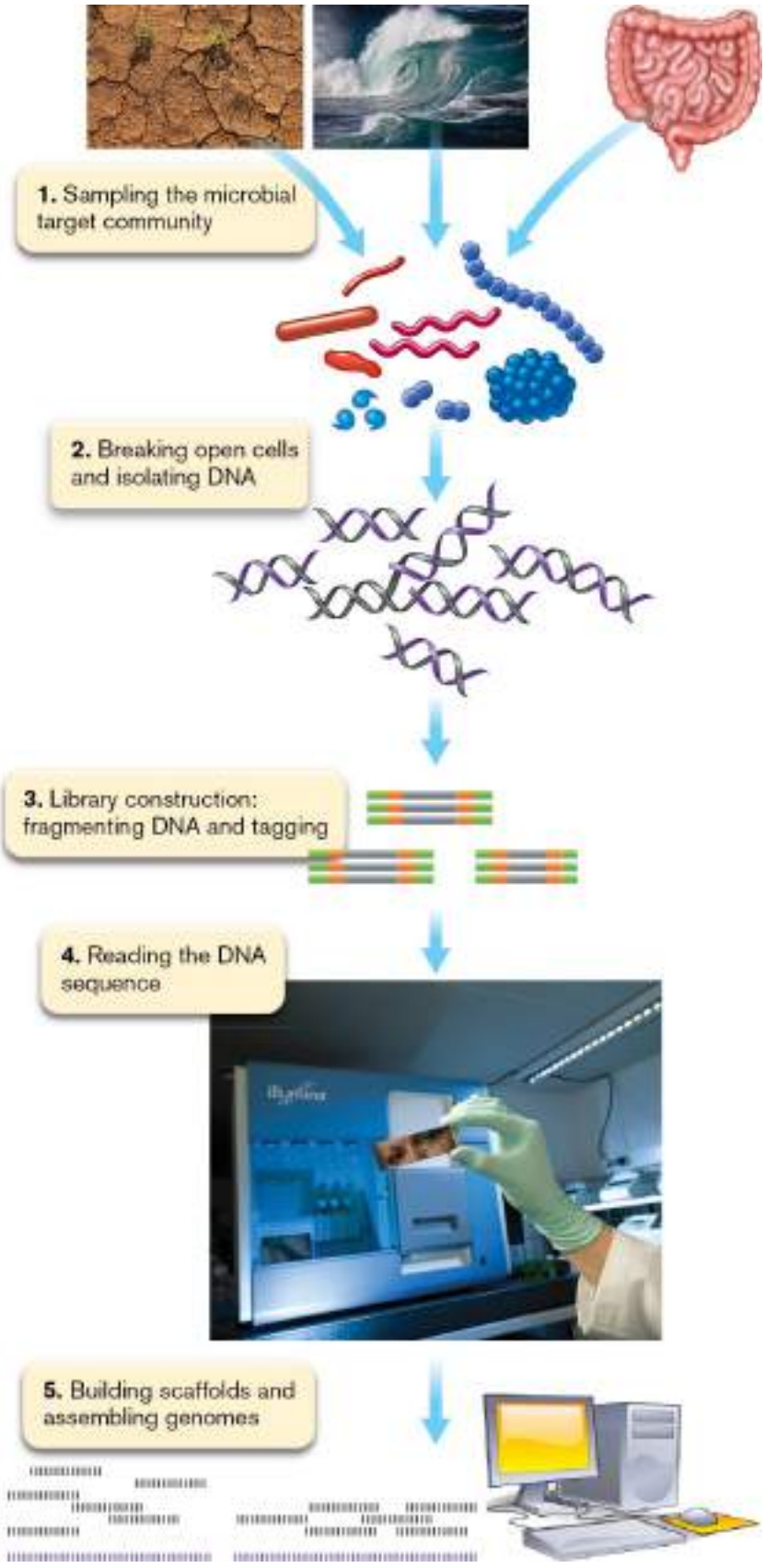

1. Sampling the microbial target community
2. Breaking open cells and isolating DNA
3. Library construction: fragmenting DNA and tagging
4. Reading the DNA sequence
5. Building scaffolds and assembling genomes

FIGURE 7.31 ■ Sequencing a metagenome. First, select a target community to sample from a habitat such as soil, water, or host plant or animal. Remove the sampled microbes from their environment (step 1) and stabilize the content. Lyse the cells and isolate pure, intact DNA (step 2). Amplify the DNA library by constructing fragments with tagged ends (step 3). Read the DNA sequence using a next-generation sequencing (NGS) sequencer (step 4). Use a computational software pipeline to build scaffolds and assemble genomes (step 5). Source: John Wooley et al. 2010. *PLoS Comput. Biol.* 6:e1000667.

FLPA/SHUTTERSTOCK
PHOTOBRET 2014/SHUTTERSTOCK
IMAGE BROKER/ALAMY STOCK PHOTO

Isolating DNA. Once separated from the physical habitat, the cells of the target community must be opened in such a way that all of the DNA is released with minimal breakage of the strands (**Fig. 7.31** , step 2). We can lyse the cells by "bead beating" or by sonication (methods discussed in Chapter 3). The DNA can then be purified by phenol extraction and precipitation with ethanol or by binding to special filters. But—unlike the analysis of a single-species genome—analysis of a metagenome requires accounting for different species that possess different kinds of enzyme inhibitors, as well as envelope, sheath, and S-layers of diverse composition.

How can we ensure that our protocol will be optimal for all the thousands of species in the community? In fact, no single best way exists to extract metagenomic DNA. Different researchers argue for one of two main approaches:

- **Use a single, universally applied method of DNA extraction for all target communities.** If all research groups sample metagenomes using a common DNA extraction method, then results may be compared across all projects.
- **Use multiple DNA extraction methods for each target.** If a research group uses multiple DNA extraction methods to sample

one target community, then the group has the best chance to maximize coverage of all the microbial genomes in the sample.

Thought Question

7.11 Suppose you are conducting a metagenomic analysis of soil sampled from different parts of a wetland. Would you use one DNA extraction method or multiple methods?

Preliminary screen for diversity. Before investing in large-scale DNA sequencing, we can perform a preliminary screen for sequence diversity. The preliminary screen may answer questions such as: Because species in the community may differ in their fractional representation by several orders of magnitude, how important are the community's rarest members? Is our aim a complete description of the target community, such as the microbiota inhabiting the human stomach? Or do we focus on a narrower goal, such as identifying soil actinomycete genes that produce novel antibiotics?

A preliminary screen of the sample involves sequencing the small-subunit rRNA (SSU rRNA) genes (discussed in Chapter 17). The genes specifying 16S rRNA (bacteria and archaea) or 18S rRNA (eukaryotes) may be amplified by PCR using universal primers known to detect a wide range of taxa. The classical approach of cloning has largely been replaced by amplification of the rRNA genes from Illumina next-generation sequencing of DNA (see eAppendix 3). This use of Illumina to obtain SSU rRNA sequences from a metagenome is called iTAG analysis or metabarcoding (as compared with barcoding, the SSU rRNA identification of single taxa). Compared to earlier methods, Illumina sequences provide a larger number of genes at lower cost.

Metabarcoding quickly identifies many taxa present in a community, including relatively rare members that do not yield full

genomes when large-scale sequencing is performed. SSU rRNA gene similarity is used to define operational taxonomic units (OTUs), a working metagenomic definition of genetically distinct organisms, in some cases equated to "species." SSU rRNA gene sequencing reveals the general categories of microbes found in a community, their relative abundance in a community, and the overall diversity of the sample.

Library construction for next-generation sequencing. Samples, such as a human tissue biopsy or marine water, will contain DNA in such small quantities that it must be amplified by a form of PCR and processed for the next-generation sequencing (NGS) reactions. The amplification and processing are called library construction (**Fig. 7.31** , step 3). The library construction depends on several factors:

- **DNA concentration.** Lower concentration requires greater amplification. However, amplification can introduce sequence errors and loss of sequences that fail to be amplified.
- **Fragmentation for size.** The DNA must be fragmented to the size required by NGS, typically the Illumina platform. Small fragment sizes may increase errors in alignment, but larger sizes may involve other kinds of errors.
- **Addition of adapters.** Specific adapter sequences called tags are added at the ends of the fragments to allow the sequencing reaction to proceed.

Metagenome sequencing and assembly. The most common type of NGS metagenome sequencing is Illumina sequencing by synthesis (**Fig. 7.31** , step 4). The process of Illumina sequencing is described in eAppendix 3 . Illumina sequencing generates millions of short DNA sequences of defined length; these short sequences are called reads . To sequence the genome requires assembly of reads whose sequences of base pairs overlap. The overlapping reads are assembled into contigs , which are regions of contiguous DNA sequence without gaps (**Fig. 7.32**). Inevitably, though, gaps remain between contigs.

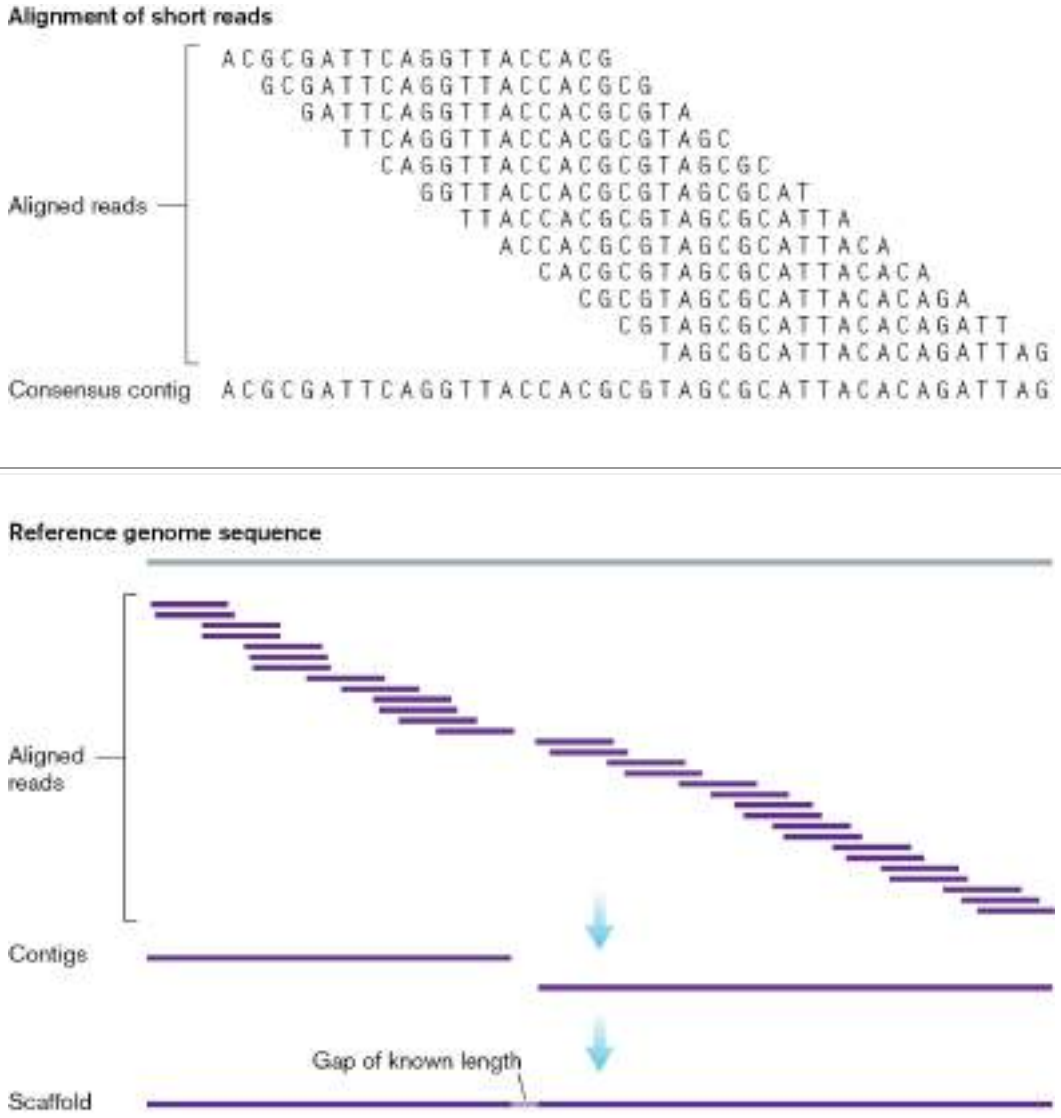


FIGURE 7.32 ■ Assembly of reads into contigs and scaffolds. Overlapping reads generate a contig. Contigs matched to a reference genome generate a scaffold. Scaffolds may still contain gaps of unknown sequence.

If the contigs show high relatedness to the genome of a known organism, this reference genome may be used to match contigs together in scaffolds containing gaps of presumed length. In effect, the assembly of contigs resembles a jigsaw puzzle with thousands of pieces. But a metagenome requires assembly of thousands of jigsaw

puzzles, whose pieces are all jumbled together. In practice, metagenomes rarely yield complete genomes of individual species, but they can yield partial genomes of previously unknown organisms with interesting properties. These partial genomes are called metagenome-assembled genomes (MAGs). A metagenome-assembled genome may define an OTU with greater resolution than those defined by a single SSU rRNA sequence.

Genome assembly requires a computational pipeline , a linear series of programs that combine mathematical tools with biological assumptions to propose assembled genomes. Mathematical tools of assembly are based on a formula known as the de Bruijn graph (**Fig. 7.33**). The de Bruijn graph compiles overlaps between sequences of base pairs in a manner that predicts the identity of the original intact sequence. The short sequences are called k-mers, where "k" refers to a defined length. The chosen k-value is generally shorter than the Illumina read length, such that sequence errors are minimized. The k-mers are assumed to start from all possible positions in the overall sequence; thus, they overlap neighboring k-mers, in all possible ways. The computational algorithm links all k-mers by their overlapping ends. Repeated k-mers cannot be distinguished, so the algorithm collapses them with overlapping connections, generating a diagram of possible sequence connections with loops as shown. An actual de Bruijn graph from a set of reads may have numerous loops of this type.

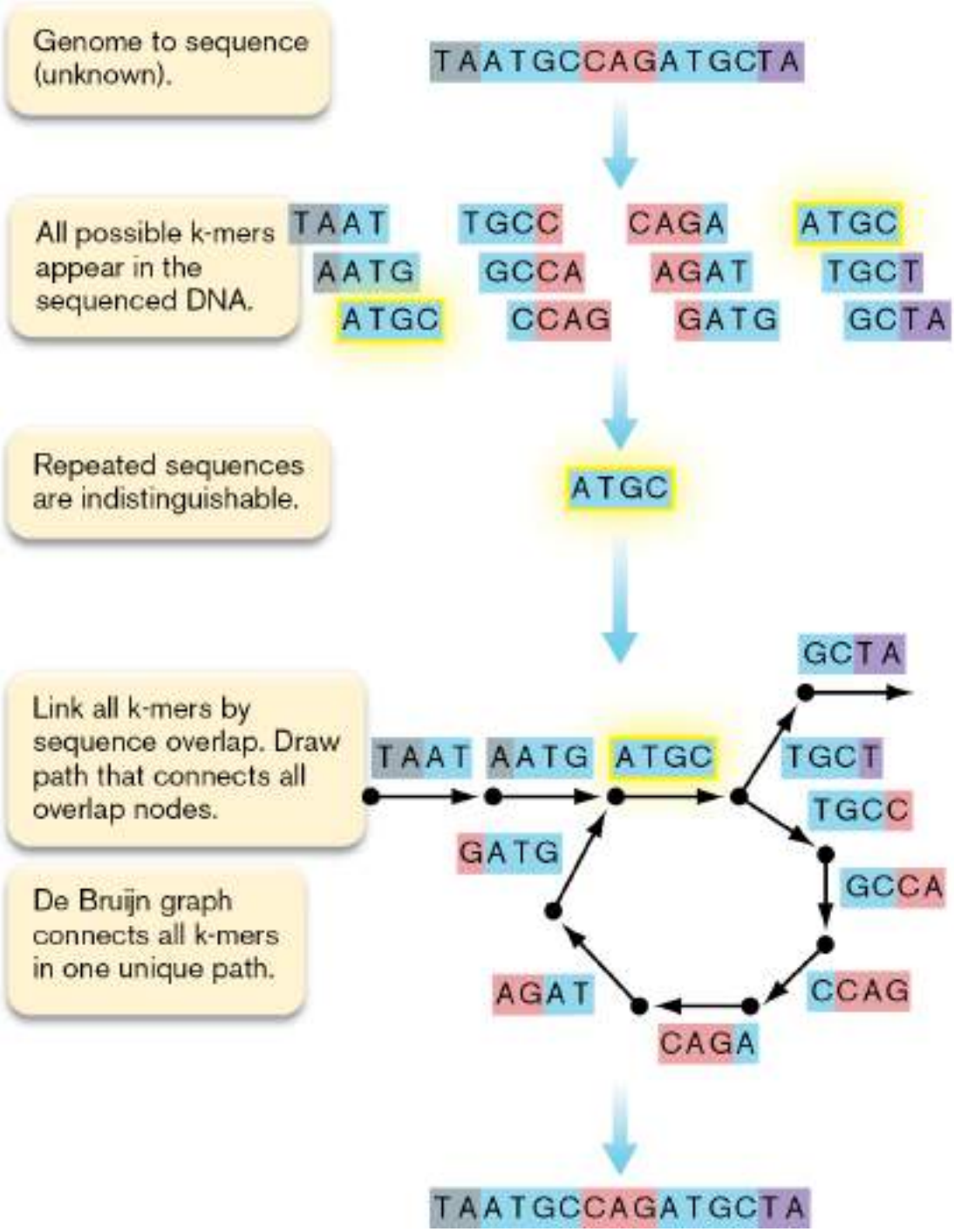


FIGURE 7.33 ■ Assembling reads on the basis of de Bruijn graph computation. A given DNA sequence generates many short fragments of defined length, called k-mers. All k-mers found are linked by their overlapping ends. Repeated k-mers cannot be distinguished, so the de Bruijn graph collapses them with overlapping connections. To reveal the sequence that

produced the fragments, find a path that passes through every k-mer exactly once. That path generates the original sequence of base pairs.

We can predict the sequence that produced the fragments by taking a path through the de Bruijn graph that passes through every k-mer exactly once. This path generates the original sequence of base pairs. In principle, the k-mer size is selected such that most reads have one unique position in the genome, and in **Figure 7.33** , computation based on the de Bruijn graph predicts the exact sequence of the original DNA. In practice, of course, there are ambiguities and "unfinished" regions caused by errors in sequence, and by multicopy repeated regions that are larger than the k-mer size.

Genome assembly can be of two types: mapping to a reference or de novo assembly. Mapping to a reference, or "resequencing," involves aligning the contigs and scaffolds with known reference genomes. Reference genomes can be effective at "pulling out" genomes of organisms present in low abundance. Without a reference, de novo assembly takes much longer, although it offers greater possibility of identifying a previously unknown type of organism.

Single-Cell Genomics Can Reveal Capabilities of Uncultured Cells

The technology of DNA sequencing has progressed to a point where the genome of a single cell can be sequenced—enabling single-cell genomics . Single-cell genomics (SCG) has several advantages over shotgun sequencing–based metagenomics. First, it requires less sample DNA—only that found within a single cell—which may be important for study sites where microbial abundance is low. Second, because all reads come from one cell type, the assignment of the reads to a single taxon is more reliable. Hence, SCG has the power to identify more complex metabolic and physiological features of cells

that may involve large suites of interacting genes. SCG requires special protocols, equipment, and reagents to isolate a single cell from an environment and, without ever growing it, extract and amplify its DNA. Once amplified, the genomic DNA is readily sequenced using the next-generation sequencing tools described in eAppendix 3 .

Using SCG, Karen Lloyd (**Fig. 7.34A**) and colleagues at the University of Tennessee, Aarhus University, and other institutions reconstructed partial genomes of archaea that are abundant in marine sediments but have thus far resisted all attempts at cultivation. Among the several metabolic pathways they were able to assemble, they discovered that these organisms have the potential for breakdown of extracellular proteins and peptides, an activity not previously recognized in sediment archaea (**Fig. 7.34B**). Bioinformatic analysis suggested that the cells secrete several classes of proteases into the environment. Consistent with this genomic evidence, sediments where these organisms were collected tested positive for the activity of these specific classes of proteases. Bioinformatics also identified candidate transporters for oligopeptide uptake and intracellular proteases that could, in concert with the secreted proteases, completely break down environmental proteins into amino acid monomers for use in both biosynthesis of new proteins and heterotrophic energy metabolism.

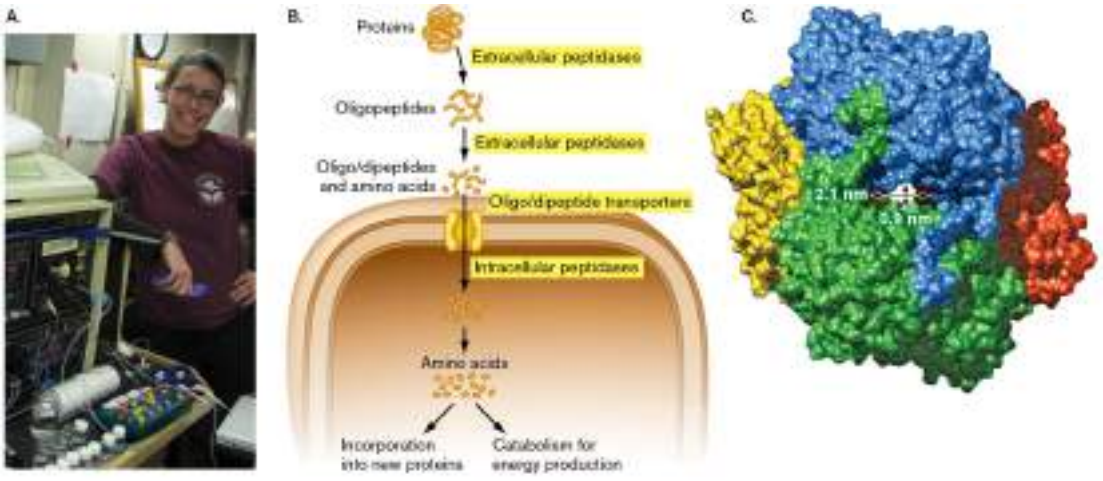


FIGURE 7.34 ■ Single-cell genomics identifies microbial functions in marine sediments. Karen Lloyd **(A)** and

colleagues used SCG **(B)** to identify enzymes of an uncultured archaeon that degrades, imports, and processes extracellular proteins scavenged from marine sediments. *E. coli* was used as a surrogate host to produce proteins from this uncultured archaeon so that the proteins could be studied in the lab. **C** . One intracellular protein, S15, whose native state was a tetramer (each color is a monomeric subunit), was confirmed to have peptidase activity. S15 was the first protein of experimentally validated structure and function characterized for this uncultured microbe. (PDB code: 2B9V)

Source: Part B modified from K. G. Lloyd et al. 2013. *Nature* **496** :7444, fig. 3.

COURTESY OF KAREN LLOYD
KAROLINA MICHALSKA ET AL. 2015. *FASEB J.* **29** :4071–4079, FIG. 2A.

Environmental DNA sequence information in the form of metagenomes and single-cell genomes can provide valuable predictions of gene function but cannot test those predictions. How, then, can the functions be validated for genes discovered in microbes that cannot be cultured? One method that has proved successful is to clone and express the gene of interest in a surrogate host, such as *E. coli* . Gene cloning for expression is described in detail in eAppendix 3 , but in brief, *E. coli* can transcribe foreign DNA into RNA, which can be translated into protein. While these proteins may lack any posttranslational processing that normally occurs with the original organism (see Section 8.4), they often retain key structural and functional properties when synthesized in the surrogate host.

Researchers at Argonne National Laboratory were interested in the potential new properties of the enzymes found in the uncultured archaea found by Lloyd's team, so they collaborated to clone and express these genes in *E. coli* . One protein of interest, S15, was predicted to function as an exopeptidase; that is, a peptidase that hydrolyzes the terminal amino acids of proteins. Purified preparations of S15 confirmed this prediction and showed that this enzyme, which assembles into a tetrameric complex (**Fig. 7.34C**), has a preference for cysteine and hydrophobic residues at the N terminus of proteins.

Enzymes such as these, isolated from uncultured microbes in exotic, often extreme environments, may have novel substrates or products and may have greater activity under extreme conditions (for example, high or low temperature or pH) that are optimal for industrial applications.

To Summarize

- **Microbiomes** are the complex microbial communities that live in environments such as the human gut, the ocean, and every other inhabited location on Earth.
- **Metagenomics** uses rapid DNA sequencing and other genomic techniques to study consortia of microbes directly in their natural environment.
- **Samples are obtained from a target community of a defined environment.** Sampling requires separating the microbes from their physical environment, breaking open the cells, and purifying the DNA.
- **The sequence reads are assembled into scaffolds and binned into partial genomes.** Assembly requires a computational pipeline that incorporates mathematical tools and biological assumptions.
- **Single-cell genomics** can analyze the partial genome of a single microbe without the need for cultivation.
- **Genes identified by metagenomics or SCG** can be cloned and expressed in other microbes as a form of bioprospecting.

Glossary

microbiota *or* microbiome
: The total community of microbes associated with an organism (such as the human body) or with a defined habitat (such as soil

or plants).

microbiota *or* microbiome
: The total community of microbes associated with an organism (such as the human body) or with a defined habitat (such as soil or plants).

metagenome
: The sum of genomes of all members of a community of organisms.

metagenomics
: The study of community genomes, or metagenomes.

operational taxonomic unit (OTU)
: A taxonomic group that is defined by a designated degree of similarity among members on the basis of DNA sequence.

target community
: A community whose genomes are sequenced for metagenomic analysis.

metabarcoding
: Also called *iTAG analysis* . The use of a short DNA sequence, such as the gene for SSU rRNA, to screen taxa from a metagenome.

library construction
: The amplification and processing of DNA for sequencing reactions, such as those of next-generation sequencing (NGS).

read
: A short DNA sequence that is generated by shotgun or next-generation sequencing methods.

assembly
: 1. In a virus, the packaging of a viral genome into the capsid to form a complete virion. 2. In metagenomics, the piecing together of DNA sequence reads into contigs, and of contigs into a scaffold.

contig
: A sequence of overlapping fragments of cloned DNA that are contiguous along a chromosome.

scaffold

The assembly of contigs (regions of contiguous sequence) into a large segment of a draft genome.

computational pipeline

A linear series of programs that combine mathematical tools with biological assumptions to interpret genetic sequences.

single-cell genomics

The study of the genome of a single cell.

eResearch Activity 7

Can Single-Cell Genomics Reveal Prebiotic Consumers of the Gut Microbiome?

Gut microbiomes can provide multiple benefits to their animal hosts, one of which is to convert an undigestible food source into one that the host can consume (see Chapter 13). One such food is the dietary fiber inulin (not to be confused with insulin), a polysaccharide composed of fructose monomers produced by some plants (**Fig. ERA 7.1**). Inulin, also called linear β(2→1) fructan, is a prebiotic that promotes the growth of beneficial gut microbes. And, while the host cannot digest this resource directly, fermentation of inulin by members of the gut microbiome releases short-chain fatty acids (SCFAs) that can be absorbed by the intestinal epithelium (see Section 21.4).

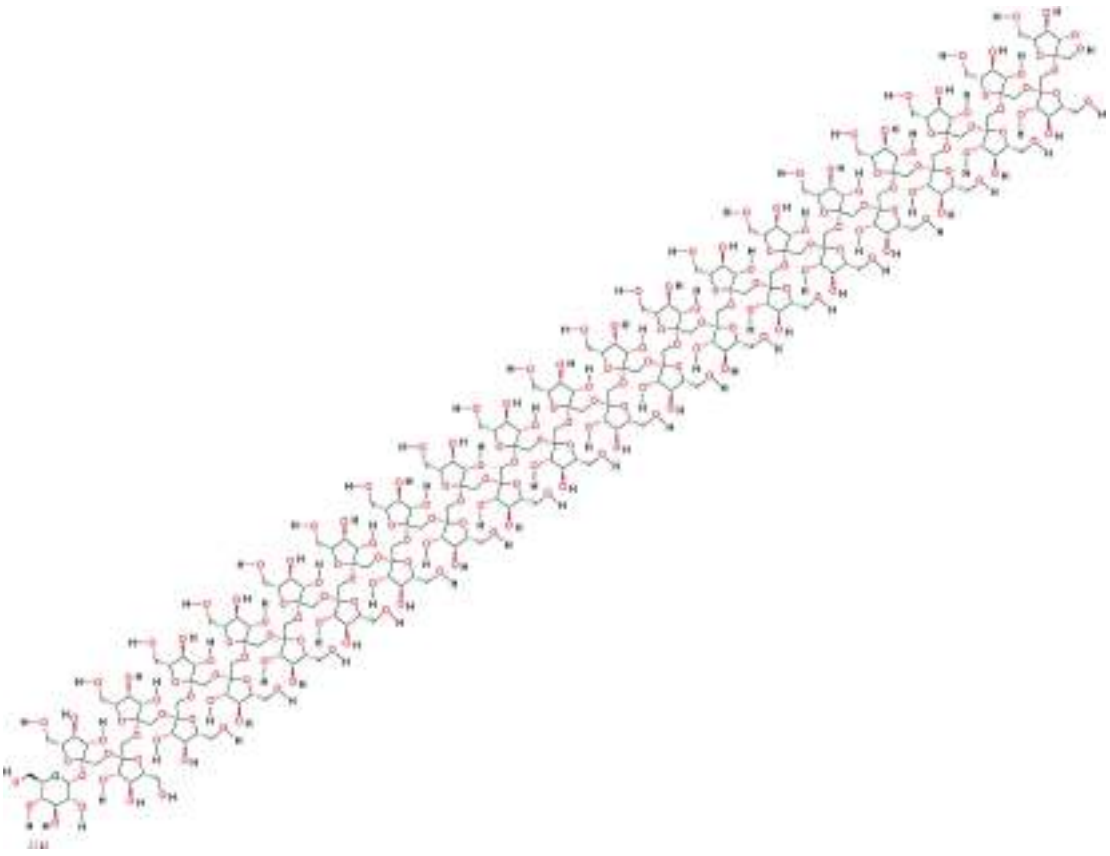
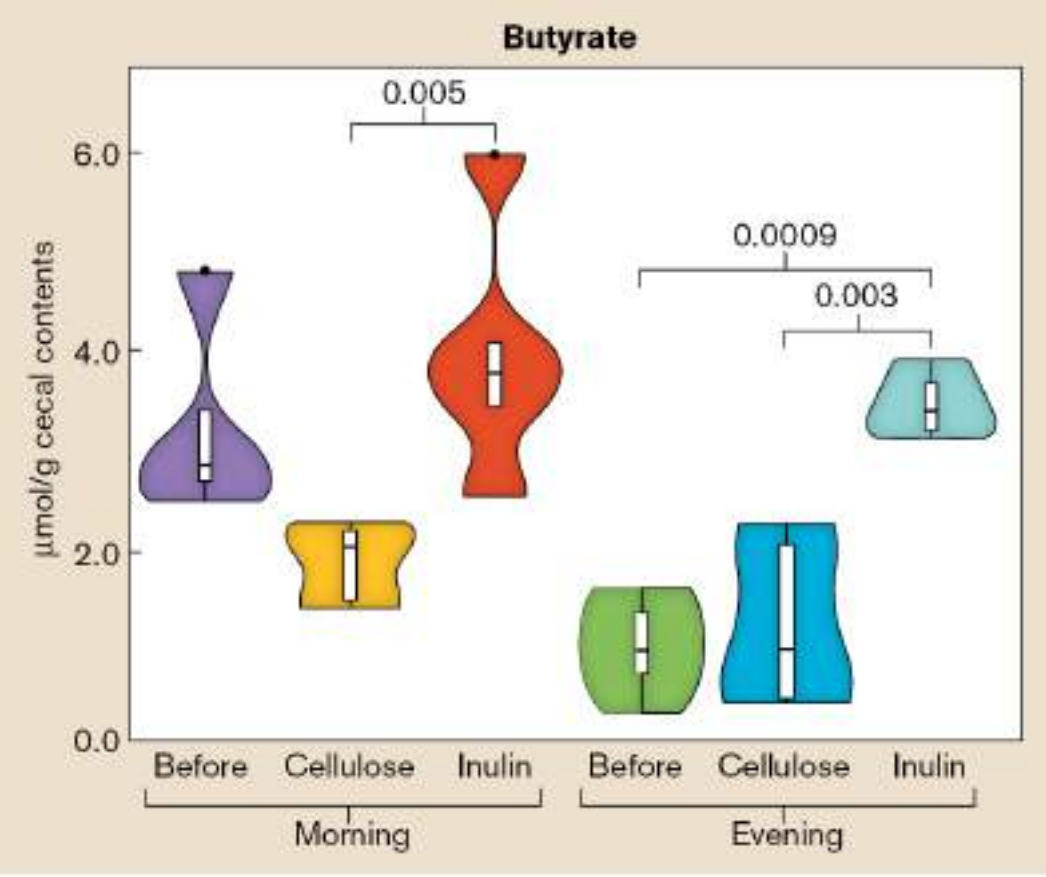


FIGURE ERA 7.1 ■ Structure of inulin, a fructose polymer made by some plants.

Because the gut microbiome is highly diverse, identifying which members consume a specific resource such as inulin can be daunting. At Waseda University in Tokyo, Japan, Rieka Chijiwa, Masahito Hosokawa, and colleagues tackled this objective by using a single-cell genomics approach and the mouse gut as their experimental system. The researchers reasoned that feeding the mice with an inulin supplement would stimulate the growth of the inulin consumers relative to the nonconsumers and lead to an increase in SCFAs in the cecum (gut). The researchers assayed for changes in relative abundances by high-throughput DNA sequencing of the community and used the 16S ribosomal RNA (rRNA) gene for taxonomic identification. They collected samples in both the morning and evening, as the mouse gut microbiome composition was known to oscillate over the course of the day.

After 2 weeks of inulin supplementation, the researchers saw large changes in both SCFA concentration in the cecum (**Fig. ERA 7.2**) and microbiome composition of feces (**Fig. ERA 7.3**), especially in the evening. Whereas cellulose supplementation had no effect, inulin supplementation led to significant increases in the SCFA butyrate and in the tricarboxylic acid (TCA) cycle intermediate succinate. Because mice do not have the enzymes to degrade inulin, this result suggests that one or more members of the microbiome performed this function, likely to serve their own metabolic needs for growth. Among the microbiota, the family Bacteroidaceae increased most dramatically in inulin-fed mice. This increase was not seen when the mice were fed a different dietary fiber, cellulose, which indicated that the response of Bacteroidaceae was specific to inulin. Could the members of the Bacteroidaceae be responsible for the breakdown of inulin and release of SCFAs?



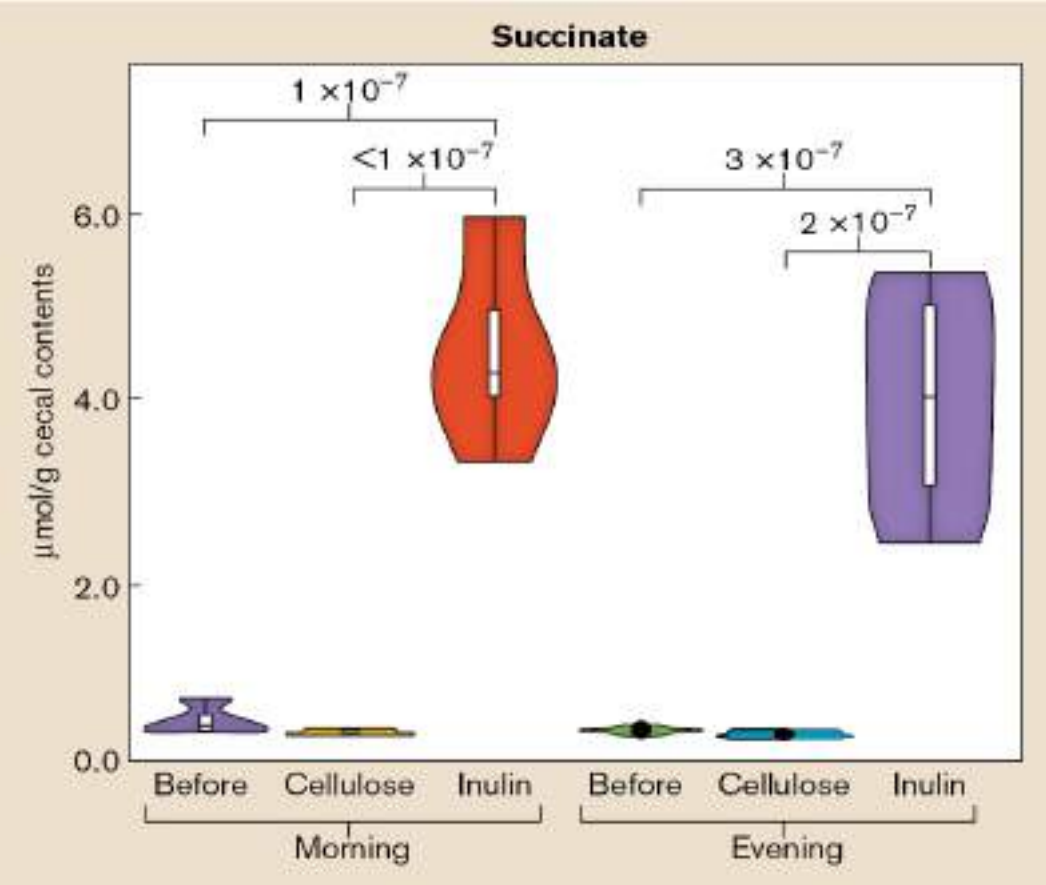


FIGURE ERA 7.2 ■ Concentration of butyrate and succinate in the mouse cecum before and after supplementation with dietary fiber. Samples were collected in the morning and evening, both before the period of supplementation and after 2 weeks of supplementation with either cellulose or inulin. Data are presented as "violin plots" that display the probability density of the data at different values along the *y* -axis. Brackets with numbers above them indicate pairs of values whose differences are statistically significant.

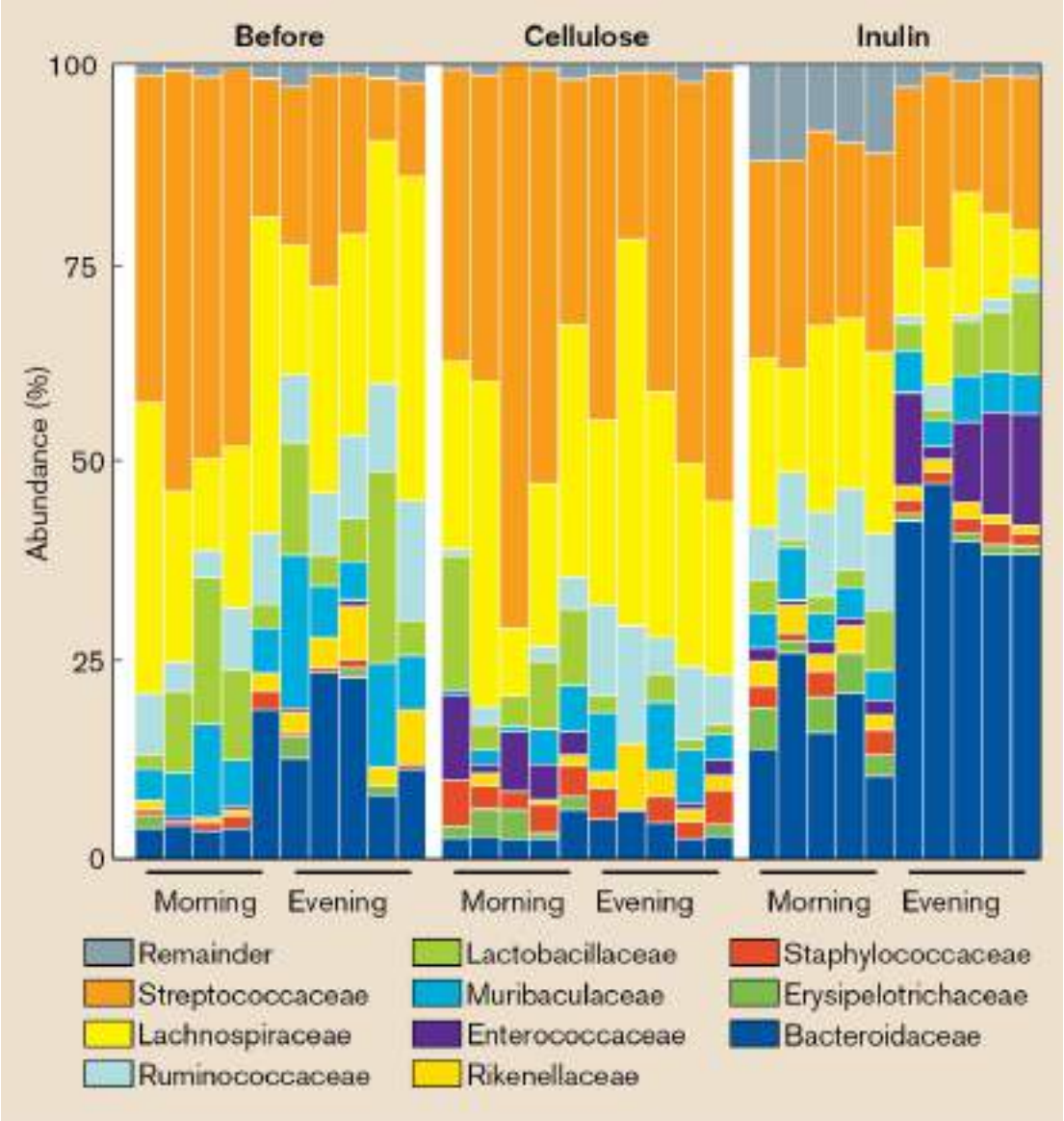


FIGURE ERA 7.3 ■ Microbial composition of mouse feces before and after supplementation with dietary fiber. High-throughput sequencing of the community's 16S rRNA genes revealed the relative abundance of the numerically dominant bacterial families. Samples were collected in the morning and evening, both before the period of supplementation and after 2 weeks of supplementation with either cellulose or inulin. Each column represents a sample acquired from a single mouse.

While 16S rRNA provides taxonomic identity, it does not reveal the presence or absence of other genes in the genome. During evolution, many genes are acquired via horizontal gene transfer or lost through deletion (see Section 9.3), and it is therefore dangerous to assume that taxonomy equals function. Regarding inulin, some members of Bacteroidaceae are known to consume it, whereas others lack this ability. With such uncertainty in mind, the next step was a whole-genome analysis of the Bacteroidaceae present in the gut of inulin-treated mice to find which, if any, had the genes needed to metabolize inulin.

Bacterial cells were isolated from the feces of the inulin-fed mice, and a microfluidic droplet generator was used to capture single cells within picoliter-sized agarose gel beads (**Fig. ERA 7.4**). Cells were then treated with enzymes to lyse the cells, releasing genomic DNA (gDNA), which remained within the gel. Whole-genome amplification of the gDNA then occurred within the gel bead, and this amplified DNA—a single-amplified genome (SAG)—was likewise trapped within the bead. Beads with amplified DNA could be distinguished from those lacking this amplification by their greater fluorescence when stained with a DNA-binding fluorophore. Flow cytometry (see Section 4.3) was then used to detect the high-fluorescing beads and sort each of them into a well within a microtiter plate. Within each well, whole-genome sequencing was performed (see Section 7.6 in the printed book and eAppendix 3), which led to the acquisition of 346 total SAGs from the inulin-treated mice. The taxonomic identity of these SAGs was then determined by examining their 16S rDNA sequence.

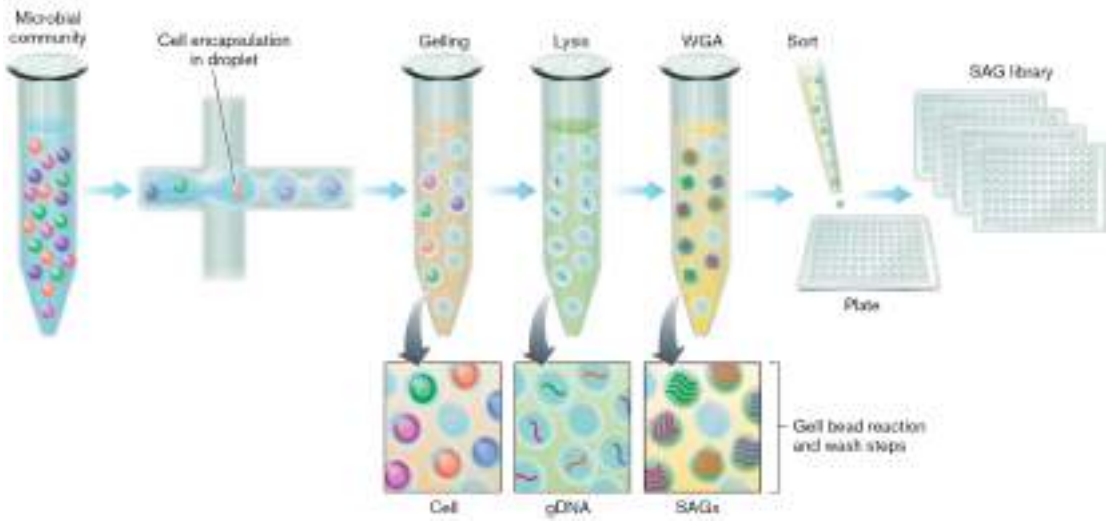


FIGURE ERA 7.4 ■ Preparation of a library of single-amplified genomes (SAGs). Random, single cells of the microbial community within the mouse feces were encapsulated into agarose gel droplets. Cells within the beads were lysed, and the trapped genomic DNA (gDNA) was subjected to whole-genome amplification (WGA). Beads with amplified DNA were distinguished by greater fluorescence of a DNA-staining fluorophore (green background) and were sorted by flow cytometry into individual wells of a microtiter plate.

Of the 346 SAGs, 24% belonged to the Bacteroidaceae—a high but expected fraction, due to the random nature of the cell encapsidation process and the high relative abundance of this family in the inulin-treated microbiome (**Fig. ERA 7.2**). Two of the Bacteroidaceae SAGs showed no potential for inulin degradation. However, two other Bacteroidaceae SAGs, IMSAGC_001 and IMSAGC_004, contained complete copies of a gene cluster known as the polysaccharide utilization locus (PUL), whose genes have functions identified by prior genetic analyses in other strains of Bacteroidaceae (**Fig. ERA 7.5**).

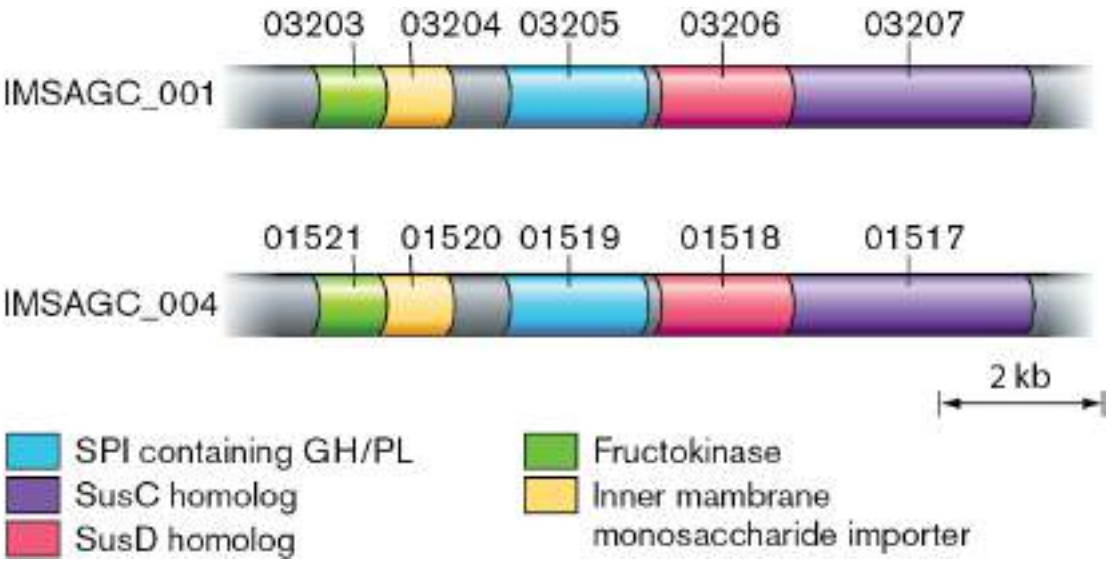


FIGURE ERA 7.5 ■ The polysaccharide utilization loci from two Bacteroidaceae SAGs. Predicted products of the genes within each PUL cluster are noted in the key. Numbers above the genes refer to their relative positions on the chromosome.

The genes in the PUL cluster work collectively to break down inulin, import the breakdown products, and finish catabolizing this resource within the cytoplasm. Inulin degradation begins outside the bacterial cell, as it is too large to import. The PUL cluster includes a gene encoding a glycoside hydrolase or polysaccharide lyase (GH/PL), which is exported from the cell because it has a signal peptide that is recognized by protein export machinery (see Section 8.5). This signal peptide has a SPI motif that is cleaved by s ignal p eptidase I once the protein enters the periplasm; once cleaved, the released GH/PL can be exported beyond the outer membrane and enter the mouse's gut, where it can break down inulin.

Once broken down into smaller oligosaccharides, inulin can be further processed by proteins encoded by the PUL cluster. The oligosaccharides enter the periplasm via the outer membrane channel formed by SusC and SusD proteins, and, once broken down to fructose, monomers enter the cytoplasm via a PUL-encoded monosaccharide importer. Once in the cell, the fructose can be

metabolized by the PUL-encoded fructokinase. Notably, in addition to the PUL clusters, the IMSAGC_001 and IMSAGC_004 SAGs both had an incomplete TCA cycle pathway. The end product of this incomplete pathway is succinate, one of the metabolites that increased in concentration in the inulin-fed mouse gut.

Follow-up studies will need to be performed to confirm that the genes encoded in the PUL clusters of these SAGs have the functions predicted by their similarity (homology) to genes of verified function. Additionally, the microbes that these SAGs represent will need to be tested for their ability to degrade inulin in the mouse gut. A major value of this SAG-based approach is that it has provided researchers with an excellent lead on which members of a complex gut microbiome are likely to provide a very specific and very important metabolic function.

Further Exploration

Because the GH/PL is released into the mouse gut, is there anything that prevents other members of the microbiome from competing with the GH/PL-exporting cells for those inulin breakdown products?

Source: **Chijiwa, R., M. Hosokawa, M. Kogawa, Y. Nishikawa, K. Ide, et al.** 2020. Single-cell genomics of uncultured bacteria reveals dietary fiber responders in the mouse gut microbiota. *Microbiome* **8** :5.

CHAPTER REVIEW

Review Questions

1. Explain the structural types of bacterial genomes.
2. What are the differences between DNA and RNA?
3. Explain DNA supercoiling. Why is it important to microbial genomes?
4. Discuss the mechanisms of topoisomerases. What drugs target the topoisomerase enzyme DNA gyrase?
5. What are the basic mechanisms of DNA replication?
6. How does the bacterial cell regulate the initiation of chromosome replication?
7. What is the clamp loader? Primase? DNA helicase (DnaB)? Helicase loader (DnaC)? DNA proofreading?
8. How is the problem of replicating both strands at a replication fork solved?
9. What is a catenane? What does it have to do with DNA replication?
10. How are chromosome dimers that form by homologous recombination resolved to separate the chromosomes?
11. How does rolling-circle replication compare with bidirectional replication?
12. What is handcuffing?
13. What distinguishes a secondary chromosome from a primary chromosome or a plasmid?

14. What are the similarities and differences in genome structure for the eukaryotes and archaea, relative to the bacteria?
15. What is a metagenome, how is it constructed, and how can it be used for studies of microbiomes?
16. Under what situations may single-cell genomics be preferred over metagenome sequencing?

Thought Questions

1. Why might metagenomics or single-cell genomics be more successful at bioprospecting for novel microbial activities than traditional cultivation approaches are?
2. During rapid growth, why would a bacterial cell die if an antibiotic drug formed, as the text says, "a physical barrier in front of the DNA replication complex" (Section 7.2)?
3. If you were to synthesize a microbe from scratch, perhaps for industrial production of antibiotics, how would you design the genome? Single chromosome or multiple chromosomes? Plasmids included or not? Linear or circular DNA?

Key Terms

- antiparallel (253)
- assembly (282)
- catenane (268)
- computational pipeline (282)
- contig (282)
- denature (254)

- DNA ligase (267)
- DNA replication (262)
- enhancer (276)
- essential gene (274)
- exonuclease (265)
- gene (250)
- genome (250)
- histone (276)
- homologous recombination (269)
- hybridization (255)
- intron (276)
- library construction (281)
- metabarcoding (281)
- metagenome (278)
- metagenomics (278)
- microbiome (277)
- nitrogenous base (253)
- nucleobase (253)
- nucleoid (255)
- Okazaki fragments (263)
- operational taxonomic unit (OTU) (279)
- origin (*oriC*) (262)
- phosphodiester link (253)
- plasmid (270)
- pre-catenane (268)
- primase (265)
- promoter (276)
- proofreading (265)
- pseudogene (276)
- purine (253)
- pyrimidine (253)
- quinolone (260)
- read (282)

- replication fork (261)
- replicon (275)
- replisome (263)
- scaffold (282)
- secondary chromosome (274)
- semiconservative (261)
- single-cell genomics (282)
- sliding clamp (265)
- supercoiled DNA (256)
- target community (280)
- telomerase (275)
- telomere (275)
- termination (*ter*) site (262)
- topoisomerase (258)
- transformation (250)

Recommended Reading

Anda, Mizue, Yoshiyuki Ohtsubo, Takashi Okubo, Masayuki Sugawara, Yuji Nagata, et al. 2015. Bacterial clade with a ribosomal RNA operon on a small plasmid rather than the chromosome. *Proceedings of the National Academy of Sciences USA* **112** :14343–14347.

Ausiannikava, Darya, and Thorsten Allers. 2017. Diversity of DNA replication in the Archaea. *Genes* **8** :56.

Bouet, Jean-Yves, Mathiu Stouf, Elise Lebailly, and Francois Cornet. 2014. Mechanisms for chromosome segregation. *Current Opinion in Microbiology* **22** :60–65.

Bury, Katarzyna, Katarzyna Wegrzyn, and Igor Konieczny. 2017. Handcuffing reversal is facilitated by proteases and replication initiator monomers. *Nucleic Acids Research* **45** :3953–3966.

Campbell, Christopher S., and R. Dyche Mullins. 2007. In vivo visualization of type II plasmid segregation: Bacterial actin filaments pushing plasmids. *Journal of Cell Biology* **179** :1059–1066.

Costa, Allesandro, Iris V. Hood, and James M. Berger. 2013. Mechanisms for initiating cellular DNA replication. *Annual Reviews of Biochemistry* **82** :25–54.

Dame, Remus T., Olga J. Kalmykowa, and David C. Grainger. 2011. Chromosomal macrodomains and associated proteins: Implications for DNA organization and replication in Gram-negative bacteria. *PLoS Genetics* **7** :e1002123.

Dame, Remus T., Fatema-Zahra M. Rashid, and David C. Grainger. 2020. Chromosome organization in bacteria: Mechanistic insights into genome structure and function. *Nature Reviews Genetics* **21** :227–242.

Dewar, James M., and Johannes C. Walter. 2017. Mechanisms of DNA replication termination. *Nature Reviews. Molecular Cell Biology* **18** :507–516.

diCenzo, George C., and Turlough M. Finan. 2017. The divided bacterial genome: Structure, function, and evolution. *Microbiology and Molecular Biology Reviews* **81** :e00019-17.

Duggin, Ian G., R. Gerry Wake, Stephen D. Bell, and Thomas M. Hill. 2008. The replication fork trap and termination of chromosome replication. *Molecular Microbiology* **70** :1323–1333.

Georgescu, Roxana E., Isabel Kurth, and Mike E. O'Donnell. 2011. Single-molecule studies reveal the function of a third polymerase in the replisome. *Nature Structural & Molecular Biology* **19** :113–116.

Goodrich, J. K., J. L. Waters, A. C. Poole, J. L. Sutter, O. Koren, et al. 2014. Human genetics shape the gut microbiome. *Cell* **159** :789–799.

Hedlund, Brian P., Jeremy A. Dodsworth, Senthil K. Murugapiran, Christian Rinke, and Tanja Woyke. 2014. Impact of single-cell genomics and metagenomics on the

emerging view of extremophile "microbial dark matter." *Extremophiles* **18** :865–875.

Kuzminov, Andrei. 2014. The precarious prokaryotic chromosome. *Journal of Bacteriology* **196** :1793–1806.

Le, Tung Bk, and Michael T. Laub. 2014. New approaches to understanding the spatial organization of bacterial genomes. *Current Opinion in Microbiology* **22** :15–21.

Lindas, A. C., and R. Bernander. 2013. The cell cycle of archaea. *Nature Reviews. Microbiology* **11** :627–638.

Männik, Jaana, Matthew W. Bailey, Jordan C. O'Neill, and Jaan Männik. 2017. Kinetics of large-scale chromosomal movement during asymmetric cell division in *Escherichia coli* . *PloS Genetics* **13** :e1006638.

McHenry, Charles S. 2011. Bacterial replicases and related polymerases. *Current Opinion in Chemical Biology* **15** :587–594.

Michalska, Karolina, Andrew D. Steen, Gekleng Chhor, Michael Endres, Austen T. Webber, et al. 2015. New aminopeptidase from "microbial dark matter" archaeon. *FASEB Journal* **29** :4071–4079.

Reyes-Lamothe, Rodrigo, and David J. Sherratt. 2019. The bacterial cell cycle, chromosome inheritance and cell growth. *Nature Reviews. Microbiology* **17** :467–478.

Reyes-Lamothe, Rodrigo, Tung Tran, Diane Meas, Laura Lee, Alice M. Li, et al. 2014. High-copy bacterial plasmids diffuse in the nucleoid-free space, replicate stochastically and are randomly partitioned at cell division. *Nucleic Acids Research* **42** :1042–1051.

Robinson, Andrew, and Antoine M. van Oijen. 2013. Bacterial replication, transcription and translation: Mechanistic insights from single-molecule biochemical studies. *Nature Reviews. Microbiology* **11** :303–315.

Sanchez-Romero, Maria A., Ignatio Cota, and Josep Casadesus. 2015. DNA methylation in bacteria: From the

methyl group to the methylome. *Current Opinion in Microbiology* **25** :9–16.

Song, Dan, and Joseph J. Loparo. 2015. Building bridges within the bacterial chromosome. *Trends in Genetics* **31** :164–173.

Sunagawa, Shinichi, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, et al. 2015. Structure and function of the global ocean microbiome. *Science* **348** :6237.

Val, Marie-Eve, Martial Marbouty, Francisco de Lemos Martins, Sean P. Kennedy, Harry Kemble, et al. 2016. A checkpoint control orchestrates the replication of the two chromosomes of *Vibrio cholerae* . *Science Advances* **2** :e1501914.

Wegrzyn, Katarzyna E., Marta Gross, Urszula Uciechowska, and Igor Konieczny. 2016. Replisome assembly at bacterial chromosomes and iteron plasmids. *Frontiers in Molecular Biosciences* **3** :39.

Wolanski, Marcin, Rafal Donczew, Anna Zawilak-Pawlik, and Jolanta Zakrzewska-Czerwinska. 2014. *oriC* -encoded instructions for the initiation of bacterial chromosome replication. *Frontiers in Microbiology* **5** :735.

Glossary

gene
: A sequence of nucleotides that has a distinct function (regulatory) or whose encoded product (protein or RNA) has a distinct function. The functional unit of heredity.

transformation
: The internalization of free DNA from the environment into bacterial cells.

genome
: The complete genetic content of an organism. The sequence of all the nucleotides in a haploid set of chromosomes.

nucleobase
: Also called *nitrogenous base* . A planar, heteroaromatic, nitrogen-containing base that forms a nucleotide of nucleic acids; nucleobases determine the information content of DNA and RNA. There are five nucleobases: adenine, cytosine, guanine, thymine, and uracil.

nitrogenous base
: See **nucleobase** .

phosphodiester link
: The linkage between two adjacent nucleotides in a nucleic acid. A phosphate forms ester bonds with the 5′ carbon of one (deoxy-)ribose and the 3′ carbon of the other (deoxy-)ribose.

antiparallel
: Oriented such that the two strands are in opposite directions. Commonly refers to a nucleic acid double helix with one strand in the 5′-to-3′ orientation and the other strand in the 3′-to-5′ orientation.

purine
: A nitrogenous base with fused rings (that is, a bicyclic nucleobase) found in nucleotides; examples are adenine and guanine.

pyrimidine
: A single-ring nitrogenous base (that is, a monocyclic nucleobase) found in nucleotides; examples are cytosine, thymine, and uracil.

denature
: To lose secondary and tertiary structure in a protein or nucleic acid because of high temperature or chemical treatment.

hybridization
: The annealing of a nucleic acid strand with another nucleic acid strand containing a complementary sequence of bases. The binding of one nucleic acid strand with a complementary strand.

nucleoid
: The looped coils of a bacterial chromosome.

supercoiled DNA

A physical state of circular DNA where the number of twists of the two strands around each other is either less than or greater than the number found in the relaxed state of the double-stranded helix. The resulting tension generates a higher-order helix that compacts the DNA double-stranded helix.

topoisomerase
: An enzyme that can change the supercoiling of DNA.

quinolone
: A type of antibiotic drug that inhibits DNA synthesis by targeting bacterial topoisomerases such as DNA gyrase.

semiconservative
: Describing the mode of DNA replication whereby each new double helix contains one old, parental strand and one newly synthesized daughter strand.

replication fork
: During DNA synthesis, the region of the chromosome that is being unwound.

DNA replication
: The biological process of making an identical copy of double-stranded DNA using existing DNA as a template.

origin (*oriC*)
: The region of a bacterial or archaeal chromosome where DNA replication initiates.

termination (*ter*) site
: A bacterial sequence of DNA that halts replication of DNA by DNA polymerases elongating from both directions around the circular genome.

Okazaki fragments
: Short fragments of DNA that are synthesized on the lagging strand during DNA synthesis.

replisome
: A complex of DNA polymerase and other accessory molecules that performs DNA replication.

primase
: An RNA polymerase that synthesizes short RNA primers complementary to a DNA template to launch DNA replication.

sliding clamp
: A protein that keeps DNA polymerase affixed to DNA during replication.

proofreading
: An enzymatic activity of some nucleic acid polymerases that attempts to correct mispaired bases.

exonuclease
: An enzyme that cleaves DNA from the end.

DNA ligase
: An enzyme that cells use to form a covalent bond at a nick in the phosphodiester backbone. It is also used in molecular biology laboratories to join pieces of DNA.

pre-catenane
: A knot of intertwined DNA that is generated during chromosome replication.

catenane
: A pair of linked rings of DNA that occurs as a by-product of replication of circular chromosomes.

homologous recombination
: The process by which two DNA molecules exchange arms by cutting and splicing their helix backbones. Exchange occurs between sequences that are identical or nearly identical, as the machinery requires complementary base pairing to exchange the DNA molecules.

plasmid
: An extrachromosomal genetic element that may be present in some cells. Plasmids carry no essential genes.

secondary chromosome
: A plasmid-like small chromosome that carries at least one essential gene.

essential gene
: A gene that is required for cell viability under all environmental conditions.

replicon
: A nucleic acid molecule such as a chromosome or plasmid that is replicated autonomously, with its own origin of replication.

telomere
: The DNA segment at either end of a eukaryotic chromosome.

telomerase
: A reverse transcriptase enzyme complex that reads RNA as a template to synthesize DNA.

histone
: A protein that binds eukaryotic DNA and compacts chromosomes in nucleosomes.

enhancer
: A noncoding DNA regulatory region in eukaryotes that can lead to activation of transcription when bound by an appropriate transcription factor. Its location on the chromosome can be far removed from the regulated gene.

promoter
: A noncoding DNA regulatory region immediately upstream of a structural gene that is needed for transcription initiation.

intron
: In eukaryotic genes, an intervening sequence that does not code for protein and is spliced out of the mRNA prior to translation.

pseudogene
: A nonfunctional gene-like sequence that evolved by degenerative evolution.

microbiota *or* microbiome
: The total community of microbes associated with an organism (such as the human body) or with a defined habitat (such as soil or plants).

metagenome
: The sum of genomes of all members of a community of organisms.

metagenomics
: The study of community genomes, or metagenomes.

operational taxonomic unit (OTU)
: A taxonomic group that is defined by a designated degree of similarity among members on the basis of DNA sequence.

target community

A community whose genomes are sequenced for metagenomic analysis.

metabarcoding

Also called *iTAG analysis* . The use of a short DNA sequence, such as the gene for SSU rRNA, to screen taxa from a metagenome.

library construction

The amplification and processing of DNA for sequencing reactions, such as those of next-generation sequencing (NGS).

read

A short DNA sequence that is generated by shotgun or next-generation sequencing methods.

assembly

1. In a virus, the packaging of a viral genome into the capsid to form a complete virion. 2. In metagenomics, the piecing together of DNA sequence reads into contigs, and of contigs into a scaffold.

contig

A sequence of overlapping fragments of cloned DNA that are contiguous along a chromosome.

scaffold

The assembly of contigs (regions of contiguous sequence) into a large segment of a draft genome.

computational pipeline

A linear series of programs that combine mathematical tools with biological assumptions to interpret genetic sequences.

single-cell genomics

The study of the genome of a single cell.