Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

CHAPTER 3

Genomes

WHAT'S IN THIS CHAPTER?

- We start by describing the diversity of virus genomes.
- To understand how this affects virus replication, we consider the major genetic mechanisms that affect viruses.
- We finish by looking at representative virus genomes to illustrate the various possibilities.

CONTENTS

The Structure and Complexity of Virus Genomes 55
Molecular genetics..57
Virus genetics.....61
Virus mutants.....63
Spontaneous mutations 63
Induced mutations 64
Types of mutant viruses..... 64
Genetic interactions between viruses.....66
Nongenetic interactions between viruses.....69
Small DNA genomes.....70
Large DNA genomes.....75
Positive-strand RNA viruses.....78
Picornaviruses..... 79
Togaviruses..... 80
Flaviviruses 80
Coronaviruses 80
Positive-sense RNA plant viruses..... 81
Negative-strand RNA viruses.....81
Bunyaviruses..... 83
Arenaviruses 83
Orthomyxoviruses 83

THE STRUCTURE AND COMPLEXITY OF VIRUS GENOMES

Unlike the **genomes** of all cells, which are composed of DNA, virus genomes may contain their genetic information encoded in either DNA or RNA. The chemistry and structures of virus genomes are more varied than any of those seen in the entire bacterial, plant, or animal kingdoms. The nucleic acid making up the genome may be single stranded or double stranded, and it may have a linear, circular, or segmented structure. Single-stranded virus genomes may be either positive-sense (i.e., the same polarity or nucleotide sequence as the **mRNA**), negative-sense, or **ambisense** (a mixture of the two). Virus genomes range in size from approximately 2500 nucleotides (nt) (e.g., the geminivirus tobacco yellow dwarf virus at 2580 nt of single-stranded DNA) to approximately 1.2 million base pairs of double-stranded DNA (2,400,000 nt), in the case of Mimivirus, which is twice as big as the smallest bacterial genome (e.g., *Mycoplasma genitalum* at 580,000 base pairs). Some of the simpler **bacteriophages** are good examples of the smallest and least complex genomes. At the other end of the scale, the genomes of the largest double-stranded DNA viruses such as herpesviruses and poxviruses are sufficiently complex to still have escaped complete functional analysis (even though the complete nucleotide sequences of the genomes of a large number of examples are now known).

Principles of Molecular Virology. DOI: 10.1016/B978-0-12-384939-7.10003-1

Paramyxoviruses..... 83
Rhabdoviruses 84
Segmented and multipartite virus genomes.....84
Reverse transcription and transposition.....88
Evolution and epidemiology 97
Summary 100
Further Reading 100

Whatever the composition of a virus genome, each must follow one rule. Because viruses are obligate intracellular parasites only able to replicate inside the appropriate host cells, the genome must contain information encoded in a way that can be recognized and decoded by the particular type of host cell. The genetic code used by the virus must match or at least be recognized by the host organism. Similarly, the control signals that direct the expression of virus genes must be appropriate to the host. Many of the DNA viruses of **eukaryotes** closely resemble their host cells in terms of the biology of their genomes. Chapter 4 describes the ways in which virus genomes are replicated, and Chapter 5 deals in more detail with the mechanisms that regulate the expression of virus genetic information. The purpose of this chapter is to describe the diversity of virus genomes and to consider how and why this variation may have arisen.

Virus genome structures and nucleotide sequences have been intensively studied in recent decades because the power of recombinant DNA technology has focused much attention in this area. It would be wrong to present molecular biology as the only means of addressing unanswered problems in virology, but it would be equally foolish to ignore the opportunities that it offers and the explosion of knowledge that has resulted from it in recent years. As noted in Chapter 1, this has been (almost) matched by an explosion in digital bioinformatics techniques to process and make sense of all this data.

Some DNA virus genomes are complexed with cellular histones to form a **chromatin**-like structure inside the virus particle. Once inside the nucleus of the host cell, these genomes behave like miniature satellite chromosomes, controlled by cellular enzymes and the cell cycle:

- Vaccinia virus mRNAs were found to be polyadenylated at their 3′ ends by Kates in 1970—the first time this observation had been made in any organism.
- Split genes containing noncoding **introns**, protein-coding **exons**, and spliced mRNAs were first discovered in adenoviruses by Roberts and Sharp in 1977.

Introns in **prokaryotes** were first discovered in the genome of bacteriophage T4 in 1984. Several examples of this phenomenon have now been discovered in T4 and some other phages. This raises an important point. The conventional view is that prokaryote genomes are smaller and replicate faster than those of **eukaryotes** and hence can be regarded as streamlined. The genome of phage T4 consists of 160 kbp of double-stranded DNA and is highly compressed; for example, **promoters** and translation control sequences are nested within the coding regions of overlapping upstream genes. The presence of introns in **bacteriophage** genomes, which are under constant ruthless pressure to exclude junk sequences, suggests that these genetic elements must have evolved mechanisms to escape or neutralize this pressure and to persist as parasites

within parasites. All virus genomes experience pressure to minimize their size. Viruses with **prokaryotic** hosts must be able to replicate sufficiently quickly to keep up with their host cells, and this is reflected in the compact nature of many (but not all) bacteriophages. Overlapping genes are common, and the maximum genetic capacity is compressed into the minimum genome size. In viruses with **eukaryotic** hosts there is also pressure on genome size. Here, however, the pressure is mainly from the packaging size of the virus particle (i.e., the amount of nucleic acid that can be incorporated into the **virion**). Therefore, these viruses commonly show highly compressed genetic information when compared with the low density of information in eukaryotic cellular genomes.

There are exceptions to this rule. Some bacteriophages (e.g., the family *Myoviridae*, such as T4) have relatively large genomes, up to 170 kbp. The largest virus genome currently known is that of Mimivirus at approximately 1.2 Mbp, which contains around 1200 open reading frames, only 10% of which show any similarity to proteins of known function. Among viruses of eukaryotes, herpesviruses and poxviruses also have relatively large genomes, up to 235 kbp. It is notable that these virus genomes contain many genes involved in their own replication, particularly enzymes concerned with nucleic acid metabolism. These viruses partially escape the restrictions imposed by the biochemistry of the host cell by encoding additional biochemical equipment. The penalty is that they have to encode all the information necessary for a large and complex particle to package the genome, which is also an upward pressure on genome size. Later sections of this chapter contain detailed descriptions of both small, and compact, and large complex virus genomes.

BOX 3.1. IT'S NOT THE SIZE OF YOUR GENOME THAT COUNTS, IT'S WHAT YOU DO WITH IT

Traditionally it was thought that virus genomes were smaller than bacterial genomes. Often that is true, but not always. So does having a bigger genome make a better virus? Not in my opinion. As discussed in this chapter, some virus genomes are as complex as bacterial genomes, and larger than some of the smaller ones. This means they have nearly the same capabilities as bacteria, but not quite. No virus genome contains all the genes needed to make ribosomes, so in the end they are still parasites. Personally, my admiration goes to those stripped down miniature marvels that contain only a handful of genes and yet still manage to take over a cell and replicate themselves successfully. Now that's impressive.

Molecular genetics

As already described, the techniques of molecular biology have been a major influence on concentrating much attention on the virus genome. It is beyond

the scope of this book to give detailed accounts of these methods. However, it is worth taking some time here to illustrate how some of these techniques have been applied to virology, remembering that these newer techniques are complementary to and do not replace the classical techniques of virology. Initially, any investigation of a virus genome will usually include questions about the following:

- Composition: DNA or RNA, single stranded or double stranded, linear or circular
- Size and number of segments
- Nucleotide sequence
- Terminal structures
- Coding capacity: open reading frames
- Regulatory signals: transcription **enhancers**, **promoters**, and terminators

It is possible to separate the molecular analysis of virus genomes into two types of approaches: physical analysis of structure and nucleotide sequence, essentially performed *in vitro*, and a more biological approach to examine the structure–function relationships of intact virus genomes and individual genetic elements, usually involving analysis of the virus phenotype *in vivo*.

The conventional starting point for the physical analysis of virus genomes has been the isolation of nucleic acids from virus preparations of varying degrees of purity. To some extent, this is still true of molecular biology techniques, although the emphasis on extensive purification has declined as techniques of molecular cloning have become more advanced. DNA virus genomes can be analyzed directly by restriction endonuclease digestion without resorting to molecular cloning, and this approach was achieved for the first time with SV40 DNA in 1971. The first pieces of DNA to be molecularly cloned were restriction fragments of bacteriophage λ DNA, which were cloned into the DNA genome of SV40 by Berg and colleagues in 1972. This means that virus genomes were both the first cloning vectors and the first nucleic acids to be analyzed by these techniques. In 1977, the genome of bacteriophage ϕX174 was the first **replicon** to be completely sequenced.

Subsequently, phage genomes such as M13 were highly modified for use as vectors in DNA sequencing. The enzymology of RNA-specific nucleases was comparatively advanced at this time, such that a spectrum of enzymes with specific cleavage sites could be used to analyze and even determine the sequence of RNA virus genomes (the first short nucleotide sequences of tRNAs having been determined in the mid-1960s). However, direct analysis of RNA by these methods was laborious and notoriously difficult. RNA sequence analysis did not begin to advance rapidly until the widespread use of reverse transcriptase (isolated from retroviruses) to convert RNA into cDNA in the 1970s.

Since the 1980s, polymerase chain reaction (PCR) has further accelerated the investigation of virus genomes (Chapter 1).

In addition to molecular cloning, other techniques of molecular analysis have also been valuable in virology. Direct analysis by electron microscopy, if calibrated with known standards, can be used to estimate the size of nucleic acid molecules. Hybridization of complementary nucleotide sequences can also be used in a number of ways to analyze virus genomes (Chapter 1). Perhaps the most important single technique has been gel electrophoresis (Figure 3.1). The earliest gel matrix employed for separating molecules was based on starch and gave relatively poor resolution. It is now most common to use agarose gels to separate large nucleic acid molecules, which may be very large indeed—several megabases (million base pairs) in the case of techniques such as pulsed-field gel electrophoresis (PFGE) and polyacrylamide gel electrophoresis (PAGE) to separate smaller pieces (down to sizes of a few nucleotides). Apart from the fact that sequencing depends on the ability to separate molecules that differ from each other by only one nucleotide in length, gel electrophoresis has been of great

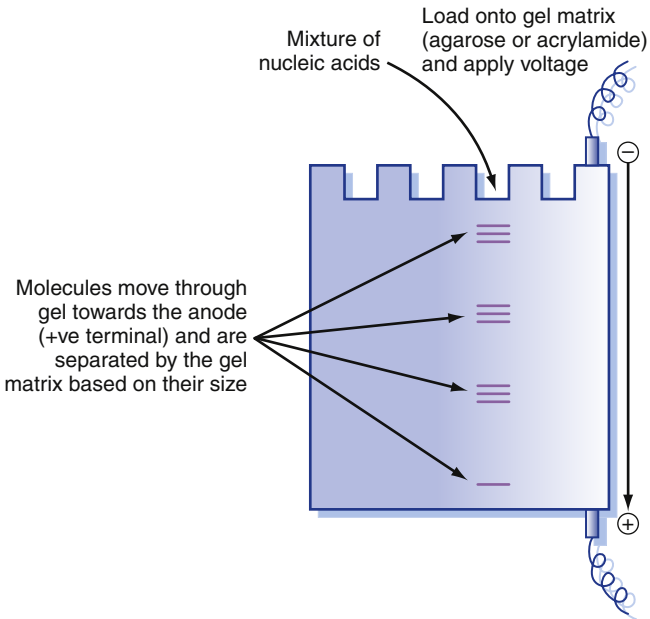

FIGURE 3.1 Gel electrophoresis.

In gel electrophoresis, a mixture of nucleic acids (or proteins) is applied to a gel, and they move through the gel matrix when an electric field is applied. The net negative charge due to the phosphate groups in the backbone of nucleic acid molecules results in their movement away from the cathode and toward the anode. Smaller molecules are able to slip through the gel matrix more easily and thus migrate farther than larger molecules, which are retarded, resulting in a net separation based on the size of the molecules.

value in analyzing intact virus genomes, particularly the analysis of viruses with segmented genomes (see later discussion). The most recent and most powerful sequence analysis techniques such as pyrosequencing have done away with electrophoresis and rely on light detection from fluorescent compounds.

Phenotypic analysis of virus populations has long been a standard technique of virology. In modern terms, this might be considered functional genomics. Examination of variant viruses and naturally occurring spontaneous mutants is an old method for determining the function of virus genes. Molecular biology has added to this the ability to design and create specific mutations, deletions, and recombinants *in vitro*. This site-directed mutagenesis is a very powerful tool. Although genetic coding capacity can be examined *in vitro* by the use of cell-free extracts to translate **mRNAs**, complete functional analysis of virus genomes can be performed only on intact viruses. Fortunately, the relative simplicity of most virus genomes (compared with even the simplest cell) offers a major advantage here—the ability to rescue infectious viruses from purified or cloned nucleic acids. Infection of cells caused by nucleic acid alone is referred to as **transfection**.

Virus genomes that consist of positive-sense RNA are infectious when the purified RNA (vRNA) is applied to cells in the absence of any virus proteins. This is because positive-sense vRNA is essentially mRNA, and the first event in a normally infected cell is to translate the vRNA to make the virus proteins responsible for genome replication. In this case, direct introduction of RNA into cells circumvents the earliest stages of the replicative cycle (Chapter 4). Virus genomes that are composed of double-stranded DNA are also infectious. The events that occur here are a little more complex, because the virus genome must first be transcribed by host polymerases to produce mRNA. This is relatively simple for phage genomes introduced into **prokaryotes**, but for viruses that replicate in the nucleus of **eukaryotic** cells, such as herpesviruses, the DNA must first find its way to the appropriate cellular compartment. Most of the DNA that is introduced into cells by transfection is degraded by cellular nucleases. However, irrespective of its sequence, a small proportion of the newly introduced DNA finds its way into the nucleus, where it is transcribed by cellular polymerases.

Unexpectedly, cloned cDNA genomes of positive-sense RNA viruses (e.g., picornaviruses) are also infectious, although less efficient at infecting cells than the vRNA. This is presumably because the DNA is transcribed by cellular enzymes to make RNA. Synthetic RNA transcribed *in vitro* from the cDNA template of the genome is much more efficient at initiating infection. Such experiments are referred to as reverse genetics—that is, the manipulation of a virus via a cloned intermediate. Using these techniques, viruses can be rescued from cloned genomes, including those that have been manipulated *in vitro*.

Originally, this type of approach was not possible for analysis of viruses with negative-sense genomes. This is because all negative-sense virus

particles contain a virus-specific polymerase. The first event when these virus genomes enter the cell is that the negative-sense genome is copied by the polymerase, forming either positive-sense transcripts that are used directly as **mRNA** or a double-stranded molecule, known either as the replicative intermediate (RI) or replicative form (RF), which serves as a template for further rounds of mRNA synthesis. Therefore, because purified negative-sense genomes cannot be directly translated by the host cell and are not replicated in the absence of the virus polymerase, these genomes are inherently noninfectious. However, systems have now been developed that permit the rescue of viruses with negative-sense genomes from purified or cloned nucleic acids.

All such systems rely on a ribonucleoprotein complex that can serve as a template for genome replication by RNA-dependent RNA polymerase, but they fall into one of two approaches:

- *In vitro* complex formation: Virus proteins purified from infected cells are mixed with RNA transcribed from cloned cDNAs to form complexes that are then introduced into susceptible cells to initiate an infection. This method has been used for paramyxoviruses, rhabdoviruses, and bunyaviruses.
- *In vivo* complex formation: Ribonucleoprotein complexes formed *in vitro* are introduced into cells infected with a helper virus strain. This method has been used for influenza virus, bunyaviruses, and double-stranded RNA viruses such as reoviruses and birnaviruses.

Such developments open up possibilities for genetic investigation of negative- and double-stranded RNA viruses that have not previously existed, and are of particular interest because of their potential for vaccine development (see Chapter 6).

Virus genetics

Although nucleotide sequencing now dominates the analysis of virus genomes, functional genetic analysis of animal viruses is based largely on the isolation and analysis of mutants, usually achieved using plaque purification (biological cloning). In the case of viruses for which no such systems exist (because they either are not cytopathic or do not replicate in culture), little genetic analysis was possible before the development of molecular genetics. However, certain tricks make it possible to extend standard genetic techniques to noncytopathic viruses:

- **Biochemical analysis:** Use of metabolic inhibitors to construct genetic maps; inhibitors of translation (such as puromycin and cycloheximide) and transcription (actinomycin D) can be used to decipher genetic regulatory mechanisms.

- **Focal immunoassays:** Replication of noncytopathic viruses visualized by immune staining to produce visual foci (e.g., human immunodeficiency virus).
- **Physical analysis:** Use of high-resolution electrophoresis to identify genetic polymorphisms of virus proteins or nucleic acids.
- **Transformed foci:** Production of transformed foci of cells by noncytopathic focus-forming viruses (e.g., DNA and RNA tumor viruses).

Two types of genetic maps can be constructed:

- **Recombination maps:** These represent an ordered sequence of mutations derived from the probability of **recombination** between two genetic markers, which is proportional to the distance between them—a classic genetic technique. This method works for viruses with nonsegmented genomes (DNA or RNA).
- **Reassortment maps (or groups):** In viruses with segmented genomes, the assignment of mutations to particular genome segments results in identification of genetically linked reassortment groups equivalent to individual genome segments.

Other types of maps that can be constructed include:

- **Physical maps:** Mutations or other features can be assigned to physical locations on a virus genome using the rescue of mutant genomes by small pieces of the wild-type genome after **transfection** of susceptible cells. Alternatively, cells can be cotransfected with the mutant genome plus individual restriction fragments to localize the mutation. Similarly, various polymorphisms (such as electrophoretic mobility of proteins) can be used to determine the genetic structure of a virus.
- **Restriction maps:** Site-specific cleavage of DNA by restriction endonucleases can be used to determine the structure of virus genomes. RNA genomes can be analyzed in this way after cDNA cloning.
- **Transcription maps:** Maps of regions encoding various mRNAs can be determined by hybridization of mRNA species to specific genome fragments (e.g., restriction fragments). The precise start/finish of mRNAs can be determined by single-strand-specific nuclease digestion of radiolabelled probes. Proteins encoded by individual mRNAs can be determined by translation *in vitro*. Ultraviolet (UV) irradiation of RNA virus genomes can also be used to determine the position of open reading frames because those farthest from the translation start are the least likely to be expressed by *in vitro* translation after partial degradation of the virus RNA by UV light.
- **Translation maps:** Pactamycin (an antibiotic that inhibits translation) has been used to map protein-coding regions of enteroviruses. Pulse labeling results in incorporation of radioactivity only into proteins

initiated before addition of the drug. Proteins nearest the 3′ end of the genome are the most heavily labeled; those at the 5′ end of the genome are the least heavily labeled.

Virus mutants

Mutant, *strain*, *type*, *variant*, and even *isolate* are all terms used rather loosely by virologists to differentiate particular viruses from each other and from the original *parental*, *wild-type*, or *street* isolates of that virus. More accurately, these terms are generally applied as follows:

- **Strain:** Different lines or isolates of the same virus (e.g., from different geographical locations or patients)
- **Type:** Different serotypes of the same virus (e.g., various antibody neutralization phenotypes)
- **Variant:** A virus whose phenotype differs from the original wild-type strain but the genetic basis for the difference is not known (e.g., a new clinical isolate from a patient)

Mutant viruses can arise in various ways, described next.

Spontaneous mutations

In some viruses, mutation rates may be as high as 10^{-3} to 10^{-4} per incorporated nucleotide (e.g., in retroviruses such as human immunodeficiency virus, HIV), whereas in others they may be as low as 10^{-8} to 10^{-11} (e.g., in herpesviruses), which is similar to the mutation rates seen in cellular DNA. These differences are due to the mechanism of genome replication, with error rates in RNA-dependent RNA polymerases generally being higher than in DNA-dependent DNA polymerases. Some RNA virus polymerases do have proofreading functions, but in general mutation rates are higher in most RNA viruses than in DNA viruses. For a virus, mutations are a mixed blessing. The ability to generate antigenic variants that can escape the immune response is a clear advantage, but mutation also results in many defective particles, since most mutations are deleterious. In the most extreme cases (e.g., HIV), the error rate is 10^{-3} to 10^{-4} per nucleotide incorporated. The HIV genome is approximately 9.7 kb long; therefore, there will be 0.9 to 9.7 mutations in every genome copied. Hence, in this case, the wild-type virus actually consists of a fleeting majority type that dominates the dynamic equilibrium (i.e., the population of genomes) present in all cultures of the virus. These mixtures of molecular variants are known as **quasispecies** and also occur in other RNA viruses (e.g., picornaviruses). However, the majority of these variants will be noninfectious or seriously disadvantaged and are therefore rapidly weeded out of a replicating population. This mechanism is an important force in virus evolution (see "Evolution and Epidemiology").

Induced mutations

Historically, most genetic analysis of viruses has been performed on virus mutants isolated from mutagen-treated populations. Mutagens can be divided into two types:

- *In vitro* mutagens chemically modify nucleic acids and do not require replication for their activity. Examples include nitrous acid, hydroxylamine, and alkylating agents (e.g., nitrosoguanidine).
- *In vivo* mutagens require metabolically active (i.e., replicating) nucleic acid for their activity. These compounds are incorporated into newly replicated nucleic acids and cause mutations to be introduced during subsequent rounds of replication. Examples include base analogues such as 5-bromouracil, which result in faulty base pairing; intercalating agents (e.g., acridine dyes) that stack between bases, causing insertions or deletions; and UV irradiation, which causes the formation of pyrimidine dimers, which are excised from DNA by repair mechanisms that are much more error-prone than the usual enzymes used in DNA replication.

Experiments involving chemical mutagens suffer from a number of drawbacks:

- Safety is a concern, because mutagens are usually carcinogens and are also frequently highly toxic. They are very unpleasant compounds to work with.
- The dose of mutagen used must be chosen carefully to give an average of 0.1 mutation per genome; otherwise, the resultant viruses will contain multiple mutations that can complicate interpretation of the phenotype. Therefore, most of the viruses that result will not contain any mutations, which is inefficient because screening for mutants can be very laborious.

There is no control over where mutations occur, and it is sometimes difficult or impossible to isolate mutations in a particular gene or region of interest. For these reasons, site-specific molecular biological methods such as oligonucleotide-directed mutagenesis or PCR-based mutagenesis are now much more commonly used. Together with techniques such as enzyme digestion (to create deletions) and linker scanning (to create insertions), it is now possible to introduce almost any type of mutation precisely and safely at any specific site in a virus genome.

Types of mutant viruses

The phenotype of a mutant virus depends on the type of mutation(s) it has and also upon the location of the mutation(s) within the genome. Each of the following classes of mutations can occur naturally in viruses or may be artificially induced for experimental purposes:

- **Biochemical markers:** These include drug resistance mutations, mutations that result in altered virulence, polymorphisms resulting in altered electrophoretic

mobility of proteins or nucleic acids, and altered sensitivity to inactivating agents.

- **Deletions:** Similar in some ways to nonsense mutants (see later) but may include one or more virus genes and involve noncoding control regions of the genome (**promoters**, etc.). Spontaneous deletion mutants often accumulate in virus populations as defective-interfering (D.I.) particles. These noninfectious but not necessarily genetically inert genomes are thought to be important in establishing the course and pathogenesis of certain virus infections (see Chapter 6). Genetic deletions can only revert to wild-type by **recombination**, which usually occurs at comparatively low frequencies. Deletion mutants are very useful for assigning structure–function relationships to virus genomes, since they are easily mapped by physical analysis.
- **Host range:** This term can refer either to whole animal hosts or to permissive cell types *in vitro*. **Conditional mutants** of this class have been isolated using amber-suppressor cells (mostly for phages but also for animal viruses using *in vitro* systems).
- **Nonsense:** These result from alteration of a coding sequence of a protein to one of three translation stop codons (UAG, amber; UAA, ochre; UGA, opal). Translation is terminated, resulting in the production of an amino-terminal fragment of the protein. The phenotype of these mutations can be suppressed by propagation of a virus in a cell (bacterial or, more recently, animal) with altered suppressor tRNAs. Nonsense mutations are rarely leaky (i.e., the normal function of the protein is completely obliterated) and can only revert to wild-type at the original site (see later), so they usually have a low reversion frequency.
- **Plaque morphology:** Mutants may be either large-plaque mutants, which replicate more rapidly than the wild-type, or small-plaque mutants, which are the opposite. Plaque size is often related to a temperature-sensitive (t.s.) phenotype (see next). These mutants are often useful as unselected markers in multifactorial crosses.
- **Temperature-sensitive (t.s.):** This type of mutation is very useful because it allows the isolation of conditional-lethal mutations, a powerful means of examining virus genes that are essential for replication and whose function cannot otherwise be interrupted. Temperature-sensitive mutations usually result from missense mutations in proteins (i.e., amino acid substitutions), resulting in proteins of full size with subtly altered conformation that can function at (permissive) low temperatures but not at (nonpermissive) higher ones. Generally, the mutant proteins are immunologically unaltered, which is frequently useful. These mutations are usually leaky—that is, some of the normal activity is retained even at nonpermissive temperatures. Protein function is often impaired, even at permissive temperatures, therefore a high frequency of reversion is often

a problem with this type of mutation because the wild-type virus replicates faster than the mutant.

- **Cold-sensitive (c.s.):** These mutants are the opposite of t.s. mutants and are very useful in bacteriophages and plant viruses whose host cells can be propagated at low temperatures but are less useful for animal viruses because their host cells generally will not grow at significantly lower temperatures than normal.
- **Revertants:** Reverse mutation is a valid type of mutation in its own right. Most of the previous classes can undergo reverse mutation, which may be either simple back mutations (i.e., correction of the original mutation) or second-site compensatory mutations, which may be physically distant from the original mutation and not even necessarily in the same gene as the original mutation.
- **Suppression:** Suppression is the inhibition of a mutant phenotype by a second suppressor mutation, which may be either in the virus genome or in that of the host cell. This mechanism of suppression is not the same as the suppression of chain-terminating amber mutations by host-encoded suppressor tRNAs (see earlier), which could be called informational suppression. Genetic suppression results in an apparently wild-type phenotype from a virus that is still genetically mutant—a pseudorevertant. This phenomenon has been best studied in prokaryotic systems, but examples have been discovered in animal viruses—for example, reoviruses, vaccinia, and influenza—where suppression has been observed in an attenuated vaccine, leading to an apparently virulent virus. Suppression may also be important biologically in that it allows viruses to overcome the deleterious effects of mutations and therefore be positively selected.

Mutant viruses can appear to revert to their original phenotype by three pathways:

- Back mutation of the original mutation to give a wild-type genotype/phenotype (true reversion)
- A second, compensatory mutation that may occur in the same gene as the original mutation, thus correcting it—for example, a second frameshift mutation restoring the original reading frame (intragenic suppression)
- A suppressor mutation in a different virus gene or a host gene (extragenic suppression)

Genetic interactions between viruses

Genetic interactions between viruses often occur naturally, as host organisms are frequently infected with more than one virus. These situations are generally too complicated to be analyzed successfully. Experimentally, genetic interactions can be analyzed by mixed infection (superinfection)

of cells in culture. Two types of information can be obtained from such experiments:

- The assignment of mutants to functional groups known as **complementation** groups
- The ordering of mutants into a linear genetic map by analysis of **recombination** frequencies

Complementation results from the interaction of virus gene products during superinfection that results in production of one or both of the parental viruses being increased while both viruses remain unchanged genetically. In this situation, one of the viruses in a mixed infection provides a functional gene product for another virus that is defective for that function (Figure 3.2). If both mutants are defective in the same function, enhancement of replication does not occur and the two mutants are said to be in the same complementation group. The importance of this test is that it allows functional analysis of unknown mutations if the biochemical basis of any one of the mutations in a particular complementation group is known. In theory, the number of complementation groups is equal to the number of genes in the virus genome.

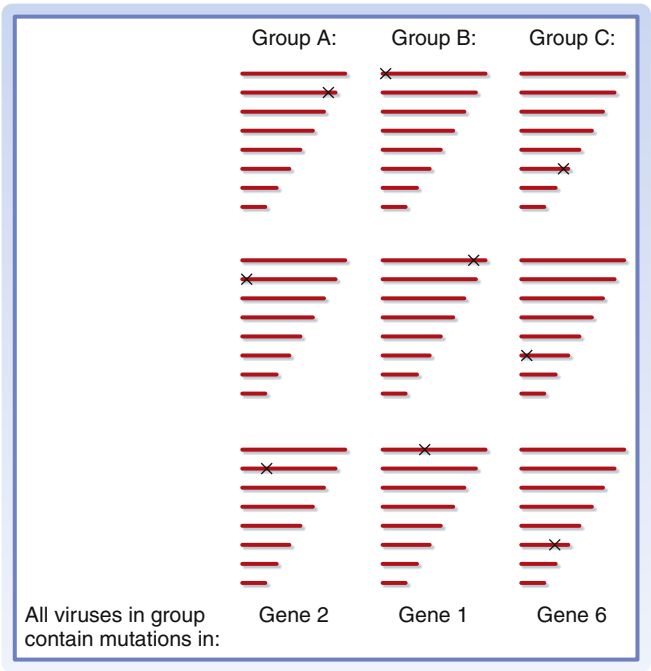

FIGURE 3.2 **Complementation groups in influenza.**
Complementation groups in influenza (or other viruses) all contain a mutation in the same virus gene, preventing the rescue of another mutant virus genome from the same complementation group.

In practice, there are usually fewer complementation groups than genes, as mutations in some genes are always lethal and other genes are nonessential and therefore cannot be scored in this type of test. There are two possible types of complementation:

- Allelic (intragenic) complementation occurs where different mutants have complementing defects in the same protein (e.g., in different functional domains) or in different subunits of a multimeric protein (although this is rare).
- Nonallelic (intergenic) complementation results from mutants with defects in different genes and is the more common type.

Complementation can be asymmetric—that is, only one of the mutant viruses is able to replicate. This can be an absolute or a partial restriction. When complementation occurs naturally, it is usually the case that a replication-competent wild-type virus rescues a replication-defective mutant. In these cases, the wild-type is referred to as a helper virus, such as in the case of defective transforming retroviruses containing **oncogenes** (see Figure 3.3 and Chapter 7). **Recombination** is the physical interaction of virus genomes during superinfection that results in gene combinations not present in either parent. There are three mechanisms by which this can occur, depending on the organization of the virus genome:

- **Intramolecular recombination via strand breakage and religation:** This process occurs in all DNA viruses and in RNA viruses that replicate via

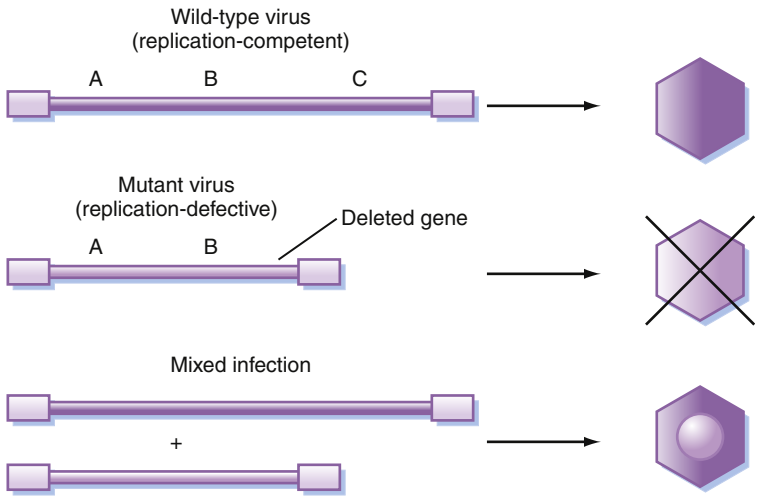

FIGURE 3.3 **Helper viruses.**
Helper viruses are replication-competent viruses that are capable of rescuing replication-defective genomes in a mixed infection, permitting their multiplication and spread.

a DNA intermediate. It is believed to be caused by cellular enzymes, since no virus mutants with specific recombination defects have been isolated.

- **Intramolecular recombination by copy-choice:** This process occurs in RNA viruses, probably by a mechanism in which the virus polymerase switches template strands during genome synthesis. There are cellular enzymes that could be involved (e.g., **splicing** enzymes), but this is unlikely and the process is thought to occur essentially as a random event. **Defective interfering** (**D.I.**) particles in RNA virus infections are frequently generated in this way (see Chapter 6).
- **Reassortment:** In viruses with segmented genomes, the genome segments can be randomly shuffled during **superinfection**. Progeny viruses receive (at least) one of each of the genome segments, but probably not from a single parent. For example, influenza virus has eight genome segments; therefore, in a mixed infection, there could be $2^8 = 256$ possible progeny viruses. Packaging mechanisms in these viruses are not well understood (see Chapter 2) but may be involved in generating reassortants.

In intramolecular recombination, the probability that breakage-reunion or strand-switching will occur between two markers (resulting in recombination) is proportional to the physical distance between them. Pairs of genetic markers can be arranged on a linear map with distances measured in map units (i.e., percentage recombination frequency). In reassortment, the frequency of recombination between two markers is either very high (indicating that the markers are on two different genome segments) or comparatively low (which means that they are on the same segment). This is because the frequency of reassortment usually swamps the lower background frequency that is due to intermolecular recombination between strands.

Reactivation is the generation of infectious (recombinant) progeny from noninfectious parental virus genomes. This process has been demonstrated *in vitro* and may be important *in vivo*. For example, it has been suggested that the rescue of defective, long-dormant HIV **proviruses** during the long clinical course of acquired immune deficiency syndrome (AIDS) may result in increased antigenic diversity and contribute to the pathogenesis of the disease. Recombination occurs frequently in nature; for example, influenza virus reassortment has resulted in worldwide epidemics (**pandemics**) that have killed millions of people (Chapter 6). This makes these genetic interactions of considerable practical interest and not merely a dry academic matter.

Nongenetic interactions between viruses

A number of nongenetic interactions between viruses occur that can affect the outcome and interpretation of the results of genetic crosses. **Eukaryotic** cells have a diploid genome with two copies of each chromosome, each bearing its own allele of the same gene. The two chromosomes may differ in allelic

markers at many loci. Among viruses, only retroviruses are truly diploid, with two complete copies of the entire genome, but some DNA viruses, such as herpesviruses, have repeated sequences and are therefore partially heterozygous. In a few (mostly **enveloped**) viruses, aberrant packaging of multiple genomes may occasionally result in multiploid particles that are heterozygous (e.g., up to 10% of Newcastle disease virus particles). This process is known as **heterozygosis** and can contribute to the genetic complexity of virus populations.

Another commonly seen nongenetic interaction between viruses is interference. This process results from the resistance to **superinfection** by a virus observed in cells already infected by another virus. Homologous interference (i.e., against the same virus) often results from the presence of D.I. particles that compete for essential cell components and block replication. However, interference can also result from other types of mutations (e.g., dominant temperature-sensitive mutations) or by sequestration of virus **receptors** due to the production of **virus-attachment proteins** by viruses already present within the cell (e.g., in the case of avian retroviruses).

Phenotypic mixing can vary from extreme cases, where the genome of one virus is completely enclosed within the **capsid** or **envelope** of another (**pseudotyping**), to more subtle cases where the capsid/envelope of the progeny contains a mixture of proteins from both viruses. This mixing gives the progeny virus the phenotypic properties (e.g., cell **tropism**) dependent on the proteins incorporated into the particle, without any genetic change. Subsequent generations of viruses inherit and display the original parental phenotypes. This process can occur easily in viruses with naked capsids (nonenveloped) that are closely related (e.g., different strains of enteroviruses) or in enveloped viruses, which need not be related to one another (Figure 3.4). In this latter case, the phenomenon is due to the nonspecific incorporation of different virus glycoproteins into the envelope, resulting in a mixed phenotype. Rescue of replication-defective transforming retroviruses by a helper virus is a form of pseudotyping. Phenotypic mixing has proved to be a very useful tool to examine biological properties of viruses. Vesicular stomatitis virus (VSV) readily forms pseudotypes containing retrovirus envelope glycoproteins, giving a plaque-forming virus with the properties of VSV but with the cell tropism of the retrovirus. This trick has been used to study the cell tropism of HIV and other retroviruses.

Small DNA genomes

Bacteriophage M13 has already been mentioned in Chapter 2. The genome of this phage consists of 6.4 kb of single-stranded, positive-sense, circular DNA and encodes ten genes. Unlike most **icosahedral** virions, the filamentous M13 **capsid** can be expanded by the addition of further protein subunits, so the

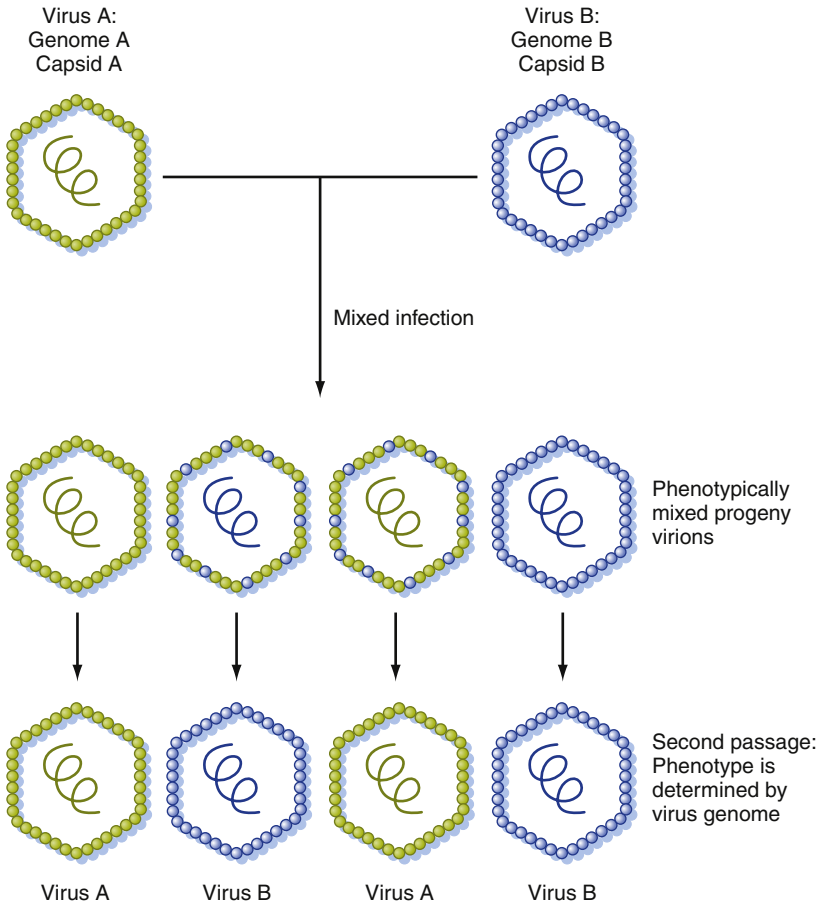

FIGURE 3.4 **Phenotypic mixing.**
Phenotypic mixing occurs in mixed infections, resulting in genetically unaltered virus particles that have some of the properties of the other parental type due to sharing a capsid.

genome length can be increased by the addition of extra sequences in the nonessential intergenic region without becoming incapable of being packaged into the capsid. In other bacteriophages, the genome packaging limits are more rigid. For example in phage λ, only DNA of between approximately 95 and 110% (approximately 46–54 kbp) of the normal genome size (49 kbp) can be packaged into the virus particle. Not all bacteriophages have such simple genomes as M13. The genome of phage T4 is about 160 kbp double-stranded DNA, and that of phage φKZ of *Pseudomonas aeruginosa* is 280 kbp.

T4 and λ also illustrate another common feature of linear virus genomes—the importance of the sequences present at the ends of the genome. In the case of

phage λ, the substrate packaged into the phage heads during assembly consists of long repetitive strings (concatemers) of phage DNA that are produced during the later stages of genome replication. The DNA is reeled in by the phage head, and when a complete genome length has been incorporated, the DNA is cleaved at a specific sequence by a phage-coded endonuclease (Figure 3.5). This enzyme leaves a 12-bp 5′ overhang on the end of each of the cleaved strands,

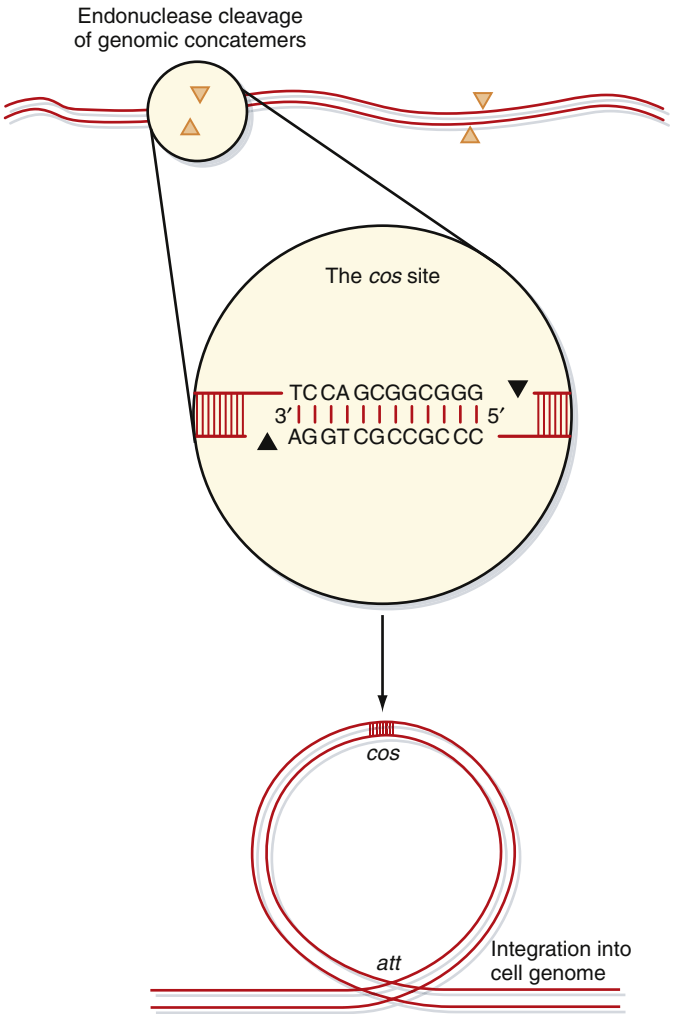

FIGURE 3.5 Integration of the bacteriophage λ genome.
The cohesive sticky ends of the *cos* site in the bacteriophage λ genome are ligated together in newly infected cells to form a circular molecule. Integration of this circular form into the *Escherichia coli* chromosome occurs by specific recognition and cleavage of the *att* site in the phage genome.

known as the *cos* site. Hydrogen bond formation between these sticky ends can result in the formation of a circular molecule. In a newly infected cell, the gaps on either side of the *cos* site are closed by DNA ligase, and it is this circular DNA that undergoes vegetative replication or integration into the bacterial chromosome. Phage T4 illustrates another molecular feature of certain linear virus genomes—**terminal redundancy**. Replication of the T4 genome also produces long concatemers of DNA. These are cleaved by a specific endonuclease, but unlike the λ genome the lengths of DNA incorporated into the particle are somewhat longer than a complete genome length (Figure 3.6). Some genes are repeated at each end of the genome, and the DNA packaged into the phage particles contains repeated information. These examples show that bacteriophage genomes are neither necessarily small nor simple!

As further examples of small DNA genomes, consider those of two groups of animal viruses: the parvoviruses and polyomaviruses. Parvovirus genomes are linear, nonsegmented, single-stranded DNA of about 5 kb. Parvovirus genomes are negative-sense, but some parvoviruses package equal amounts of (+) and (−) strands into **virions**. These are very small genomes, and even the replication-competent parvoviruses contain only two genes: *rep*, which encodes proteins involved in transcription; and *cap*, which encodes the coat proteins. However, the expression of these genes is rather complex, resembling the pattern seen in adenoviruses, with multiple **splicing** patterns seen for each

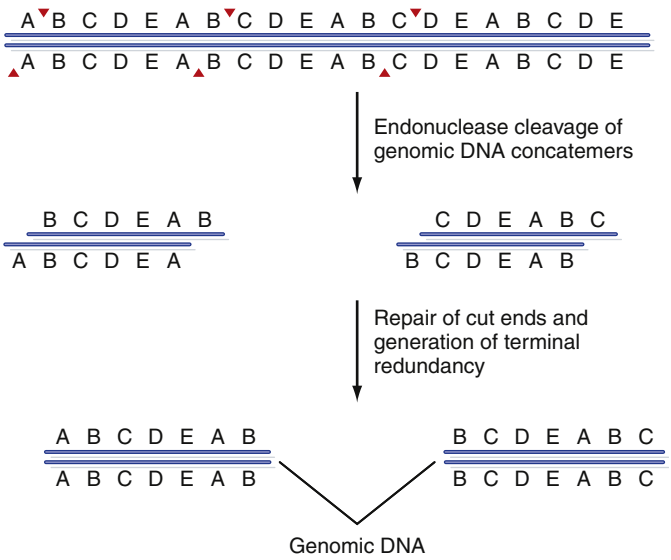

FIGURE 3.6 **Terminal redundancy.**
Terminal redundancy in the bacteriophage T4 genome results in the reiteration of some genetic information.

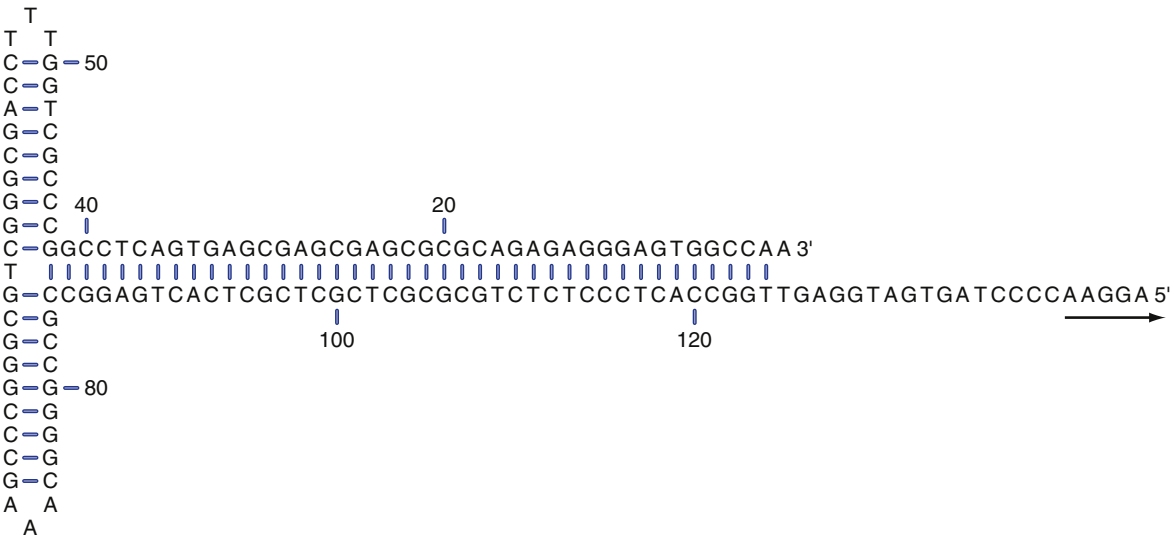

FIGURE 3.7 **Parvovirus hairpin sequences.**
Palindromic sequences at the ends of parvovirus genomes result in the formation of hairpin structures involved in the initiation of replication.

gene (Chapter 5). The ends of the genome have palindromic sequences of about 115 nt, which form "hairpins" (Figure 3.7). These structures are essential for the initiation of genome replication, again emphasizing the importance of the sequences at the ends of the genome, and also determine which strand is packaged into virus particles.

The genomes of polyomaviruses are double-stranded, circular DNA molecules of approximately 5 kbp. The architecture of the polyomavirus genome (i.e., number and arrangement of genes and function of the regulatory signals and systems) has been studied in great detail. Within the particles, the virus DNA assumes a supercoiled form and is associated with four cellular histones: H2A, H2B, H3, and H4 (see Chapter 2). The genomic organization of these viruses has evolved to pack the maximum information (six genes) into minimal space (5 kbp). This has been achieved by the use of both strands of the genome DNA and overlapping genes (Figure 3.8). VP1 is encoded by a dedicated open reading frame (ORF), but the VP2 and VP3 genes overlap so that VP3 is contained within VP2. The origin of replication is surrounded by noncoding regions that control transcription. Polyomaviruses also encode T-antigens, which are proteins that can be detected by sera from animals bearing polyomavirus-induced tumors. These proteins bind to the origin of replication and show complex activities in that they are involved both in DNA replication and in the transcription of virus genes. This topic is discussed further in Chapter 7.

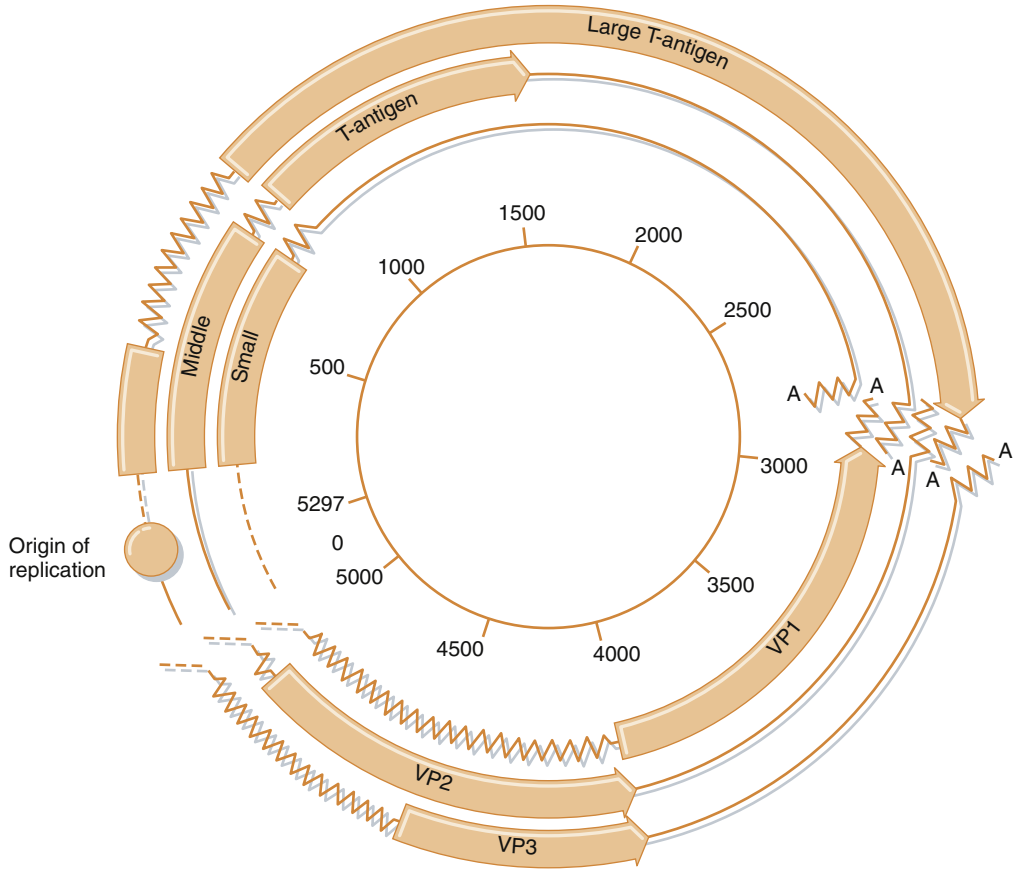

FIGURE 3.8 **Polyomavirus genome.**
The complex organization of the polyomavirus genome results in the compression of much genetic information into a relatively short sequence.

Large DNA genomes

A number of virus groups have double-stranded DNA genomes of considerable size and complexity. In many respects, these viruses are genetically similar to the host cells that they infect. The genomes of adenoviruses consist of linear, double-stranded DNA of 30 to 38 kbp, the precise size of which varies between different adenoviruses. These genomes contain 30 to 40 genes (Figure 3.9). The terminal sequence of each DNA strand is an inverted repeat of 100 to 140 bp, and the denatured single strands can form panhandle structures. These structures are important in DNA replication, as is a 55-kDa protein known as the terminal protein that is covalently attached to the 5′ end of each strand. During genome replication, this protein acts as a primer, initiating the synthesis of new DNA strands. Although adenovirus genomes are considerably smaller than

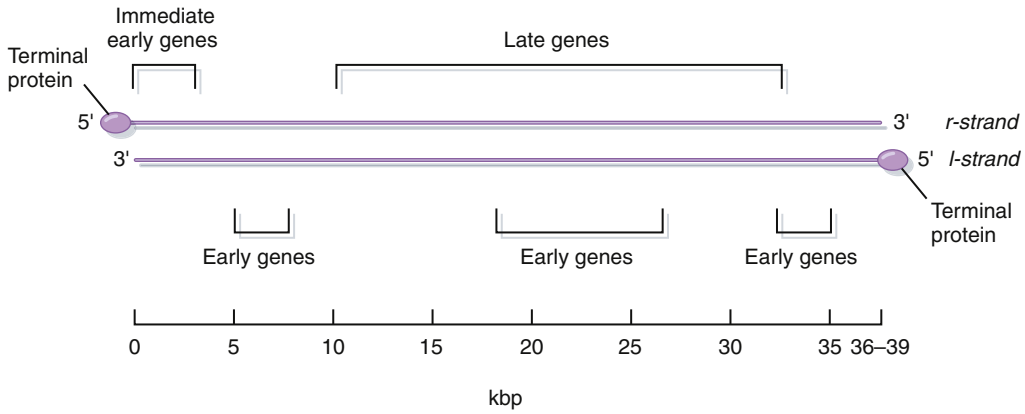

FIGURE 3.9 **Adenovirus genomes.**
Organization of the adenovirus genome.

those of herpesviruses, the expression of the genetic information is rather more complex. Clusters of genes are expressed from a limited number of shared **promoters**. Multiple spliced mRNAs and alternative **splicing** patterns are used to express a variety of polypeptides from each promoter (see Chapter 5).

The *Herpesviridae* is a large family containing more than 100 different members, at least one for most animal species that have been examined to date. There are eight human herpesviruses, all of which share a common overall genome structure but which differ in the fine details of genome organization and at the level of nucleotide sequence. The family is divided into three subfamilies, based on their nucleotide sequence and biological properties (Table 3.1). Herpesviruses have large genomes composed of up to 235 kbp of linear, double-stranded DNA and correspondingly large and complex virus particles containing about 35 virion polypeptides. All encode a variety of enzymes involved in nucleic acid metabolism, DNA synthesis, and protein processing (e.g., protein kinases). The different members of the family are widely separated in terms of genomic sequence and proteins, but all are similar in terms of structure and genome organization (Figure 3.10(a)).

Some but not all herpesvirus genomes consist of two covalently joined sections, a unique long (U_L) and a unique short (U_S) region, each bounded by inverted repeats. The repeats allow structural rearrangements of the unique region. This arrangement allows these genomes to exist as a mixture of four structural isomers, all of which are functionally equivalent (Figure 3.10(b)). Herpesvirus genomes also contain multiple repeated sequences and, depending on the number of these, the genome size of different isolates of a particular virus can vary by up to 10 kbp. The prototype member of the family is herpes simplex virus (HSV), whose genome consists of approximately 152 kbp of

Table 3.1 Human Herpesviruses

Alphaherpesvirinae	
Latent infections in sensory ganglia; genome size 120–180 kbp	
Simplexvirus	Human herpesviruses 1 and 2 (HSV-1, HSV-2)
Varicellovirus	Human herpesvirus 3 (VZV)
Betaherpesvirinae	
Restricted host range; genome size 140–235 kbp	
Cytomegalovirus	Human herpesvirus 5 (HCMV)
Roseolovirus	Human herpesviruses 6 and 7 (HHV-6, HHV-7)
Gammaherpesvirinae	
Infection of lymphoblastoid cells; genome size 105–175 kbp	
Lymphocryptovirus	Human herpesvirus 4 (EBV)
Rhadinovirus	Human herpesvirus 8 (HHV-8)

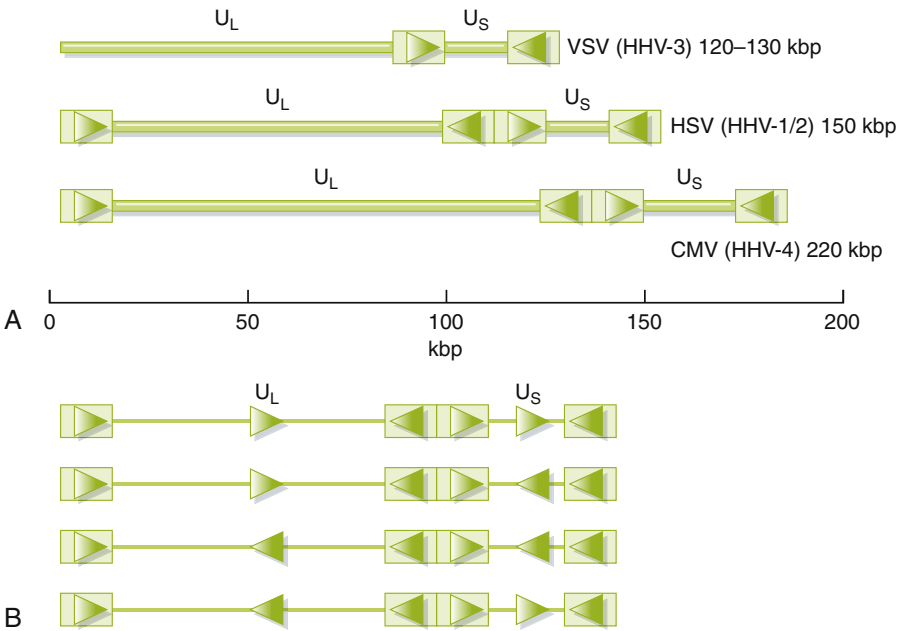

FIGURE 3.10 **Herpesvirus genomes.**
(a) Some herpesvirus genomes consist of two covalently joined sections, U_L and U_S, each bounded by inverted repeats. (b) This organization permits the formation of four different isomeric forms of the genome.

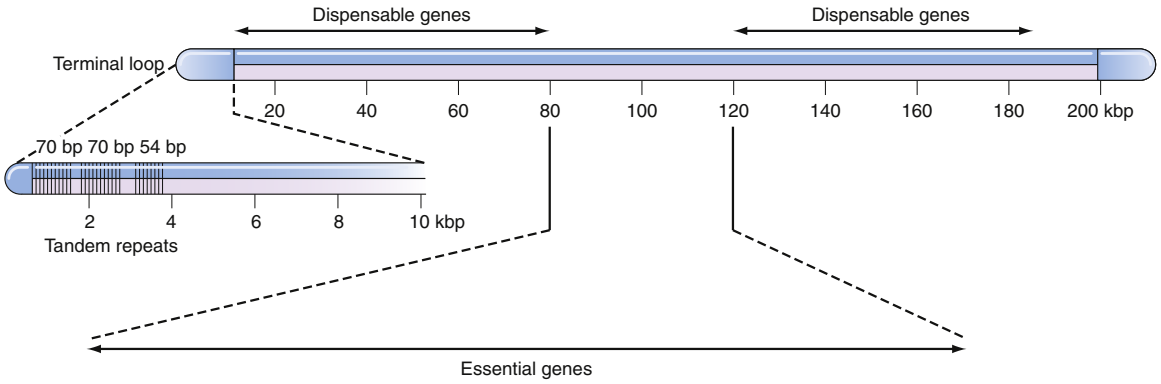

FIGURE 3.11 **Poxvirus genome organization.**
In these large and complex genomes, essential genes are located in the central region of the genome. Genes that are dispensable for replication in culture are located closer to the ends of the genome; sequences at the end of the strand contain many sequence repeats important for genome replication.

double-stranded DNA, the complete nucleotide sequence of which has now been determined. This virus contains about 80 genes, densely packed and with overlapping reading frames. Each gene is expressed from its own promoter (see adenovirus discussion earlier).

Poxvirus genomes are linear structures ranging in size from 140 to 290 kbp. As with the herpesviruses, each gene tends to be expressed from its own promoter. Characteristically, the central regions of poxvirus genomes tend to be highly conserved and to contain genes that are essential for replication in culture, while the outer regions of the genome are more variable in sequence and at least some of the genes located here are dispensable (Figure 3.11). In contrast, the noncoding nucleic acid structures at the ends of the genome are highly conserved and vital for replication. There are no free ends to the linear genome because these are closed by hairpin arrangements. Adjacent to the ends of the genome are other noncoding sequences that play vital roles in replication (see Chapter 4).

In the last few years, viruses with even larger DNA genomes have been discovered. The genomes of these viruses range from around 500 kbp in the case of Phycodnaviruses, up to 1.2 Mbp in the case of Mimivirus (see Chapter 2). Most of these viruses are aquatic and they are sometimes known as "giruses" (giant viruses), even though they are not closely related and infect a range of prokaryotic and eukaryotic hosts. There is some evidence however that at least some of these viruses may have evolved from a common ancestor.

Positive-strand RNA viruses

The ultimate size of single-stranded RNA genomes is limited by the relatively fragile nature of RNA and the tendency of long strands to break. In

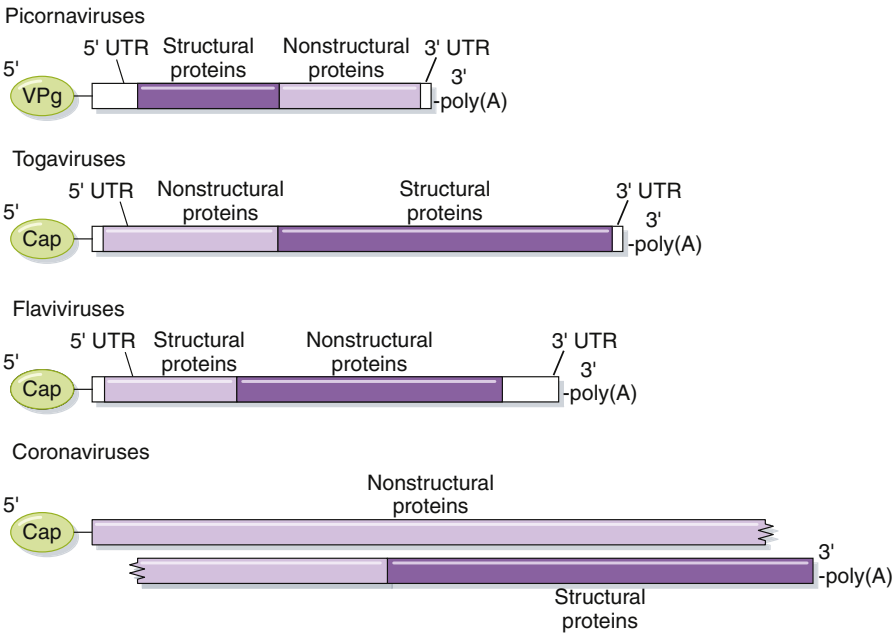

FIGURE 3.12 **Genomic organization of positive-stranded RNA viruses.**
This diagram illustrates some of the differences and similar features between different families of positive-stranded RNA viruses. Difference in patterns of gene expression are discussed in Chapter 5.

addition, RNA genomes tend to have higher mutation rates than those composed of DNA because they are copied less accurately, although the virus-encoded RNA-dependent RNA polymerases responsible for the replication of these genomes do have some repair mechanisms. These reasons tend to drive RNA viruses toward smaller genome sizes. Single-stranded RNA genomes vary in size from those of coronaviruses, which are approximately 30 kb long, to those of bacteriophages such as MS2 and Qβ, at about 3.5 kb. Although members of distinct families, most positive-sense RNA viruses of vertebrates share common features in terms of the biology of their genomes. In particular, purified positive-sense virus RNA is directly infectious when applied to susceptible host cells in the absence of any virus proteins (although it is about one million times less infectious than virus particles). On examining the features of these virus families, although the details of genomic organization vary, some repeated themes emerge (Figure 3.12).

Picornaviruses

The picornavirus genome consists of one single-stranded, positive-sense RNA molecule of between 7.2 kb in human rhinoviruses (HRVs) to 8.5 kb in

foot-and-mouth disease viruses (FMDVs), containing a number of features conserved in all picornaviruses:

- There is a long (600–1200 nt) untranslated region (UTR) at the 5′ end that is important in translation, virulence, and possibly encapsidation, as well as a shorter 3′ untranslated region (50–100 nt) that is necessary for negative-strand synthesis during replication.
- The 5′ UTR contains a clover-leaf secondary structure known as the internal ribosomal entry site (**IRES**) (Chapter 5).
- The rest of the genome encodes a single **polyprotein** of between 2100 and 2400 amino acids.

Both ends of the genome are modified—the 5′ end by a covalently attached small, basic protein VPg (23 amino acids), and the 3′ end by polyadenylation.

Togaviruses

The togavirus genome is comprised of single-stranded, positive-sense, non-segmented RNA of approximately 11.7 kb. It has the following features:

- It resembles cellular mRNAs in that it has a 5′ methylated cap and 3′ poly(A) sequences.
- Gene expression is achieved by two rounds of translation, producing first nonstructural proteins encoded in the 5′ part of the genome and later structural proteins from the 3′ part.

Flaviviruses

The flavivirus genome is a single-stranded, positive-sense RNA molecule of about 10.5 kb with the following features:

- It has a 5′ methylated cap, but in most cases the RNA is not polyadenylated at the 3′ end.
- Genetic organization differs from that of the togaviruses (previously) in that the structural proteins are encoded in the 5′ part of the genome and nonstructural proteins in the 3′ part.

Expression is similar to that of the picornaviruses, involving the production of a **polyprotein**.

Coronaviruses

The coronavirus genome consists of nonsegmented, single-stranded, positive-sense RNA, approximately 27 to 30 kb long, which is the longest of any RNA virus. It also has the following features:

- It has a 5′ methylated cap and 3′ poly(A), and the vRNA functions directly as mRNA.
- The 5′ 20-kb segment of the genome is translated first to produce a virus polymerase, which then produces a full-length negative-sense strand. This is

used as a template to produce mRNA as a nested set of transcripts, all with an identical 5′ nontranslated leader sequence of 72 nt and coincident 3′ polyadenylated ends.

- Each mRNA is **monocistronic**, with the genes at the 5′ end being translated from the longest mRNA and so on. These unusual cytoplasmic structures are produced not by **splicing** (posttranscriptional modification) but by the polymerase during transcription.

Positive-sense RNA plant viruses

The majority (but not all) of plant virus families have positive-sense RNA genomes. The genome of the tobamovirus tobacco mosaic virus (TMV) is a well-studied example (Figure 3.13):

- The TMV genome is a 6.4-kb RNA molecule that encodes four genes.
- There is a 5′ methylated cap, and the 3′ end of the genome contains extensive secondary structure but no poly(A) sequences.

Expression is reminiscent of but distinct from that of togaviruses, producing nonstructural proteins by direct translation of the open reading frame encoded in the 5′ part of the genome and the virus coat protein and further nonstructural proteins from two subgenomic RNAs encoded by the 3′ part. The similarities and differences between genomes in this class will be considered further in the discussion of virus evolution later and in Chapter 5.

Negative-strand RNA viruses

Viruses with **negative-sense** RNA genomes are more diverse than the positive-stranded RNA viruses discussed earlier. Because of the difficulties of

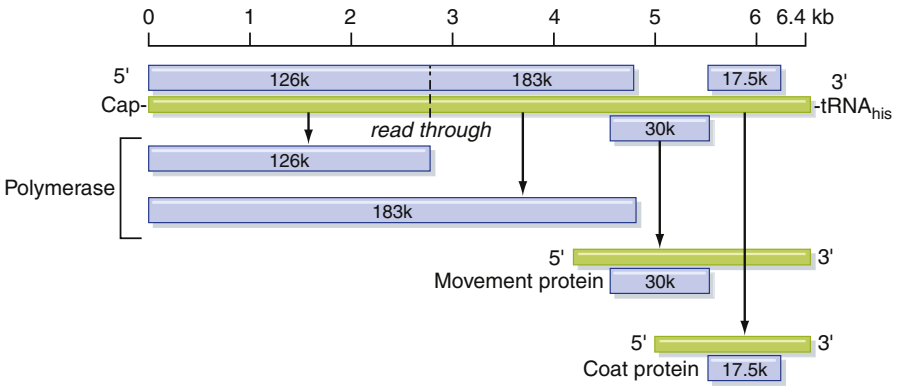

FIGURE 3.13 **Organization of the tobacco mosaic virus (TMV) genome.**
This positive-sense RNA plant virus expresses several genes via subgenomic messenger RNAs (see Chapter 5).

gene expression and genome replication, they tend to have larger genomes encoding more genetic information. Because of this, segmentation is a common, although not universal, feature of such viruses (Figure 3.14). None of these genomes is infectious as purified RNA. Although a gene encoding an RNA-dependent RNA polymerase has been found in some eukaryotic cells, most uninfected cells do not contain enough RNA-dependent RNA polymerase activity to support virus replication, and, because the negative-sense genome cannot be translated as mRNA without the virus polymerase packaged in each particle, these genomes are effectively inert. A few of the viruses described in this section are not strictly negative-sense but are **ambisense**, since they are part negative-sense and part positive-sense. Ambisense coding strategies occur in both plant viruses (e.g., the *Tospovirus* genus of the bunyaviruses, and tenuiviruses such as rice stripe virus) and animal viruses (the *Phlebovirus* genus of the bunyaviruses, and arenaviruses).

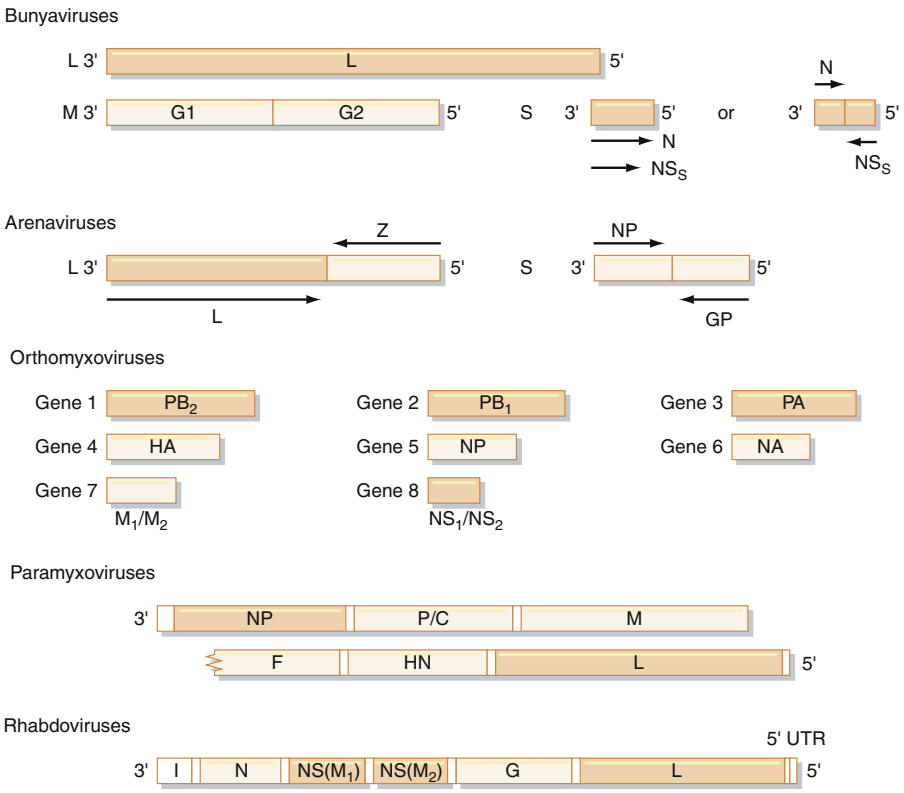

FIGURE 3.14 Genome organization of negative-stranded RNA viruses.
The fundamental distinction in the negative-strand RNA viruses is between those viruses with segmented genomes and those with nonsegmented genomes.

BOX 3.2. CAN'T MAKE YOUR MIND UP? DO BOTH!

Ambisense virus genomes contain at least one RNA segment that is part positive and part negative sense—in the same molecule. In spite of this, genetically they have more in common with negative-strand viruses than positive RNA viruses. But why on earth would any virus bother with such a complicated gene expression strategy? In general, it is more difficult for RNA viruses to control gene expression than it is for DNA viruses to upregulate and downregulate individual gene products. Most ambisense viruses can replicate in a range of hosts, such as mammals and insects or insects and plants. In their vector or reservoir host, infection is usually asymptomatic. However, in another host, multiplication of the virus can be lethal. Having two different strategies for gene expression may help them to successfully span this divide.

Bunyaviruses

Members of the *Bunyaviridae* have single-stranded, negative-sense, segmented RNA genomes with the following features:

- The genome is comprised of three molecules: L (8.5 kb), M (5.7 kb), and S (0.9 kb).
- All three RNA species are linear, but in the **virion** they appear circular because the ends are held together by base-pairing. The three segments are not present in virus preparations in equimolar amounts.
- In common with all negative-sense RNAs, the 5′ ends are not capped and the 3′ ends are not polyadenylated.

Members of the *Phlebovirus* and *Tospovirus* genera differ from the other three genera in the family (*Bunyavirus*, *Nairovirus*, and *Hantavirus*) in that genome segment S is rather larger and the overall genome organization is different—**ambisense** (i.e., the 5′ end of each segment is positive-sense, but the 3′ end is negative-sense). The *Tospovirus* genus also has an ambisense coding strategy in the M segment of the genome.

Arenaviruses

Arenavirus genomes consist of linear, single-stranded RNA. There are two genome segments: L (5.7 kb) and S (2.8 kb). Both segments have an ambisense organization, as the previous genome.

Orthomyxoviruses

See the discussion of segmented genomes in the next section.

Paramyxoviruses

Members of the *Paramyxoviridae* have nonsegmented negative-sense RNA of 15 to 16 kb. Typically, six genes are organized in a linear arrangement (3′—NP—P/C/V—M—F—HN—L—5′) separated by repeated sequences: a polyadenylation

signal at the end of the gene, an intergenic sequence (GAA), and a translation start signal at the beginning of the next gene.

Rhabdoviruses

Viruses of the *Rhabdoviridae* have nonsegmented, negative-sense RNA of approximately 11 kb. There is a leader region of approximately 50 nt at the 3′ end of the genome and a 60 nt untranslated region (UTR) at the 5′ end of the vRNA. Overall, the genetic arrangement is similar to that of paramyxoviruses, with a conserved polyadenylation signal at the end of each gene and short intergenic regions between the five genes.

Segmented and multipartite virus genomes

There is sometimes confusion between these two types of genome structures. Segmented virus genomes are those that are divided into two or more physically separate molecules of nucleic acid, all of which are then packaged into a single virus particle. In contrast, although multipartite genomes are also segmented, each genome segment is packaged into a separate virus particle. These discrete particles are structurally similar and may contain the same component proteins, but they often differ in size depending on the length of the genome segment packaged. In one sense, multipartite genomes are, of course, segmented, but this is not the strict meaning of these terms as they will be used here.

Segmentation of the virus genome has a number of advantages and disadvantages. There is an upper limit to the size of a nonsegmented virus genome that results from the physical properties of nucleic acids, particularly the tendency of long molecules to break due to shear forces (and, for each particular virus, the length of nucleic acid that can be packaged into the **capsid**). The problem of strand breakage is particularly relevant for single-stranded RNA, which is less chemically stable than double-stranded DNA. The longest single-stranded RNA genomes are those of the coronaviruses at approximately 30 kb, but the longest double-stranded DNA virus genomes are considerably longer (e.g., Mimivirus at up to 1.2 Mbp). Physical breakage of the genome results in biological inactivation, since it cannot be completely transcribed, translated, or replicated. Segmentation means that the virus avoids "having all its eggs in one basket" and also reduces the probability of breakages due to shearing, thus increasing the total potential coding capacity of the entire genome. However, the disadvantage of this strategy is that all the individual genome segments must be packaged into each virus particle or the virus will be defective as a result of loss of genetic information. In general, it is not understood how this control of packaging is achieved.

Separating the genome segments into different particles (the multipartite strategy) removes the requirement for accurate sorting but introduces a new problem in that all the discrete virus particles must be taken up by a single host

cell to establish a **productive infection**. This is perhaps the reason why multipartite viruses are found only in plants. Many of the sources of infection by plant viruses, such as inoculation by sapsucking insects or after physical damage to tissues, result in a large inoculum of infectious virus particles, providing opportunities for infection of an initial cell by more than one particle.

The genetics of segmented genomes are essentially the same as those of non-segmented genomes, with the addition of the **reassortment** of segments, as discussed earlier. Reassortment can occur whether the segments are packaged into a single particle or are in a multipartite configuration. Reassortment is a powerful means of achieving rapid generation of genetic diversity; this could be another possible reason for its evolution. Segmentation of the genome also has implications for the partition of genetic information and the way in which it is expressed, which will be considered further in Chapter 5.

To understand the complexity of these genomes, consider the organization of a segmented virus genome (influenza A virus) and a multipartite genome (geminivirus). The influenza virus genome is composed of eight segments (in influenza A and B strains; seven in influenza C) of single-stranded, negative-sense RNA (Table 3.2). The identity of the proteins encoded by each genome segment were determined originally by genetic analysis of the electrophoretic mobility of the individual segments from reassortant viruses and by analysis of

Table 3.2 Influenza Virus Genome Segments

Segment	Size (nt)	Polypeptides	Function (Location)
1	2341	PB_2	Transcriptase: cap binding
2	2341	PB_1	Transcriptase: elongation
3	2233	PA	Transcriptase: (?)
4	1778	HA	Haemagglutinin
5	1565	NP	Nucleoprotein: RNA binding; part of transcriptase complex
6	1413	NA	Neuraminidase
7	1027	M_1	Matrix protein: major component of virion
		M_2	Integral membrane protein—ion channel
8	890	NS_1	Nonstructural (nucleus): function unknown
		NS_2	Nonstructural (nucleus + cytoplasm): function unknown

a large number of mutants covering all eight segments. The eight segments have common nucleotide sequences at the 5′ and 3′ ends (Figure 3.15), which are necessary for replication of the genome (Chapter 4). These sequences are complementary to one another, and, inside the particle, the ends of the genome segments are held together by base-pairing and form a panhandle structure that again is believed to be involved in replication.

The RNA genome segments are not packaged as naked nucleic acid but in association with the gene 5 product, the nucleoprotein, and are visible in electron micrographs as **helical** structures. Here there is a paradox. Biochemically and genetically, each genome segment behaves as an individual, discrete entity; however, in electron micrographs of influenza virus particles disrupted with nonionic detergents, the **nucleocapsid** has the physical appearance of a single, long helix. Clearly, there is some interaction between the genome segments and it is this that explains the ability of influenza virus particles to select and package the genome segments within each particle with a surprisingly low error rate, considering the difficulty of the task (Chapter 2).

In many tropical and subtropical parts of the world, geminiviruses are important plant pathogens. Geminiviruses are divided into three genera based on their host plants (monocotyledons or dicotyledons) and insect vectors (leafhoppers or whiteflies). In the *Mastrevirus* and *Curtovirus* genera, the genome consists of a single-stranded DNA molecule of approximately 2.7 kb.

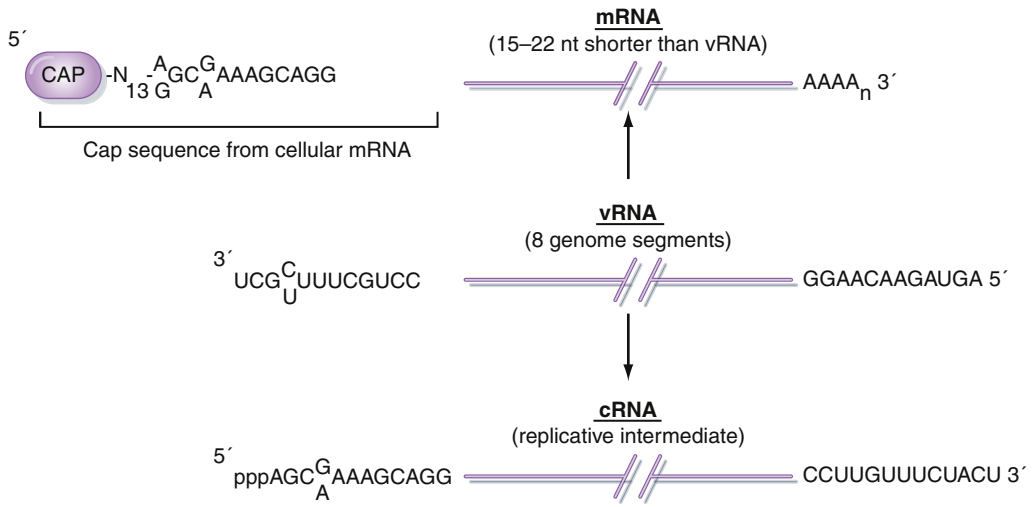

FIGURE 3.15 **Common terminal sequences of influenza RNAs.**
Influenza virus genome segments are a classic example of how sequences at the ends of liner virus genomes are crucial for gene expression and for replication.

The DNA packaged into these **virions** has been arbitrarily designated as positive-sense, although both the positive-sense and negative-sense strands found in infected cells contain protein-coding sequences. The genome of geminiviruses in the genus *Begmovirus* is bipartite and consists of two circular, single-stranded DNA molecules, each of which is packaged into a separate particle (Figure 3.16). Both of the strands comprising the genome are approximately 2.7 kb long and differ from one another completely in nucleotide sequence, except for a shared 200-nt noncoding sequence involved in DNA replication. The two genomic DNAs are packaged into entirely separate **capsids**. Because a **productive infection** requires both parts of the genome, it is necessary for a minimum of two virus particles bearing one copy of each of the genome segments to infect a new host cell. Although geminiviruses do not multiply in the tissues of their insect vectors (**nonpropagative transmission**), a sufficiently large amount of virus is ingested and subsequently deposited onto a new host plant to favor such **superinfections**.

Both of these examples show a high density of coding information. In influenza virus, genes 7 and 8 both encode two proteins in overlapping reading frames.

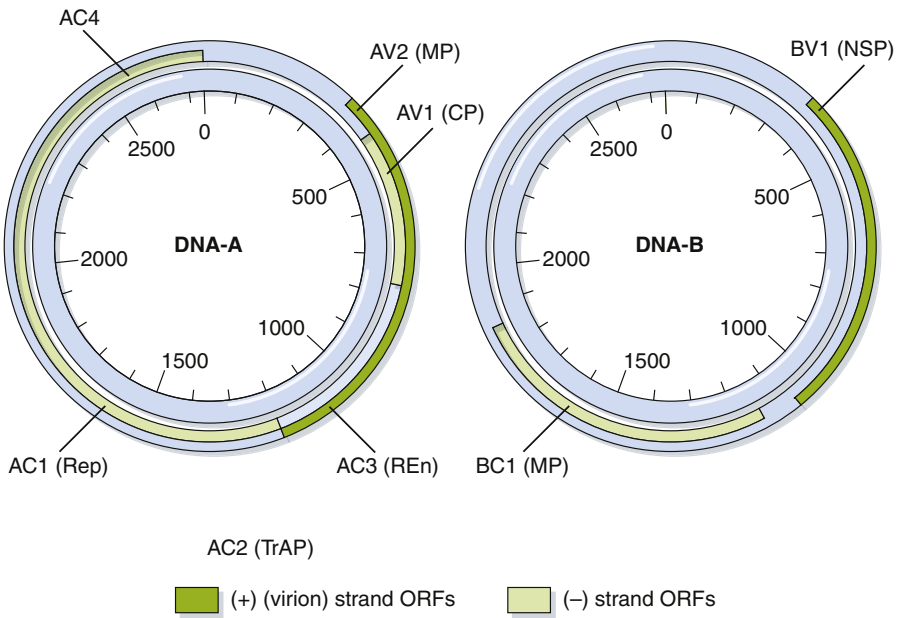

FIGURE 3.16 **Bipartite geminivirus genome.**
Organization and protein-coding potential of a bipartite begmovirus (*Geminiviridae*) genome. Proteins encoded by the positive-sense virion strand are named A or B depending on which of the two genome components they are located, or V or C depending on whether they are encoded by the positive-sense virion strand or the negative-sense complementary strand.

In geminiviruses, both strands of the virus DNA found in infected cells contain coding information, some of which is present in overlapping reading frames. It is possible that this high density of genetic information is the reason why these viruses have resorted to divided genomes, in order to regulate the expression of this information (see Chapter 5).

Reverse transcription and transposition

The first successes of molecular biology were the discovery of the double-helix structure of DNA and in understanding of the language of the genetic code. The importance of these findings does not lie in the mere chemistry but in their importance in allowing predictions to be made about the fundamental nature of living organisms. The confidence that flowed from these early triumphs resulted in the development of a grand universal theory, called the central dogma of molecular biology—namely, that all cells (and hence viruses) work on a simple organizing principle: the unidirectional flow of information from DNA, through RNA, into proteins. In the mid-1960s, however, there were rumblings that life might not be so simple.

In 1963, Howard Temin showed that the replication of retroviruses, whose particles contain RNA genomes, was inhibited by actinomycin D, an antibiotic that binds only to DNA. The replication of other RNA viruses is not inhibited by this drug. So pleased was the scientific community with an all-embracing dogma that these facts were largely ignored until 1970, when Temin and David Baltimore simultaneously published the observation that retrovirus particles contain an RNA-dependent DNA polymerase: reverse transcriptase. This finding alone was important enough, but like the earlier conclusions of molecular biology, it has subsequently had reverberations for the genomes of all organisms, and not merely a few virus families. It is now known that **retrotransposons** with striking similarities to retrovirus genomes form a substantial part of the genomes of all higher organisms, including humans. Earlier ideas of genomes as constant, stable structures have been replaced with the realization that they are, in fact, dynamic and rather fluid entities.

The concept of transposable genetic elements—specific sequences that are able to move from one position in the genome to another—was put forward by Barbara McClintock in the 1940s. Such **transposons** fall into two groups:

- Simple transposons, which do not undergo reverse transcription and are found in **prokaryotes** (e.g., the genome of enterobacteria phage Mu)
- Retrotransposons, which closely resemble retrovirus genomes and are bounded by long direct repeats (long terminal repeats, or LTRs); these move by means of a transcription/reverse transcription/integration mechanism and are found in **eukaryotes** (the *Metaviridae* and *Pseudoviridae*)

Both types show a number of similar properties:

- They are believed to be responsible for a high proportion of apparently spontaneous mutations.
- They promote a wide range of genetic rearrangements in host cell genomes, such as deletions, inversions, duplications, and translocations of the neighboring cellular DNA.
- The mechanism of insertion generates a short (3—13 bp) duplication of the DNA sequence on either side of the inserted element.
- The ends of the transposable element consist of inverted repeats, 2 to 50 bp long.
- Transposition is often accompanied by replication of the element—necessarily so in the case of **retrotransposons**, but this also often occurs with prokaryotic transposition.

Transposons control their own transposition functions, encoding proteins that act on the element in **cis** (affecting the activity of contiguous sequences on the same nucleic acid molecule) or in **trans** (encoding diffusible products acting on regulatory sites in any stretch of nucleic acid present in the cell). Bacteriophage Mu infects *E. coli* and consists of a complex, tailed particle containing a linear, double-stranded DNA genome of about 37 kb, with host-cell-derived sequences of between 0.5 and 2 kbp attached to the right-hand end of the genome (Figure 3.17). Mu is a **temperate bacteriophage** whose replication can proceed through two pathways; one involves integration of the genome into that of the host cell and results in **lysogeny**, and the other is **lytic** replication, which results in the death of the cell (see Chapter 5).

Integration of the phage genome into that of the host bacterium occurs at random sites in the cell genome. Integrated phage genomes are known as **prophage**, and integration is essential for the establishment of lysogeny. At intervals in bacterial cells lysogenic for Mu, the prophage undergoes transposition to a different site in the host genome. The mechanism leading to transposition is different from that responsible for the initial integration of

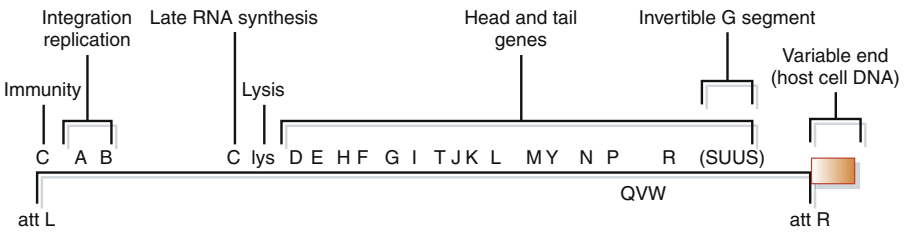

FIGURE 3.17 Bacteriophage Mu genome.
Organization of the phage Mu genome.

the phage genome (which is conservative in that it does not involve replication) and is a complex process requiring numerous phage-encoded and host-cell proteins. Transposition is tightly linked to replication of the phage genome and results in the formation of a co-integrate—that is, a duplicate copy of the phage genome flanking a target sequence in which insertion has occurred. The original Mu genome remains in the same location where it first integrated and is joined by a second integrated genome at another site. (Not all prokaryotic transposons use this process; some, such as TN10, are not replicated during transposition but are excised from the original integration site and integrate elsewhere.) There are two consequences of such a transposition. First, the phage genome is replicated during this process (advantageous for the virus), and second, the sequences flanked by the two phage genomes (which form repeated sequences) are at risk of secondary rearrangements, including deletions, inversions, duplications, and translocations (possibly but not necessarily deleterious for the host cell).

The yeast Ty viruses are representative of a class of sequences found in yeast and other **eukaryotes** known as **retrotransposons**. Unlike enterobacteria phage Mu, such elements are not true viruses but do bear striking similarities to retroviruses. The genomes of most strains of *Saccharomyces cerevisiae* contain 30 to 35 copies of the Ty elements, which are around 6 kbp long and contain direct repeats of 245 to 371 bp at each end (Figure 3.18). Within this repeat sequence is a **promoter** that leads to the transcription of a terminally redundant 5.6-kb mRNA. This contains two genes: TyA, which has homology to the *gag* gene of retroviruses, and TyB, which is homologous to the *pol* gene. The protein encoded by TyA is capable of forming a roughly spherical, 60-nm diameter, virus-like particle (VLP). The 5.6-kb RNA transcript can be incorporated into

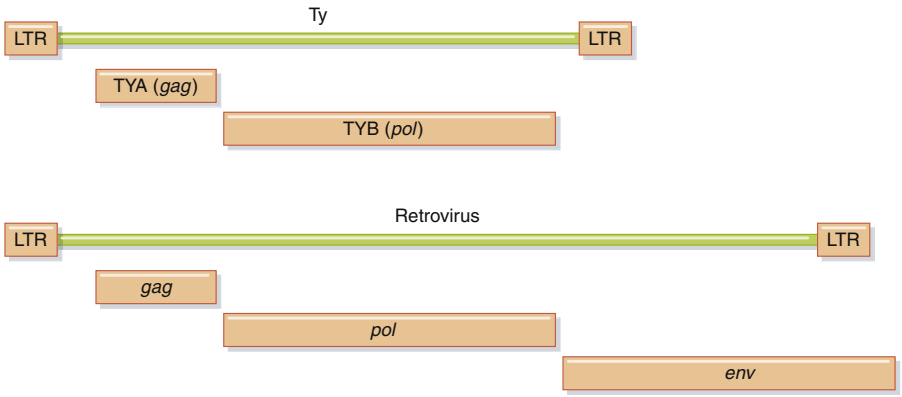

FIGURE 3.18 **Retrotransposons and retroviruses.**
The genetic organization of retrotransposons such as Ty (above) and retrovirus genomes (below) shows a number of similarities, including the presence of direct long terminal repeats (LTRs) at either end.

such particles, resulting in the formation of intracellular structures known as Ty-VLPs. Unlike true viruses these particles are not infectious for yeast cells, but if accidentally taken up by a cell they can carry out reverse transcription of their RNA content to form a double-stranded DNA Ty element, which can then integrate into the host cell genome (see later).

The most significant difference between retrotransposons such as Ty, copia (a similar element found in *Drosophila melanogaster*), and retroviruses proper is the presence of an additional gene in retroviruses, *env*, which encodes an **envelope** glycoprotein (see Chapter 2). The envelope protein is responsible for **receptor** binding and has allowed retroviruses to escape the intracellular lifestyle of retrotransposons to form a true virus particle and propagate themselves widely by infection of other cells (Figure 3.18). Retrovirus genomes have four unique features:

- They are the only viruses that are truly diploid.
- They are the only RNA viruses whose genome is produced by cellular transcriptional machinery (without any participation by a virus-encoded polymerase).
- They are the only viruses whose genome requires a specific cellular RNA (tRNA) for replication.
- They are the only positive-sense RNA viruses whose genome does not serve directly as mRNA immediately after infection.

During the process of reverse transcription (Figure 3.19), the two single-stranded positive-sense RNA molecules that comprise the virus genome are converted into a double-stranded DNA molecule somewhat longer than the RNA templates due to the duplication of direct repeat sequences at each end—the long terminal repeats (LTRs; Figure 3.20). Some of the steps in reverse transcription have remained mysteries—for example, the apparent jumps that the polymerase makes from one end of the template strand to the other. In fact, these steps can be explained by the observation that complete conversion of retrovirus RNA into double-stranded DNA only occurs in a partially uncoated core particle and cannot be duplicated accurately *in vitro* with the reagents free in solution. This indicates that the conformation of the two RNAs inside the retrovirus **nucleocapsid** dictates the course of reverse transcription—the jumps are nothing of the sort, since the ends of the strands are probably held adjacent to one another inside the core.

Reverse transcription has important consequences for retrovirus genetics. First, it is a highly error-prone process, because reverse transcriptase does not carry out the proofreading functions performed by cellular DNA-dependent polymerases. This results in the introduction of many mutations into retrovirus genomes and, consequently, rapid genetic variation (see "Spontaneous Mutations," earlier). In addition, the process of reverse transcription promotes

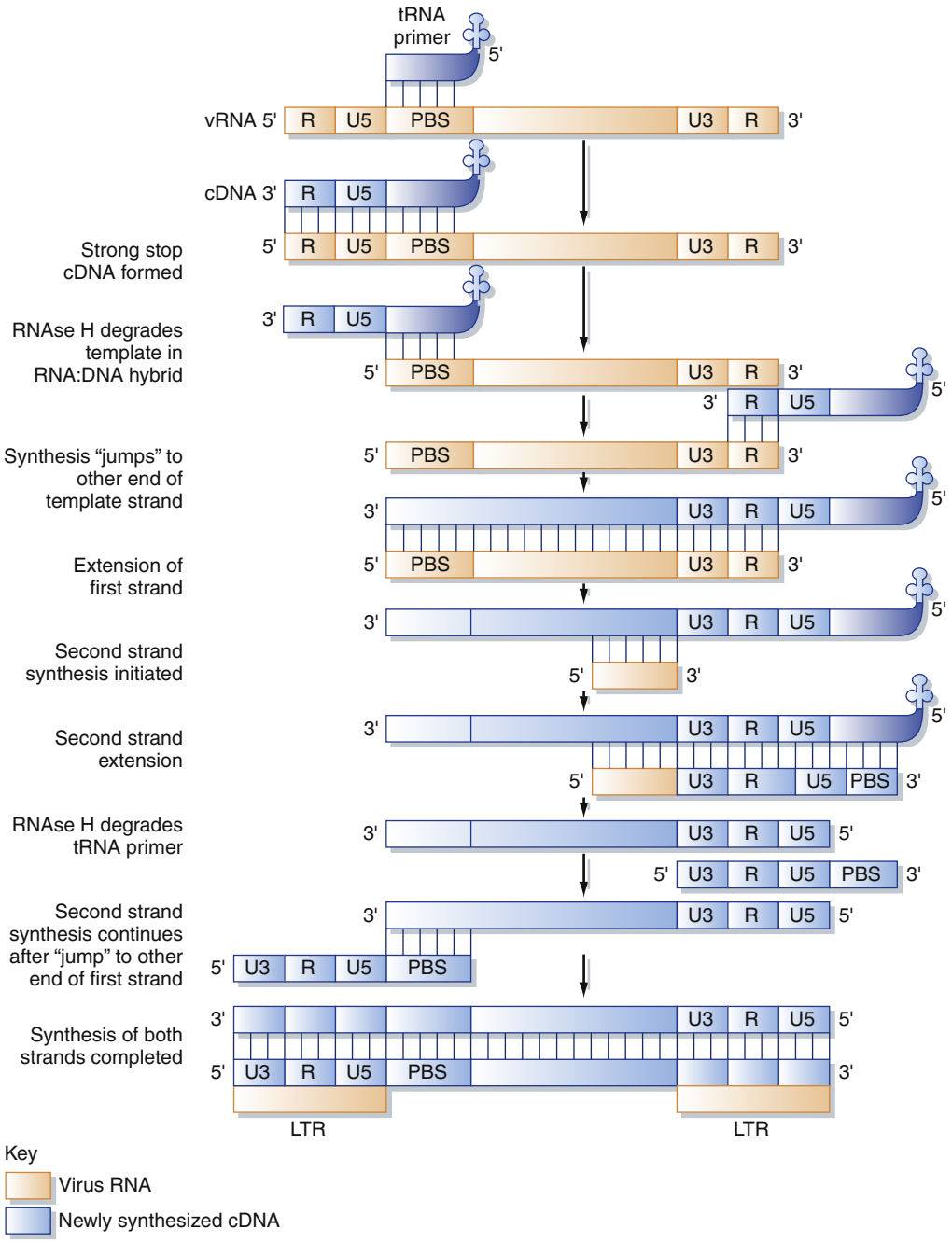

FIGURE 3.19 Reverse transcription.

Mechanism of reverse transcription of retrovirus RNA genomes, in which two molecules of RNA are converted into a single (terminally redundant) double-stranded DNA provirus.

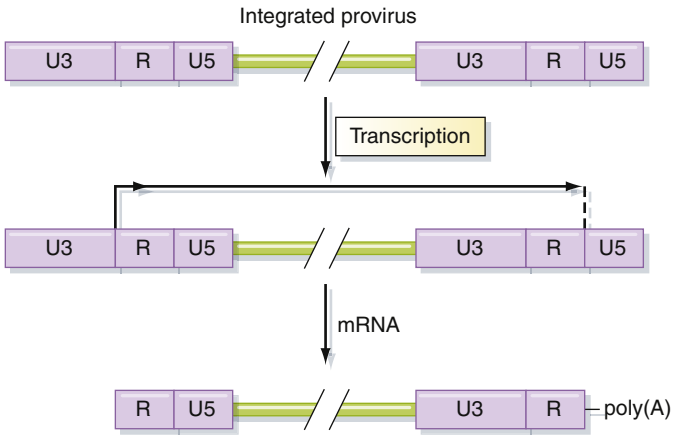

FIGURE 3.20 **Long terminal repeats.**
Generation of repeated information in retrovirus long terminal repeats (LTRs). In addition to their role in reverse transcription, these sequences contain important control elements involved in the expression of the virus genome, including a transcriptional promoter in the U3 region and polyadenylation signal in the R region.

genetic **recombination**. Because two RNAs are packaged into each **virion** and used as the template for reverse transcription, recombination can and does occur between the two strands. Although the mechanism responsible for this is not clear, if one of the RNA strands differs from the other (for example, by the presence of a mutation) and recombination occurs, then the resulting virus will be genetically distinct from either of the parental viruses.

After reverse transcription is complete, the double-stranded DNA migrates into the nucleus, still in association with virus proteins. The mature products of the *pol* gene are, in fact, a complex of polypeptides that include three distinct enzymatic activities: reverse transcriptase and RNAse H, which are involved in reverse transcription, and integrase, which catalyses integration of virus DNA into the host cell **chromatin**, after which it is known as the **provirus** (Figure 3.21). Three forms of double-stranded DNA are found in retrovirus-infected cells following reverse transcription: linear DNA and two circular forms that contain either one or two LTRs. From the structure at the ends of the provirus, it was previously believed that the two-LTR circle was the form used for integration. In recent years, systems that have been developed to study the integration of retrovirus DNA *in vitro* show that it is the linear form that integrates. This discrepancy can be resolved by a model in which the ends of the two LTRs are held in close proximity by the reverse transcriptase—integrase complex. The net result of integration is that 1 to 2 bp are lost from the end of each LTR and 4 to 6 bp of cellular DNA are duplicated on either side of the provirus. It is unclear whether there is any specificity regarding the site of integration into

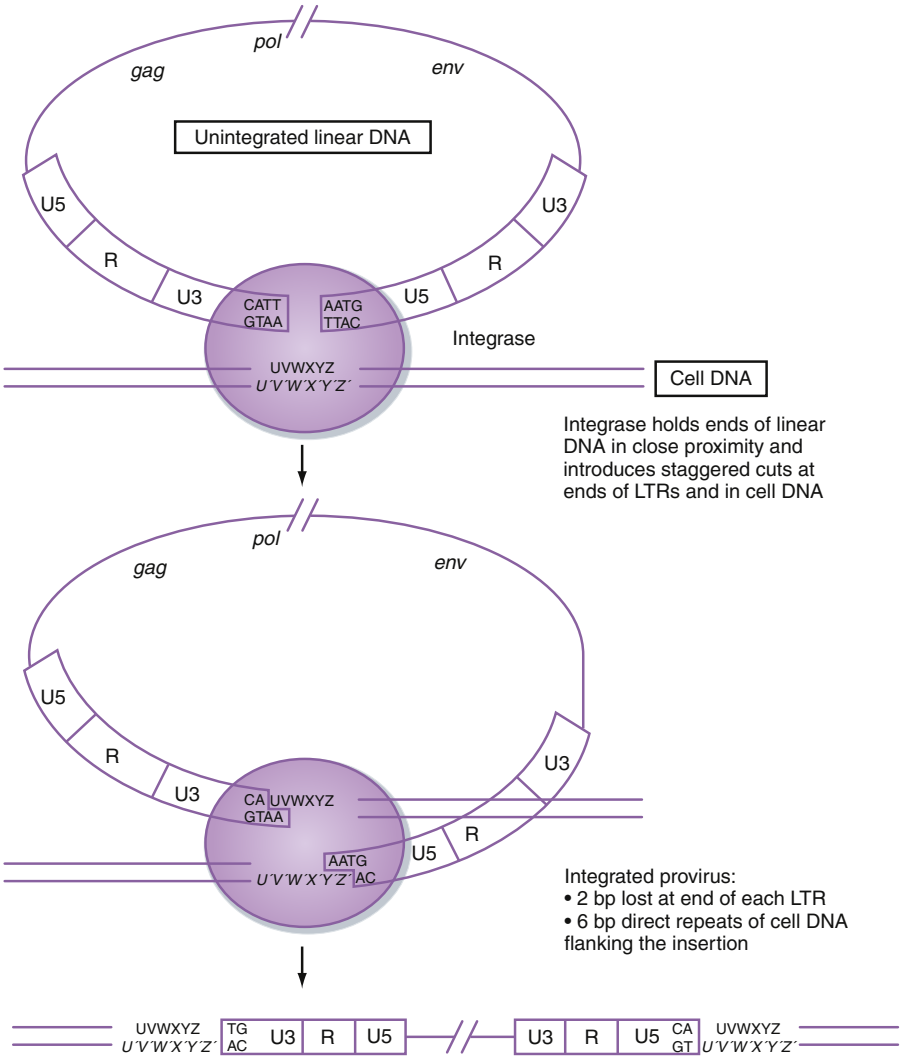

FIGURE 3.21 **Retrovirus integration.**
Mechanism of integration of retrovirus genomes into the host cell chromatin.

the cell genome. What is obvious is that there is no simple target sequence, but it is possible that there may be (numerous) regions or sites in the eukaryotic genome that are more likely to be integration sites than others.

Following integration, the DNA provirus genome becomes essentially a collection of cellular genes and is at the mercy of the cell for expression. There is no mechanism for the precise excision of integrated proviruses, some of which are known to have been fossilized in primate genomes through millions of

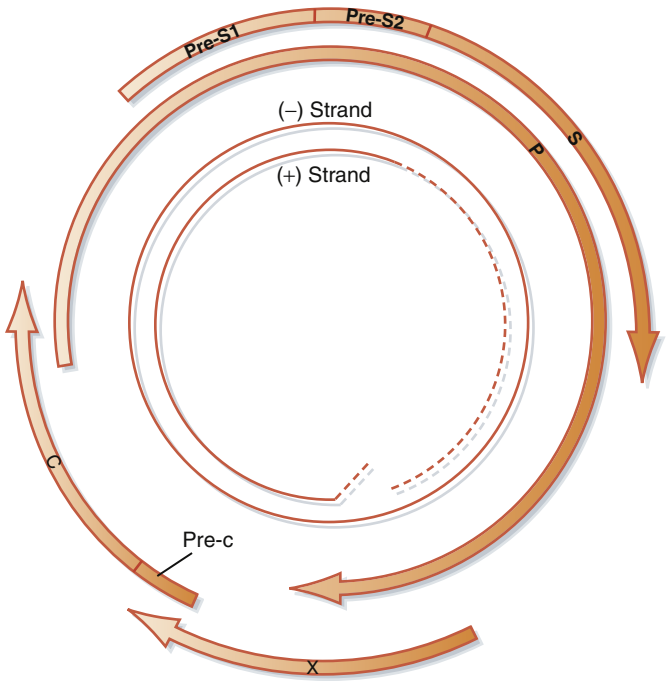

FIGURE 3.22 **Hepatitis B virus (HBV) genome.**
Structure, organization, and proteins encoded by the hepatitis B virus (HBV) genome.

years of evolution, although proviruses may sometimes be lost or altered by modifications of the cell genome. The only way out for the virus is transcription, forming what is essentially a full-length mRNA (minus the terminally redundant sequences from the LTRs). This RNA is the vRNA, and two copies are packaged into **virions** (Figure 3.20).

There are, however, two different groups of viruses whose replication involves reverse transcription. It is at this point that the difference between them becomes obvious. One strategy, as used by retroviruses and described earlier, culminates in the packaging of RNA into virions as the virus genome. The other, used by hepadnaviruses and caulimoviruses, switches the RNA and DNA phases of replication and results in DNA virus genomes inside virus particles. This is achieved by utilizing reverse transcription, not as an early event in replication as retroviruses do, but as a late step during formation of the virus particle.

Hepatitis B virus (HBV) is the prototype member of the family *Hepadnaviridae*. HBV **virions** are spherical, lipid-containing particles, 42 to 47 nm in diameter, which contain a partially double-stranded (gapped) DNA genome, plus an RNA-dependent DNA polymerase (i.e., reverse transcriptase; Figure 3.22).

Hepadnaviruses have very small genomes consisting of a negative-sense strand of 3.0 to 3.3 kb (varies between different hepadnaviruses) and a positive-sense strand of 1.7 to 2.8 kb (varies between different particles). On infection of cells, three major genome transcripts are produced: 3.5-, 2.4-, and 2.1-kb mRNAs. All have the same polarity (i.e., are transcribed from the same strand of the virus genome) and the same 3′ ends but have different 5′ ends (i.e., initiation sites). These transcripts are heterogeneous in size, and it is not completely clear which proteins each transcript encodes, but there are four known genes in the virus:

- C encodes the core protein.
- P encodes the polymerase.
- S encodes the three polypeptides of the surface antigen: pre-S1, pre-S2, and S (which are derived from alternative start sites).
- X encodes a transactivator of virus transcription (and possibly cellular genes).

Closed circular DNA is found soon after infection in the nucleus of the cell and is probably the source of these transcripts. This DNA is produced by repair of the gapped virus genome as follows:

- Completion of the positive-sense strand
- Removal of a protein primer from the negative-sense strand and an oligoribonucleotide primer from the positive-sense strand
- Elimination of **terminal redundancy** at the ends of the negative-sense strand
- Ligation of the ends of the two strands

The 3.5-kb RNA transcript, core antigen, and polymerase form core particles, and the polymerase converts the RNA to DNA in the particles as they form in the cytoplasm.

The genome structure and replication of cauliflower mosaic virus (CaMV), the prototype member of the *Caulimovirus* genus, is reminiscent of that of hepadnaviruses, although there are differences between them. The CaMV genome consists of a gapped, circular, double-stranded DNA molecule of about 8 kbp, one strand of which is known as the α-strand and contains a single gap, and a complementary strand, which contains two gaps (Figure 3.23). There are eight genes encoded in this genome, although not all eight products have been detected in infected cells. Replication of the CaMV genome is similar to that of HBV. The first stage is the migration of the gapped virus DNA to the nucleus of the infected cell where it is repaired to form a covalently closed circle. This DNA is transcribed to produce two polyadenylated transcripts, one long (35S) and one shorter (19S). In the cytoplasm, the 19S mRNA is translated to produce a protein that forms large **inclusion bodies** in the cytoplasm of infected cells,

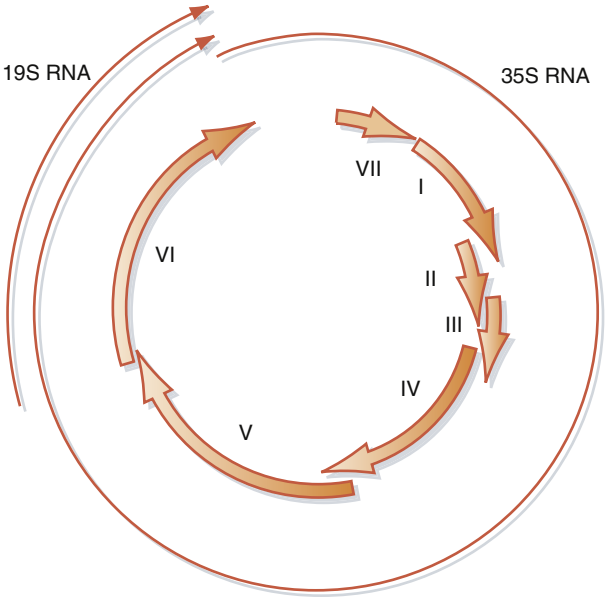

FIGURE 3.23 **Cauliflower mosaic virus (CaMV) genome.**
Structure, organization, and proteins encoded by the cauliflower mosaic virus (CaMV) genome.

and it is in these sites that the second phase of replication occurs. In these replication complexes, some copies of the 35S mRNA are translated while others are reverse transcribed and packaged into **virions** as they form. The differences between reverse transcription of these virus genomes and those of retroviruses are summarized in Table 3.3.

Evolution and epidemiology

Epidemiology is concerned with the distribution of disease and the developing strategies to reduce or prevent it. Virus infections present considerable difficulties for this process. Except for **epidemics** where acute symptoms are obvious, the

Table 3.3 Reverse Transcription of Virus Genomes

Features	Caulimoviruses	Hepadnaviruses	Retroviruses
Genome	DNA	DNA	RNA
Primer for (−)strand synthesis	tRNA	Protein	tRNA
Terminal repeats (LTRs)	No	No	Yes
Specific integration of virus genome	No	No	Yes

major evidence of virus infection available to the epidemiologist is the presence of antivirus antibodies in patients. This information frequently provides an incomplete picture, and it is often difficult to assess whether a virus infection occurred recently or at some time in the past. Techniques such as the isolation of viruses in experimental plants or animals, are laborious and impossible to apply to large populations. Although the use of PCR for virus detection is growing, it still lags behind standard serological methods of diagnosis. Molecular biology provides sensitive, rapid, and sophisticated techniques to detect and analyze the genetic information stored in virus genomes and has resulted in a new area of investigation: molecular epidemiology.

One drawback of molecular genetic analysis is that some knowledge of the nature of a virus genome is necessary before it can be investigated. However, we now possess a great deal of information about the structure and nucleotide sequences at least representative of some members of the known virus groups. This information allows virologists to look in two directions: back to where viruses came from and forward to chart the course of future epidemics and diseases. Sensitive detection of nucleic acids by amplification techniques such as the polymerase chain reaction is already having a major impact on this type of epidemiological investigation.

At least three theories seek to explain the origin of viruses:

- **Regressive evolution:** This theory states that viruses are degenerate life forms that have lost many functions that other organisms possess and have only retained the genetic information essential to their parasitic way of life.
- **Cellular origins:** In this theory, viruses are thought to be subcellular, functional assemblies of macromolecules that have escaped their origins inside cells.
- **Independent entities:** This theory suggests that viruses evolved on a parallel course to cellular organisms from the self-replicating molecules believed to have existed in the primitive, prebiotic RNA world.

Similarities in the coat protein structures of archaeal viruses and those of eubacterial and animal virus suggest that at least some present-day viruses may have a common ancestor that precedes the division into three domains of life over three billion years ago, suggesting that viruses have lineages that can be traced back to near the root of the universal tree of life. While each of these theories has its devotees and this subject provokes fierce disagreements, the fact is that viruses exist, and we are all infected with them. The practical importance of the origin of viruses is that this issue may have implications for virology here and now. Genetic and nucleotide sequence relationships between viruses can reveal the origins not only of individual viruses, but also of whole families and possible superfamilies. In a number of groups of viruses previously thought to be unrelated, genome sequencing

has revealed that functional regions appear to be grouped together in a similar way. The extent to which there is any sequence similarity between these regions in different viruses varies, although clearly the active sites of enzymes such as virus **replicases** are strongly conserved. The emphasis in these groups is more on functional and organizational similarities. The original classification scheme for viruses did not recognize a higher level grouping than the family (see Appendix 2 WEB), but there are now six groups of related virus families equivalent to the orders of formal biological nomenclature (Table 3.4).

BOX 3.3. WHAT DO ORDERS TELL US ABOUT EVOLUTION?

When the International Committee on Nomenclature of Viruses (ICTV) was created in 1966, we knew hundreds of viruses but little about most of them. This made it difficult to see how they were related to each other. Eventually it was agreed that some viruses were sufficiently similar to allow them to be grouped together as a genus—in the same way that horses (*Equus caballus*) are in the same genus as donkeys (*Equus asinus*). The next step was to group similar genera (plural of genus) together as families. At that point, there was a pause for some years until it was agreed that similar virus families could be grouped into orders, of which six have now been recognized. This change happened after enough nucleotide sequence data had been accumulated to make the faint evolutionary relationships between distantly related viruses apparent. Why does it matter? In part because this is a window on the past allowing us to look back millions of years through these genetic fossils, but much more importantly because it points to what viruses are capable of and where they might be going in the future. And that's something we should all worry about.

Knowledge drawn from taxonomic relationships allows us to predict the properties and behavior of new viruses or to develop drugs based on what is already known about existing viruses. It is believed these shared patterns suggest the descent of present-day viruses from a limited number of primitive ancestors. Although it is tempting to speculate on events that may have occurred before the origins of life as it is presently recognized, it would be unwise to discount the pressures that might result in viruses with diverse origins assuming common genetic solutions to common problems of storing, replicating, and expressing genetic information. This is particularly true now that we appreciate the plasticity of virus and cellular genomes and the mobility of genetic information from virus to virus, cell to virus, and virus to cell. There is no reason to believe that virus evolution has stopped, and it is dangerous to do so. The practical consequences of ongoing evolution and the concept of **emergent viruses** are described in Chapter 7.

Table 3.4 Orders of Related Virus Families

Order	Families
Caudovirales (tailed bacteriophages)	*Myoviridae, Podoviridae, Siphoviridae*
Herpesvirales (herpesvirus-like)	*Alloherpesviridae, Herpesviridae, Malacoherpesviridae*
Mononegavirales (nonsegmented negative-sense RNA viruses)	*Bornaviridae, Filoviridae, Paramyxoviridae, Rhabdoviridae*
Nidovirales (nested viruses, because of their pattern of transcription)	*Arteriviridae, Coronaviridae, Roniviridae*
Picornavirales (picornavirus-like)	*Dicistroviridae, Iflaviridae, Marnaviridae, Picornaviridae, Secoviridae*
Tymovirales (tymovirus-like)	*Alphaflexiviridae, Betaflexiviridae, Gammaflexiviridae, Tymoviridae*

SUMMARY

Molecular biology has put much emphasis on the structure and function of the virus genome. At first sight, this tends to emphasize the tremendous diversity of virus genomes. On closer examination, similarities and unifying themes become more apparent. Sequences and structures at the ends of virus genomes are in some ways functionally more significant than the unique coding regions within them. Common patterns of genetic organization seen in virus superfamilies suggest that many viruses have evolved from common ancestors and that exchange of genetic information between viruses has resulted in common solutions to common problems.

Further Reading

Barr, J.N., Fearns, R., 2010. How RNA viruses maintain their genome integrity. J. Gen. Virol. 91 (6), 1373–1387.

Beck, J., Nassal, M., 2007. Hepatitis B virus replication. World J. Gastroenterol. 13 (1), 48–64.

Bieniasz, P.D., 2009. The cell biology of HIV-1 virion genesis. Cell Host Microbe. 5 (6), 550–558. doi:10.1016/j.chom.2009.05.015.

Craig, N.L., et al., 2002. Mobile DNA. ASM Press, Washington, D.C.

Domingo, E., Webster, R.G., Holland, J.J., 2000. Origin and Evolution of Viruses. Academic Press, San Diego, CA.

Forterre, P., Prangishvili, D., 2009. The great billion-year war between ribosome- and capsid-encoding organisms (cells and viruses) as the major source of evolutionary novelties. Ann. N. Y. Acad. Sci. 1178, 65–77. doi: 10.1111/j.1749-6632.2009.04993.x.

Hutchinson, E.C., von Kirchbach, J.C., Gog, J.R., Digard, P., 2010. Genome packaging in influenza A virus. J. Gen. Virol. 91, 313–328. doi 10.1099/vir.0.017608-0.

Mertens, P., 2004. The dsRNA viruses. Virus Res. 101, 3–13.

Miller, E.S., et al., 2003. Bacteriophage T4 genome. Microbiol. Mol. Biol. Rev. 67, 86–156.

Moya, A., et al., 2004. The population genetics and evolutionary epidemiology of RNA viruses. Nat. Rev. Microbiol. 2, 279–288.

Nguyen, M., Haenni, A.L., 2003. Expression strategies of ambisense viruses. Virus Res. 93, 141–150. doi: 10.1016/S0168-1702(03)00094-7.

Raoult, D., et al., 2004. The 1.2-megabase genome sequence of Mimivirus. Science 306, 1344–1350.

Rice, G., et al., 2004. The structure of a thermophilic archaeal virus shows a double-stranded DNA viral capsid type that spans all domains of life. Proc. Nat. Acad. Sci. U. S. A. 101, 7716–7720.

Steinhauer, D.A., Skehel, J.J., 2002. Genetics of influenza viruses. Annu. Rev. Genet. 36, 305–332.

Van Etten, J.L., Lane, L.C., Dunigan, D.D., 2010. DNA Viruses: The Really Big Ones (Giruses). Annu. Rev. Microbiol. 64, 83–99. doi: 10.1146/annurev.micro.112408.134338.

Wagner, M., et al., 2002. Herpesvirus genetics has come of age. Trends Microbiol. 10, 318–324.