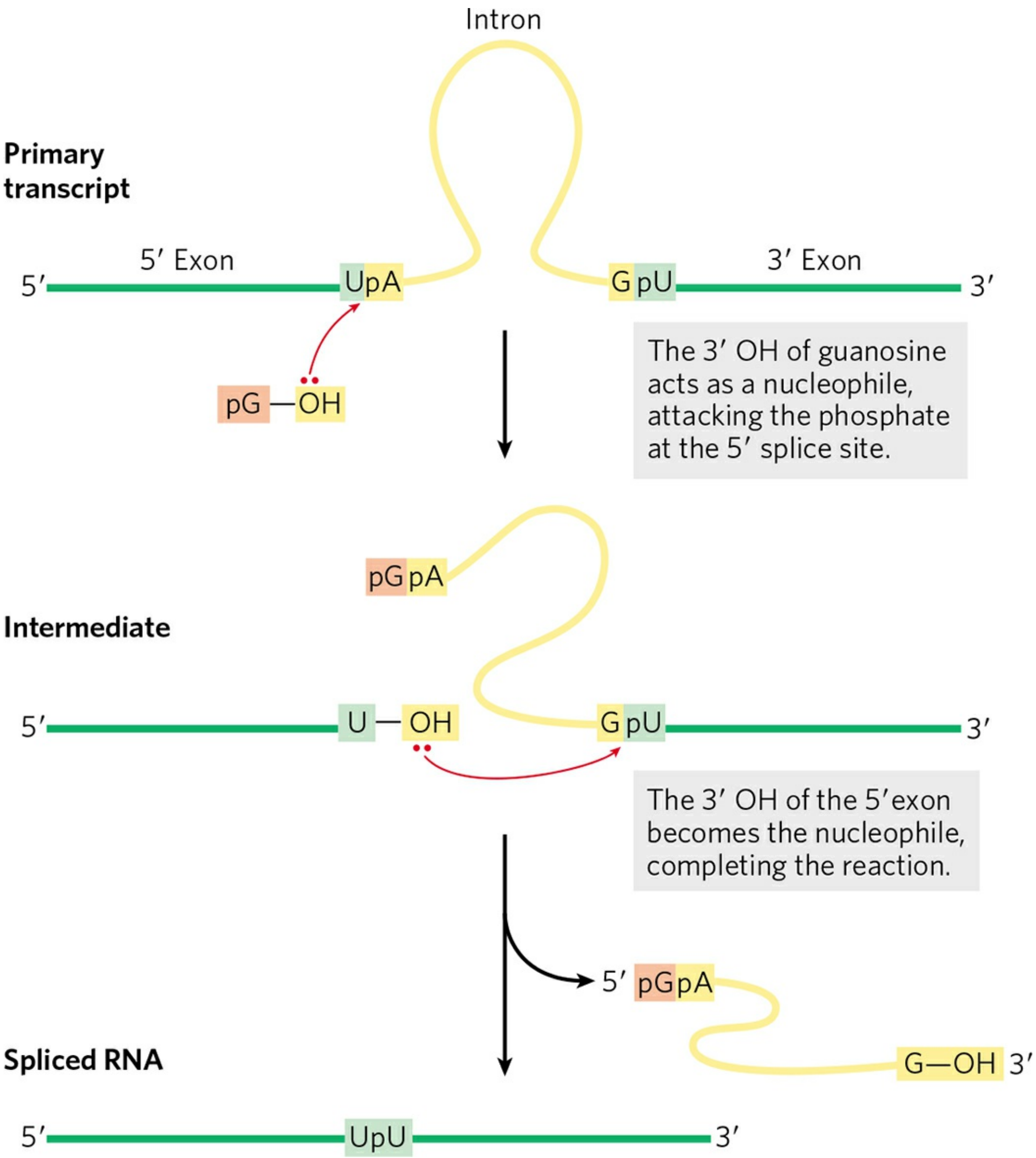

Intron
Primary transcript
5′ Exon
5′
UpA
G pU
3′ Exon
3′
pG—OH
The 3′ OH of guanosine acts as a nucleophile, attacking the phosphate at the 5′ splice site.
pG pA
Intermediate
5′
U—OH
G pU
3′
The 3′ OH of the 5′exon becomes the nucleophile, completing the reaction.
5′ pGpA
G—OH 3′
Spliced RNA
5′
UpU
3′

FIGURE 26-14 Splicing mechanism of group I introns. The nucleophile in the first step may be guanosine, GMP, GDP, or GTP. The spliced intron is eventually degraded.

In eukaryotes, most introns undergo splicing by the same lariat-forming mechanism as the group II introns. However, the intron splicing takes place within a large protein complex called a **spliceosome**, and these introns, the **spliceosomal introns**, are not assigned a group number. A spliceosome is made up of multiple specialized RNA-protein complexes called *s*mall *n*uclear *r*ibo*n*ucleo*p*roteins (snRNPs, often pronounced *snurps*). Each snRNP contains one of a class of eukaryotic RNAs, 100 to 200 nucleotides long, known as **small nuclear RNAs (snRNAs)**. Five snRNAs (U1, U2, U4, U5, U6) involved in splicing reactions are generally found in abundance in eukaryotic nuclei. In yeast, the various snRNPs include about 100 different proteins, most of which have close homologs in all other eukaryotes. In humans, these conserved protein components are augmented by more than 200 additional proteins. Spliceosomes are thus among the most complex macromolecular machines in any eukaryotic cell. The RNA components of a spliceosome are the catalysts of the various splicing steps. The overall complex can be considered a highly flexible nucleoprotein chaperone that can adapt to the great diversity in size and sequence of nuclear mRNAs.

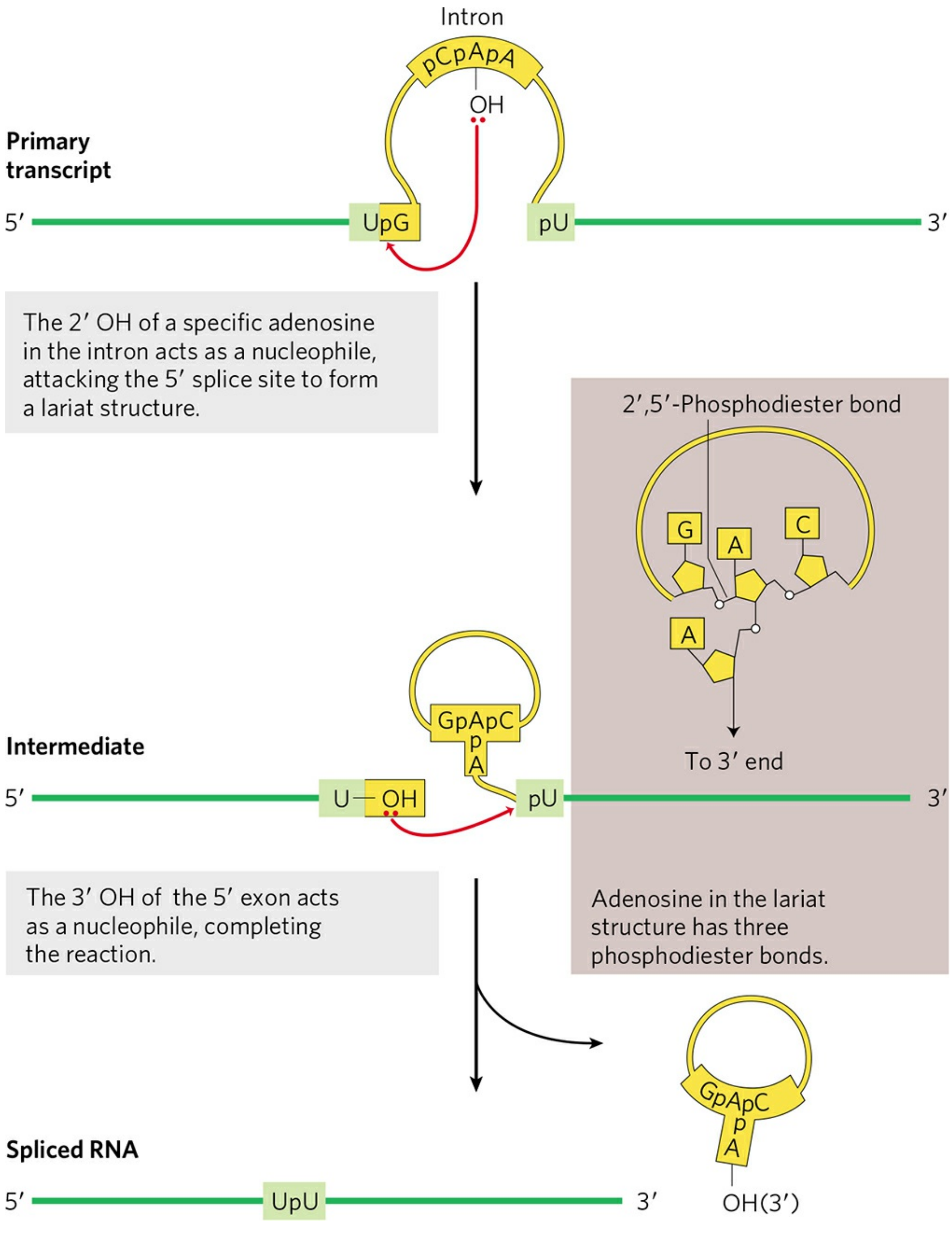

Intron
pCpApA
OH
Primary transcript
5′
UpG
pU
3′
The 2′ OH of a specific adenosine in the intron acts as a nucleophile, attacking the 5′ splice site to form a lariat structure.
2′,5′-Phosphodiester bond
G
A
C
A
To 3′ end
Intermediate
5′
U
OH
GpApC
p
A
pU
3′
The 3′ OH of the 5′ exon acts as a nucleophile, completing the reaction.
Adenosine in the lariat structure has three phosphodiester bonds.
GpApC
p
A
OH(3′)
Spliced RNA
5′
UpU
3′

FIGURE 26-15 Splicing mechanism of group II introns. The chemistry is similar to that of group I intron splicing, except for the identity of the nucleophile in the first step and the formation of a lariatlike intermediate, in which one branch is a 2′,5′-phosphodiester bond.

Spliceosomal introns generally have the dinucleotide sequence GU at the 5′ end and AG at the 3′ end, and these sequences mark the sites where splicing occurs. The U1 snRNA contains a sequence complementary to sequences near the 5′ splice site of nuclear mRNA introns **(Fig. 26-16a)**, and the U1 snRNP binds to this region in the primary transcript. A U2 snRNP binds to the 3′ end. Addition of the U4, U5, and U6 snRNPs leads to formation of the spliceosome (Fig. 26-16b). Key parts of the splicing active site found in U6 are initially sequestered by base pairing to parts of U4 to prevent aberrant cleavage of nontarget phosphodiester bonds. The U6 and U4 snRNAs must be unwound and separated to expose the active site needed for the first step in splicing. All steps in the process are reversible. Individual proteins associated with the various snRNPs sometimes have multiple functions: splicing, transport of the mRNA to the cytoplasm, translation, and eventual degradation of the mRNA. ATP is required for assembly of the spliceosome, but the RNA cleavage-ligation reactions do not require ATP. Some mRNA introns are spliced by a less common type of spliceosome, in which the U1 and U2 snRNPs are replaced by the U11 and U12 snRNPs. Whereas U1- and U2-containing spliceosomes remove introns with (5′)GU and AG(3′) terminal sequences, as shown in Figure 26-16, the U11- and U12-containing spliceosomes remove a rare class of introns that have (5′)AU and AC(3′) terminal sequences to mark the splice sites.

Some components of the splicing apparatus are tethered to the CTD of RNA polymerase II, indicating that splicing, like other RNA processing reactions, is tightly coordinated with transcription (Fig. 26-16c). As the first splice junction is synthesized, it is bound by a tethered spliceosome. The second splice junction is then captured by this complex as it passes, facilitating juxtaposition of the intron ends and the subsequent splicing process. After splicing, the intron remains in the nucleus and is eventually degraded.

The spliceosomes used in nuclear RNA splicing almost certainly evolved from more ancient group II introns, with the snRNPs contributing much greater levels of catalytic flexibility and regulation relative to their self-splicing ancestors.

A fourth and final class of introns, found in certain tRNAs, is distinguished from the group I and II introns in that the splicing reaction requires ATP and an endonuclease. The splicing endonuclease cleaves the phosphodiester bonds at both ends of the intron, and the two exons are joined by a mechanism similar to the DNA ligase reaction (see Fig. 25-16).

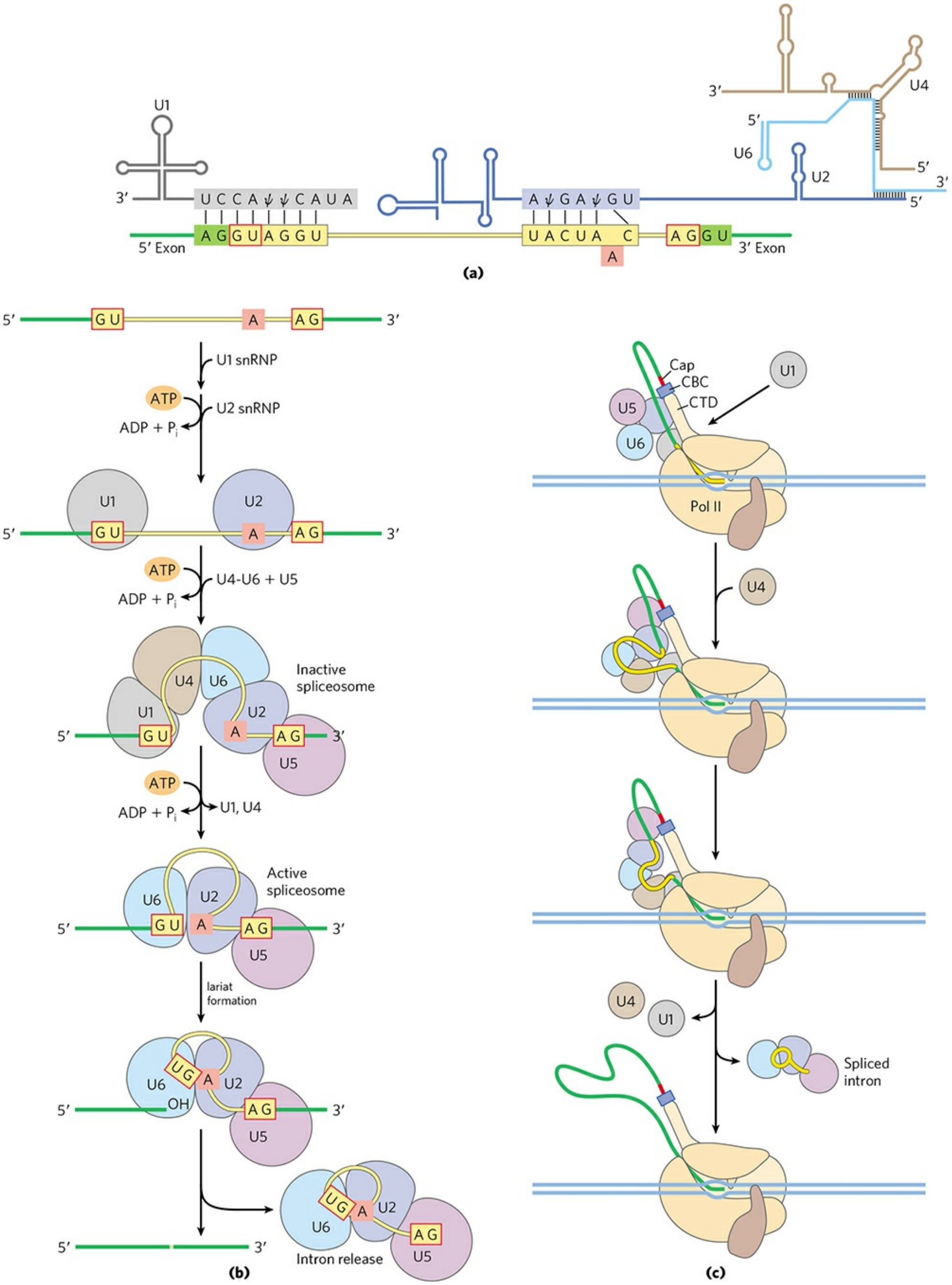


FIGURE 26-16 Splicing mechanism in mRNA primary transcripts. (a) RNA pairing interactions in the formation of spliceosome complexes. The U1 snRNA has a sequence near its 5′ end that is complementary to the splice site at the 5′ end of the intron. Base pairing of U1 to this region of the primary transcript helps define the 5′ splice site during spliceosome assembly (*ψ* is

pseudouridine; see Fig. 26-22). U2 is paired to the intron at a position encompassing the A residue (shaded light red) that becomes the nucleophile during the splicing reaction. Base pairing of U2 snRNA causes a bulge that displaces and helps to activate the adenylate, the 2′ OH of which will form the lariat structure through a 2′,5′-phosphodiester bond.

(b) Assembly of spliceosomes. All steps are reversible, but are shown proceeding in the forward direction for simplicity. The U1 and U2 snRNPs bind, then the remaining snRNPs (the U4-U6 complex and U5) bind to form an inactive spliceosome. Internal rearrangements convert this species to an active spliceosome in which U1 and U4 have been expelled and U6 is paired with both the 5′ splice site and U2. This is followed by the catalytic steps, which parallel those of the splicing of group II introns (see Fig. 26-15). The active spliceosome complex is illustrated on the cover of this book.

(c) Coordination of splicing and transcription brings the two splice sites together. See the text for details. All steps are reversible. The spliceosome is more than twice the size of RNA polymerase II.

Although spliceosomal introns seem to be limited to eukaryotes, the other three intron classes are not. Genes with group I and II introns have now been found in both bacteria and bacterial viruses. Bacteriophage T4, for example, has several protein-coding genes with group I introns. Introns may be more common in archaea than in bacteria.

Eukaryotic mRNAs Have a Distinctive 3′ End Structure

At their 3′ end, most eukaryotic mRNAs undergoing translation in the cell cytoplasm have a string of A residues, about 30 residues in yeast and 50 to 100 in animals, called the **poly(A) tail**. This tail serves as a binding site for one or more specific proteins. The poly(A) tail and its associated proteins have a variety of roles in coordinating transcription and translation, and may help protect mRNA from enzymatic destruction. Many bacterial mRNAs also acquire poly(A) tails, but these tails stimulate decay of mRNA rather than protecting it from degradation.

The poly(A) tail is added in a multistep process. The transcript is extended beyond the site where the poly(A) tail is to be added, then is cleaved at the poly(A) addition site by an endonuclease component of a large enzyme complex, again associated with the CTD of RNA polymerase II **(Fig. 26-17)**. The mRNA site where cleavage occurs is marked by two sequence elements: the highly conserved sequence (5′)AAUAAA(3′), 10 to 30 nucleotides on the 5′ side (upstream) of the cleavage site, and a less well-defined sequence rich in G and U residues, 20 to 40 nucleotides downstream of the cleavage site. Cleavage generates the free 3′-hydroxyl group that defines the end of the mRNA, to which A residues are immediately added by **polyadenylate polymerase**, which catalyzes the reaction

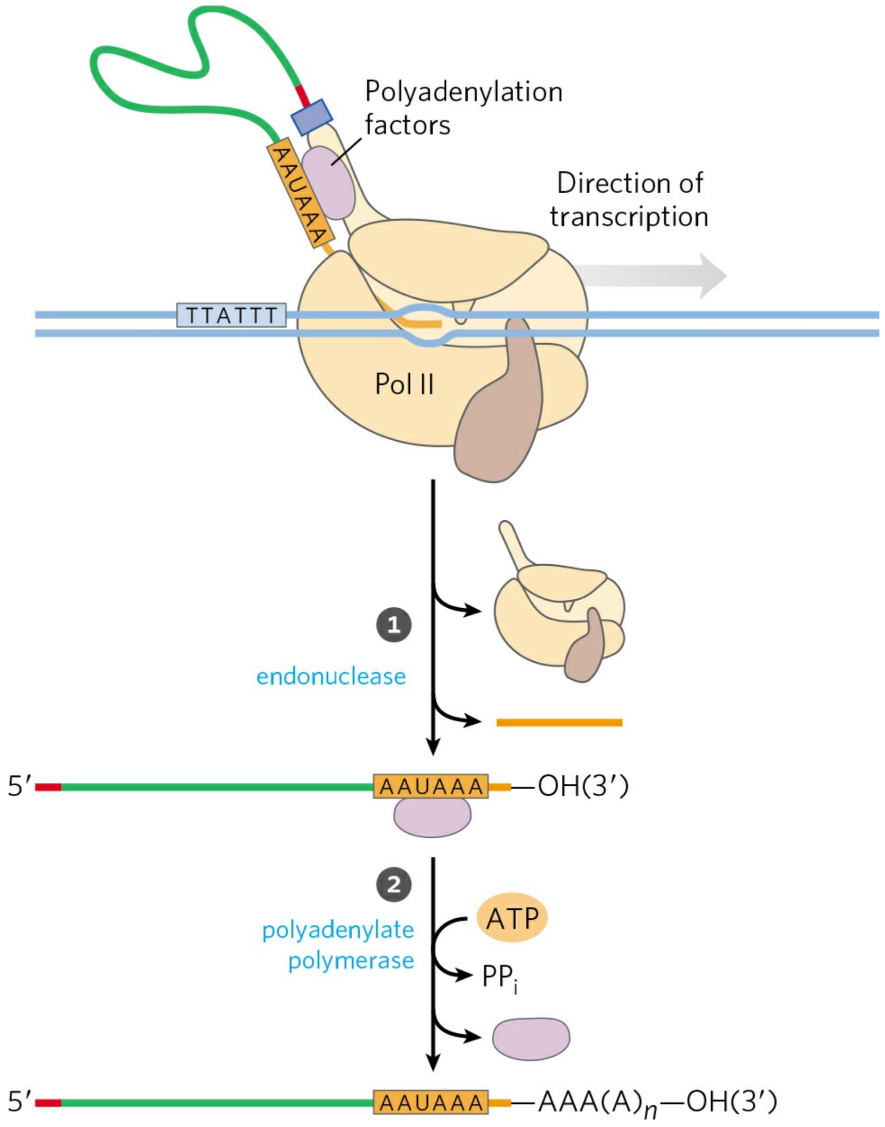


FIGURE 26-17 Addition of the poly(A) tail to the primary RNA transcript of eukaryotes. Pol II synthesizes RNA beyond the segment of the transcript containing the cleavage signal sequences, including the highly conserved upstream sequence (5′)AAUAAA. This cleavage signal sequence is bound by an enzyme complex that includes an endonuclease, a polyadenylate polymerase, and several other multisubunit proteins involved in sequence recognition, stimulation of cleavage, and regulation of the length of the poly(A) tail; all of these proteins are tethered to the CTD. 1 The RNA is cleaved by the endonuclease at a point 10 to 30 nucleotides 3′ to (downstream of) the

sequence AAUAAA. ❷ The polyadenylate polymerase synthesizes a poly(A) tail 80 to 250 nucleotides long, beginning at the cleavage site.

$$\mathrm{RNA} + \mathrm{nATP} \rightarrow \mathrm{RNA}\text{-}(\mathrm{AMP})_n + n\mathrm{PP_i}$$

where n = 80 to 250. This enzyme does not require a template but does require the cleaved mRNA as a primer. These longer poly(A) tails are added in the nucleus, and then shortened significantly after the mRNA is transported to the cytoplasm.

The overall processing of a typical eukaryotic mRNA is summarized in **Figure 26-18**. In some cases the polypeptide-coding region of the mRNA is also modified by RNA "editing" (see Section 27.1 for details). This editing includes processes that add or delete bases in the coding regions of primary transcripts or that change the sequence (such as by enzymatic deamination of a C residue to create a U residue). A particularly dramatic example occurs in trypanosomes, which are parasitic protists: large regions of an mRNA are synthesized without any uridylate, and the U residues are inserted later by RNA editing.

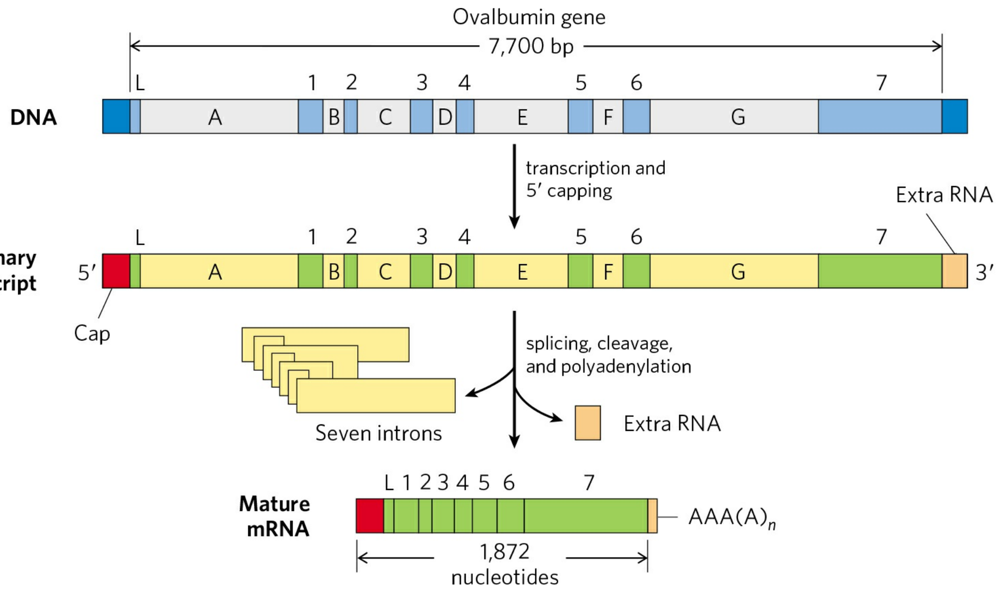
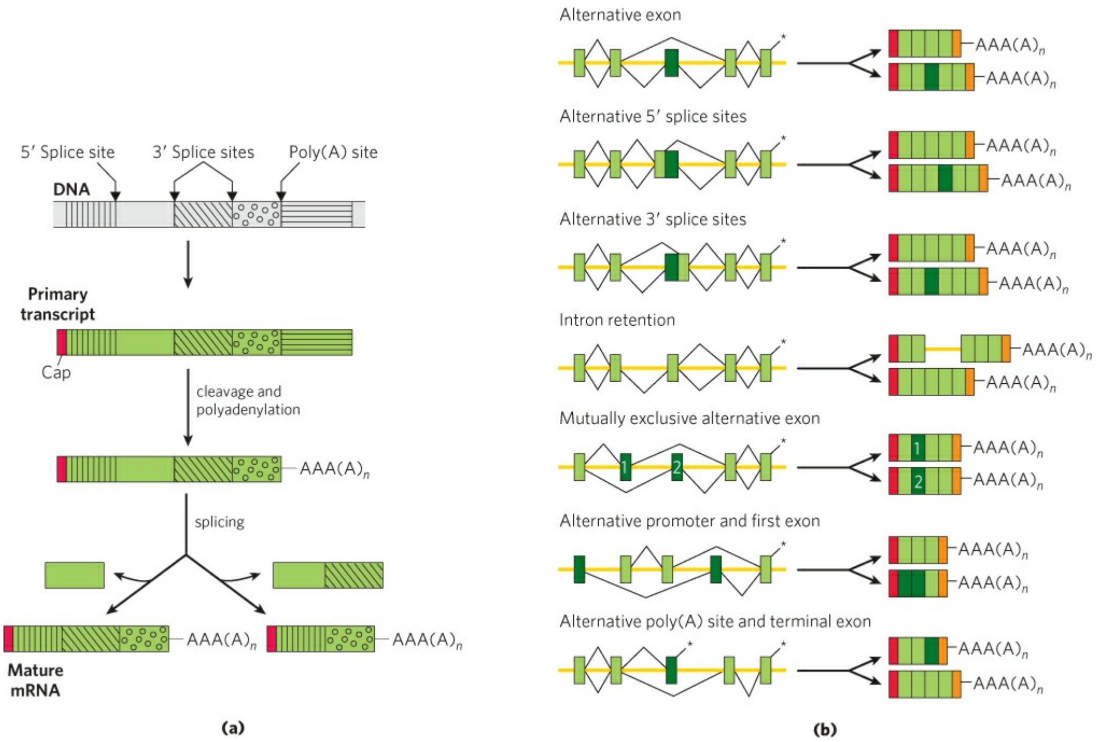


FIGURE 26-18 Overview of the processing of a eukaryotic mRNA. The ovalbumin gene, shown here, has introns A to G and exons 1 to 7 and L (L encodes a signal peptide sequence that targets the protein for export from the cell; see Fig. 27-40). About three-quarters of the RNA is removed during processing. Pol II extends the primary transcript well beyond the cleavage and polyadenylation site ("extra RNA") before terminating transcription. Termination signals for Pol II have not yet been defined.

A Gene Can Give Rise to Multiple Products by Differential RNA Processing

One of the paradoxes of modern genomics is that the apparent complexity of organisms does not correlate with the number of protein-coding genes, or even the amount of genomic DNA. Some eukaryotic mRNA transcripts can be processed in more than one way to produce *different* mRNAs and thus different polypeptides. Much of the variability in processing is the result of **alternative splicing**, in which a particular exon may or may not be incorporated into the mature mRNA transcript. Alternative splicing occurs in a relatively small number of genes in yeast, but in more than 95% of human genes.


5′ Splice site
3′ Splice sites
Poly(A) site
DNA
Primary transcript
Cap
cleavage and polyadenylation
AAA(A)$_n$
splicing
Mature mRNA
(a)
Alternative exon
Alternative 5′ splice sites
Alternative 3′ splice sites
Intron retention
Mutually exclusive alternative exon
1
2
Alternative promoter and first exon
Alternative poly(A) site and terminal exon
AAA(A)$_n$
(b)

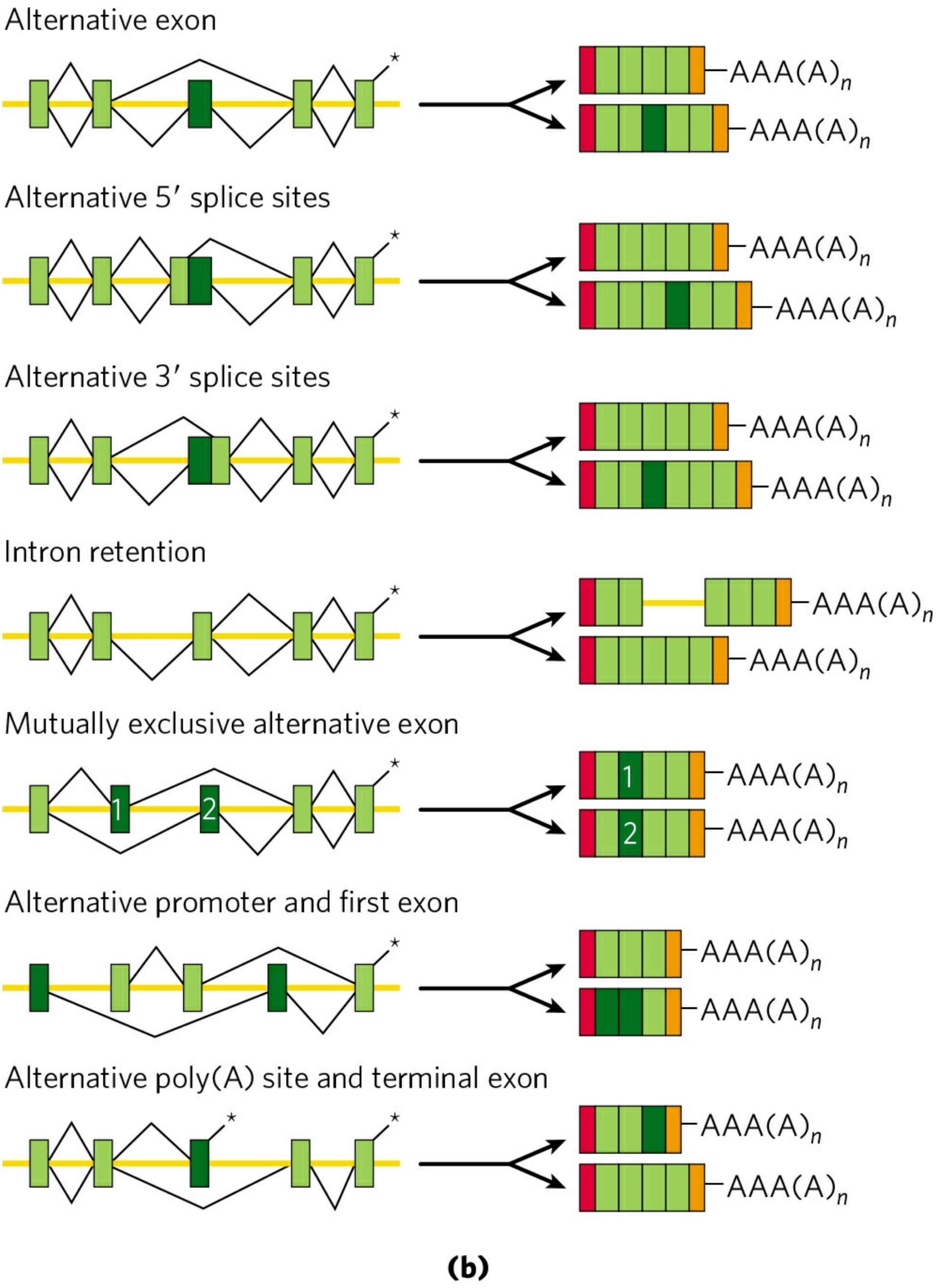


FIGURE 26-19 Alternative splicing in eukaryotes. (a) Alternative splicing patterns. Two different 3′ splice sites are shown. Different mature mRNAs are produced from the same primary transcript. **(b)** Summary of splicing patterns. Exons are shown in shades of green, and introns/untranslated regions as yellow lines. Positions where polyadenosine is to be added are marked with asterisks.

Exons joined in a particular splicing scheme are linked with black lines. For each transcript, the alternative linkage patterns shown above and below the transcript produce the top and bottom spliced mRNA, respectively. In the products, 5′ caps are represented by red boxes, 3′ untranslated regions by orange boxes.

[Source: (b) Information from B. J. Blencowe, *Cell* 126:37, 2006, Fig. 2.]

Figure 26-19a illustrates how alternative splicing patterns can produce more than one protein from a common primary transcript. The primary transcript contains molecular signals for all the alternative processing pathways, and the pathway favored in a given cell or metabolic situation is determined by processing factors, RNA-binding proteins that promote one particular path. For example, the protein composition of the snRNPs involved in splicing may vary somewhat in the spliceosomes that participate in the processing of different genes, and may change further so that the processing of a particular gene is altered at different stages of animal development. As one example, such alternative processing produces three different forms of the myosin heavy chain at different stages of fruit fly development. There are many additional patterns of alternative splicing (Fig. 26-19b).

Complex transcripts can also have more than one site where poly(A) tails can form. If there are two or more sites for cleavage and polyadenylation, use of the one closest to the 5′ end will remove more of the primary transcript sequence **(Fig. 26-20)**. This mechanism, called **poly(A) site choice**, generates diversity in the variable domains of immunoglobulin heavy chains (see Fig. 25-43).

Both alternative splicing and poly(A) site choice come into play in the processing of many genes. For example, a single RNA transcript is processed using both mechanisms to produce two different hormones: the calcium-regulating hormone calcitonin in rat thyroid and calcitonin-gene-related peptide (CGRP) in rat brain **(Fig. 26-21)**. Together, alternative splicing and poly(A) site choice greatly increase the number of different proteins generated from the genomes of higher eukaryotes.

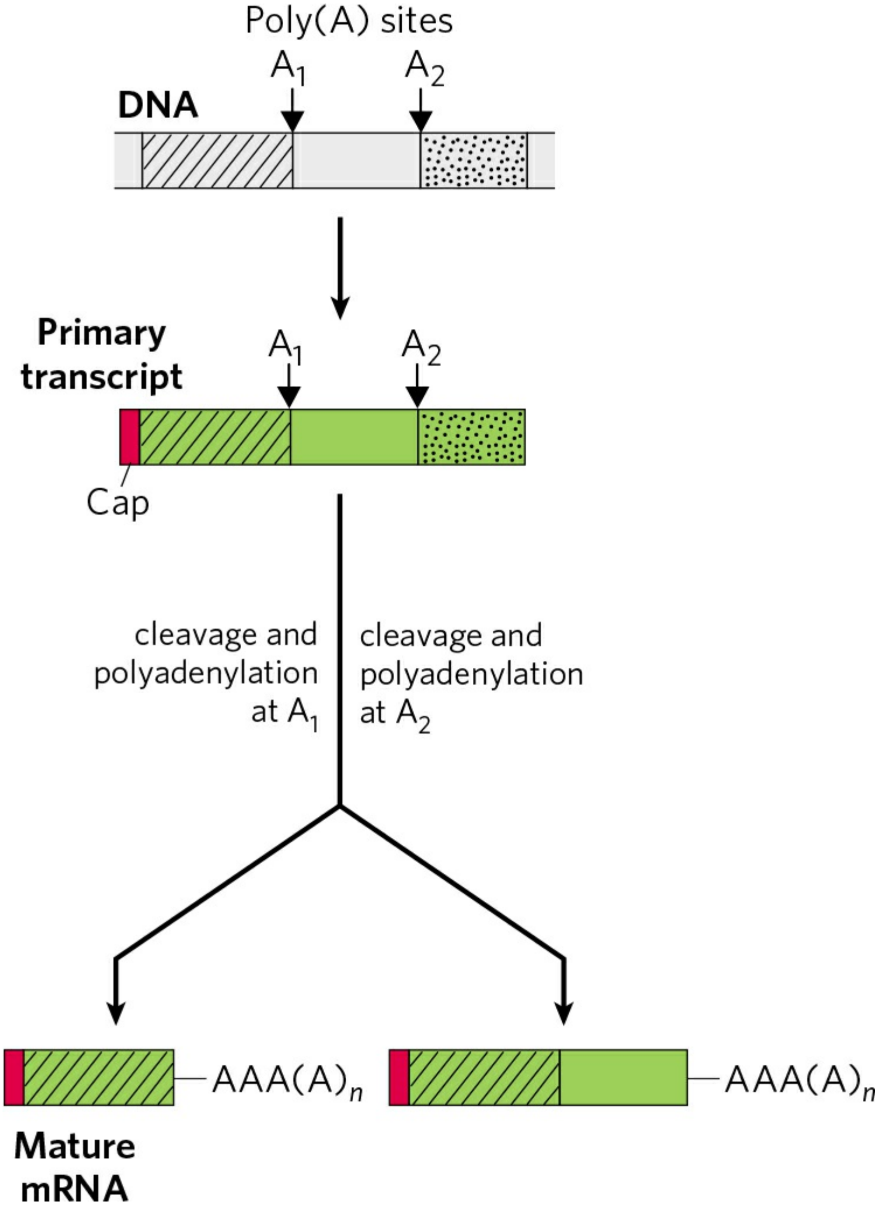


FIGURE 26-20 Poly(A) site choice. Two alternative cleavage and polyadenylation sites, A_1 and A_2, are shown.

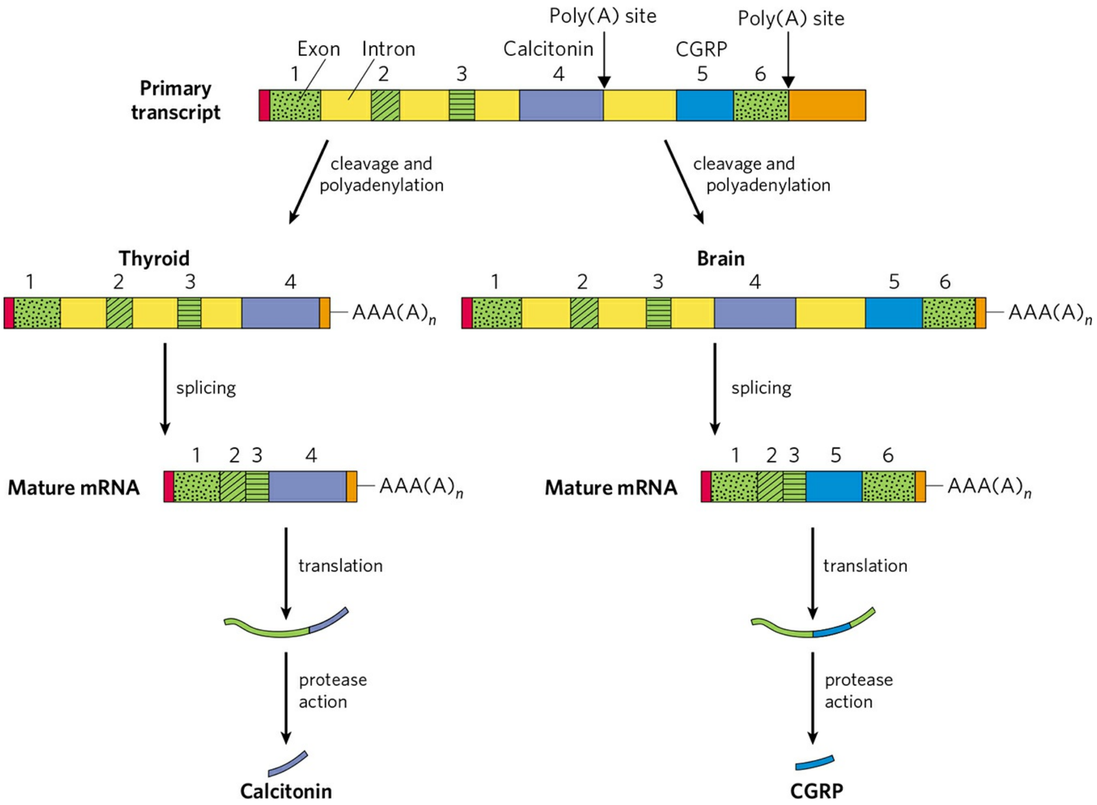


FIGURE 26-21 Alternative processing of the calcitonin gene transcript in rats. The primary transcript has two poly(A) sites; one predominates in the brain, the other in the thyroid. In the brain, splicing eliminates the calcitonin exon (exon 4); in the thyroid, this exon is retained. The resulting peptides are processed further to yield the final hormone products: calcitonin-gene-related peptide (CGRP) in the brain and calcitonin in the thyroid.

26.3 RNA-Dependent Synthesis of RNA and DNA

In our discussion of DNA and RNA synthesis up to this point, the role of the template strand has been reserved for DNA. However, some enzymes use an RNA template for nucleic acid synthesis. With the important exception of viruses with an RNA genome, these enzymes play only a modest role in information pathways. RNA viruses are the source of most RNA-dependent polymerases characterized so far.

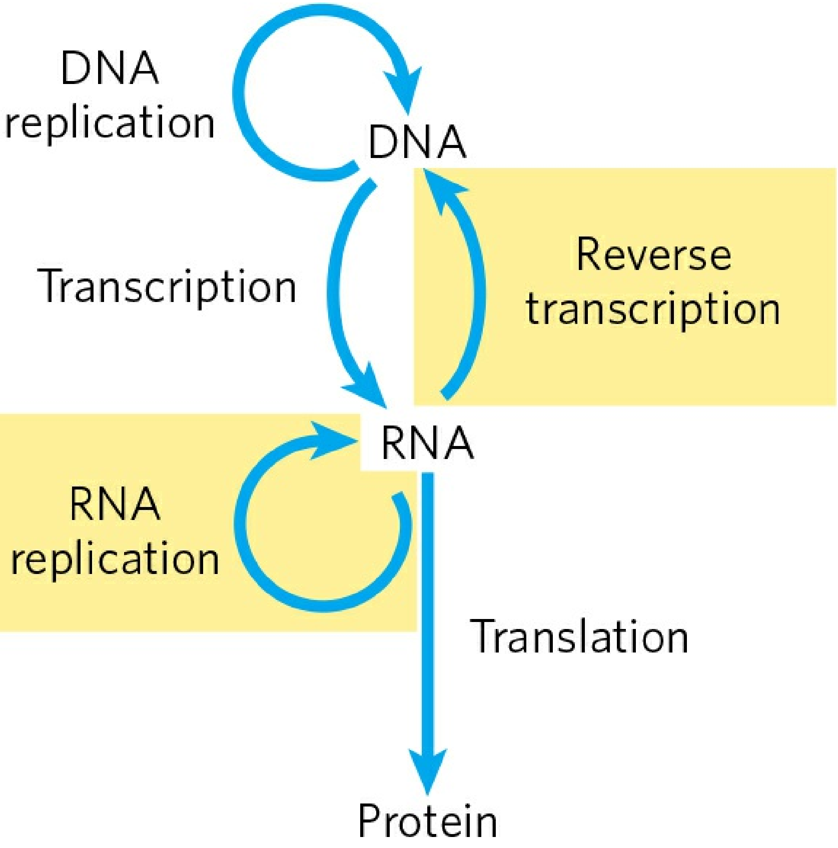


FIGURE 26-31 Extension of the central dogma to include RNA- dependent synthesis of RNA and DNA.

The existence of RNA replication requires an elaboration of the central dogma (**Fig. 26-31**). The enzymes ofthe RNA replication process have profound implications for investigationsinto the nature of self-replicating molecules that may have existed inprebiotic times.

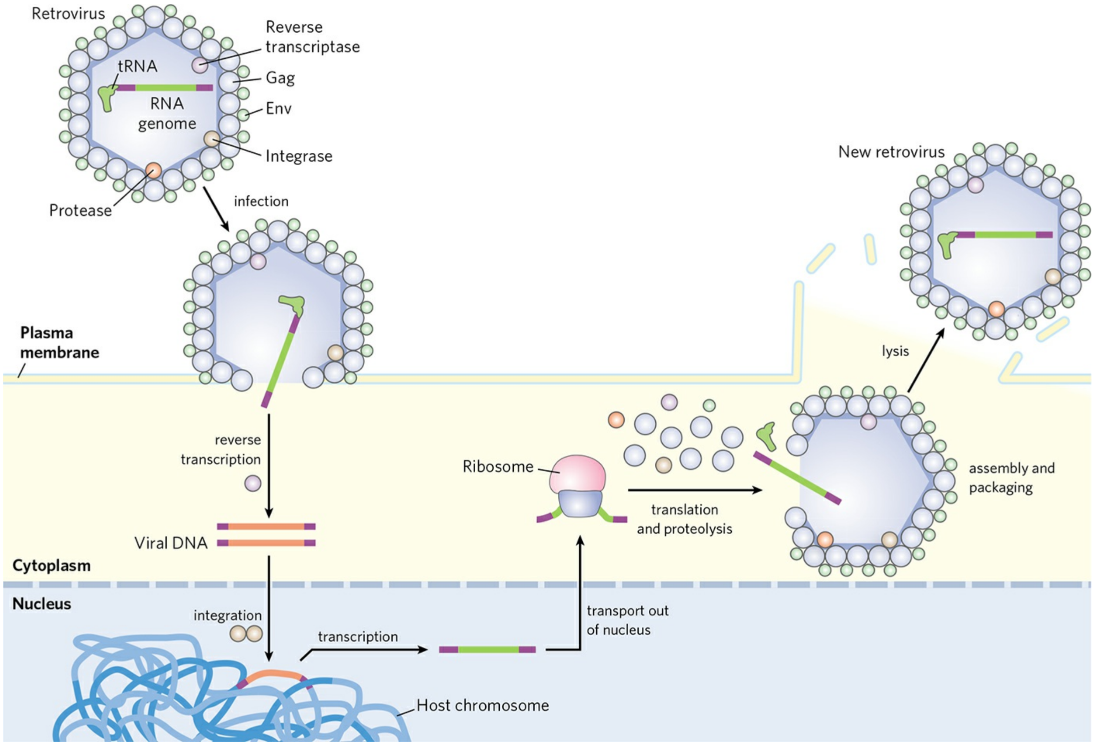


FIGURE 26-32 Retroviral infection of a mammalian cell and integration of the retrovirus into the host chromosome. Viral particles entering the host cell carry viral reverse transcriptase and a cellular tRNA (picked up from a former host cell) already base-paired to the viral RNA. The purple segments represent the long terminal repeats on the viral RNA. The tRNA facilitates immediate conversion of viral RNA to double-stranded DNA by the action of reverse transcriptase, as described in the text. The double-stranded DNA enters the nucleus and is integrated into the host genome. The integration is catalyzed by a virally encoded integrase. Integration of viral DNA into host DNA is mechanistically similar to the insertion of transposons in bacterial chromosomes. For example, a few base pairs of host DNA become duplicated at the site of integration, forming short repeats of 4 to 6 bpat each end of the inserted retroviral DNA (not shown). On transcription and translation of the integrated viral DNA, new viruses are formed and released by cell lysis (right). In the viruses, the viral RNA is enclosed by capsid proteins called Gag and outer envelope proteins called Env. Additional viral proteins (reverse transcriptase, integrase, and a viral protease needed for posttranslational processing of viral proteins) are packaged within the virus particle with the RNA.

Reverse Transcriptase Produces DNA from Viral RNA

Certain RNA viruses that infect animal cells carry within the viral particle an RNA-dependent DNA polymerase called **reverse transcriptase**. On infection, the single-stranded RNA viral genome (~10,000 nucleotides) and the enzyme enter the host cell. The reverse transcriptase first catalyzes the synthesis of a DNA strand complementary to the viral RNA **(Fig. 26-32)**, then degrades the RNA strand of the viral RNA-DNA hybrid and replaces it with DNA. The resulting duplex DNA often becomes incorporated into the genome of the eukaryotic host cell. These integrated (and dormant) viral genes can be activated and transcribed, and the gene products—viral proteins and the viral RNA genome itself—are packaged as new viruses. The RNA viruses that contain reverse transcriptases are known as **retroviruses** (*retro* is the Latin prefix for "backward").

The existence of reverse transcriptases in RNA viruses was predicted by Howard Temin in 1962, and the enzymes were ultimately detected by Temin and, independently, by David Baltimore in 1970. Their discovery aroused much attention as dogma-shaking proof that genetic information can flow "backward" from RNA to DNA.

Retroviruses typically have three genes: *gag* (a name derived from the historical designation group associated antigen), *pol*, and *env* **(Fig. 26-33)**. The transcript that contains *gag* and *pol* is translated into a long "polyprotein," a single large polypeptide that is cleaved into six proteins with distinct functions. The proteins derived from the *gag* gene make up the interior core of the viral particle. The *pol* gene encodes the protease that cleaves the long polypeptide, an integrase that inserts the viral DNA into the host chromosomes, and reverse transcriptase. Many reverse transcriptases have two subunits, α and β. The *pol* gene specifies the β subunit (M_r 90,000), and the α subunit (M_r 65,000) is simply a proteolytic fragment of the β subunit. The *env* gene encodes the proteins of the viral envelope. At each end of the linear RNA genome are long terminal repeat (LTR) sequences of a few hundred nucleotides. Transcribed into the duplex DNA, these sequences facilitate integration of the viral chromosome into the host DNA and contain promoters for viral gene expression.

Reverse transcriptases catalyze three different reactions: (1) RNA-dependent DNA synthesis, (2) RNA degradation, and (3) DNA-dependent DNA synthesis. Like many DNA and RNA polymerases, reverse transcriptases contain Zn^{2+}. Each transcriptase is most active with the RNA of its own virus, but each can be used experimentally to make DNA complementary to a variety of RNAs. The DNA and RNA synthesis and RNA degradation activities use separate active sites on the protein. For DNA synthesis to begin, the reverse transcriptase requires a primer, a cellular tRNA obtained during an earlier infection and carried in the viral particle. This tRNA is base-paired at its 3′ end with a complementary sequence in the viral RNA. The new DNA strand is synthesized in the 5′ → 3′ direction, as in all RNA and DNA polymerase reactions. Reverse transcriptases, like RNA polymerases, do not have 3′ → 5′ proofreading exonucleases. They generally have error rates of about 1 per 20,000 nucleotides added. An error rate this high is extremely unusual in DNA replication and seems to be a characteristic of most enzymes that replicate the genomes of RNA viruses. A consequence is a higher mutation rate and faster rate of viral evolution, which is a factor in the frequent appearance of new strains of disease-causing retroviruses.

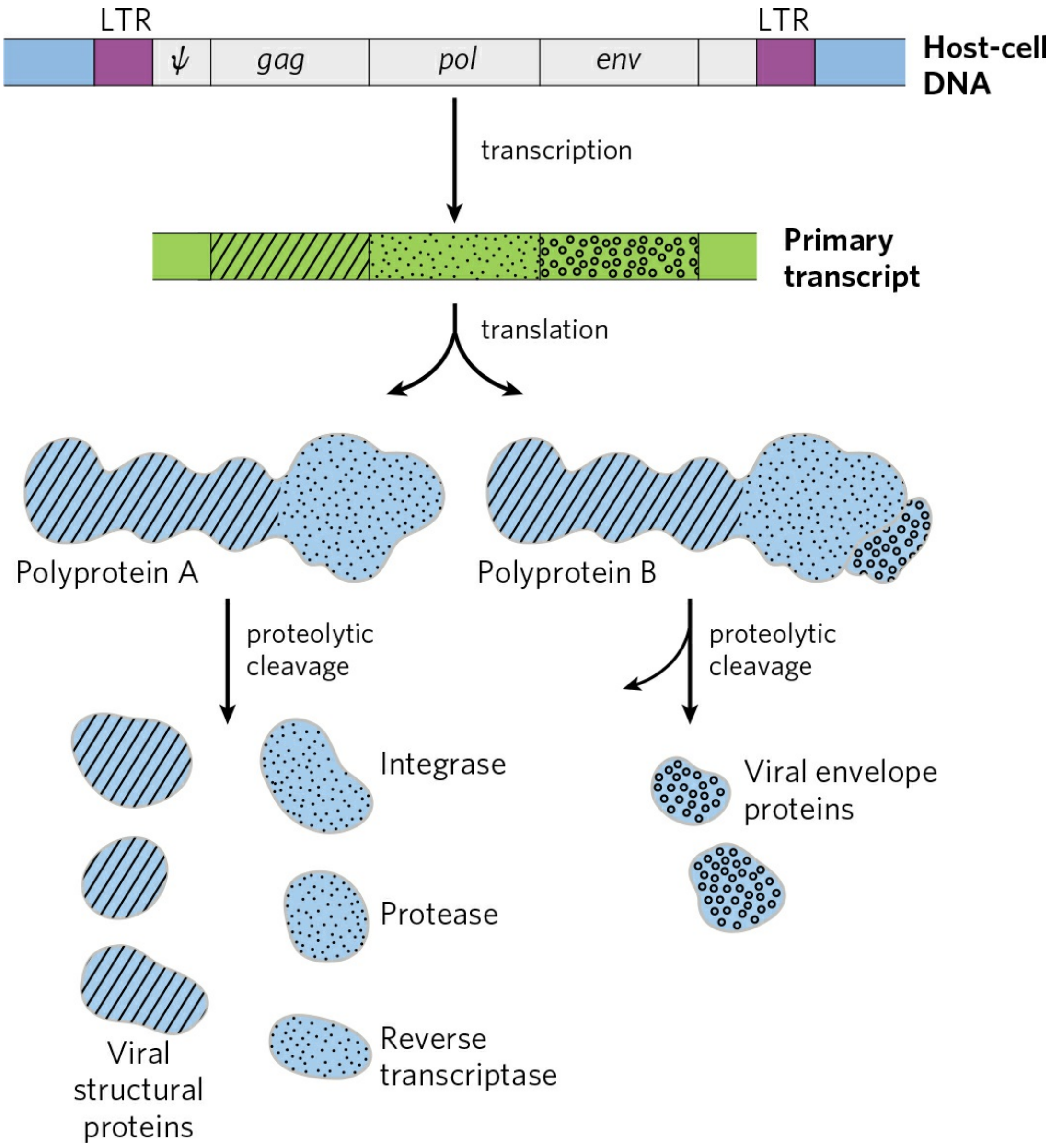


FIGURE 26-33 Structure and gene products of an integrated retroviral genome. The long terminal repeats (LTRs) have sequences needed for the regulation and initiation of transcription. The sequence denoted *ψ* is required for packaging of retroviral RNAs into mature viral particles. Transcription of the retroviral DNA produces a primary transcript encompassing the *gag, pol*, and *env* genes. Translation produces a polyprotein, a single long polypeptide derived from the *gag* and *pol* genes, which is cleaved into six distinct proteins. Splicing of the primary transcript yields an mRNA derived largely from the *env* gene, which is also translated into a polyprotein, then cleaved to generate viral envelope proteins.

Reverse transcriptases have become important reagents in the study of DNA-RNA relationships and in DNA cloning techniques. They make possible the synthesis of DNA complementary to an mRNA template, and synthetic DNA prepared in this manner, called **complementary DNA (cDNA)**, can be used to clone cellular genes .

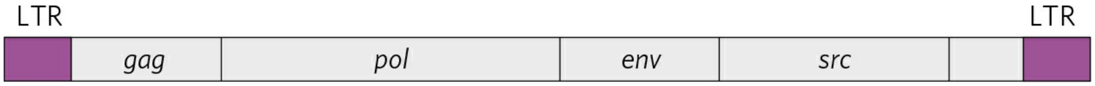


FIGURE 26-34 **Rous sarcoma virus genome.** The *src* gene encodes a tyrosine kinase, one of a class of enzymes that function in systems affecting cell division, cell-cell interactions, and intercellular communication .The same gene is found in chicken DNA (the usual host for this virus) and inthe genomes of many other eukaryotes, including humans. When associated with the Rous sarcoma virus, this oncogene is often expressed at abnormally high levels, contributing to unregulated cell division and cancer.

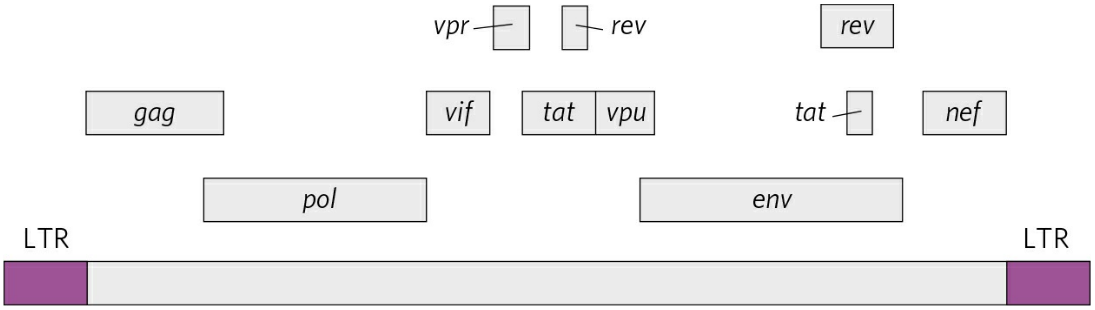


FIGURE 26-35 **The genome of HIV, the virus that causes AIDS.** In addition to the typical retroviral genes, HIV contains several small genes with a variety of functions (not identified here and not all known). Some of these genes overlap. Alternative splicing mechanisms produce many different proteins from this small (9.7 × 10^3 nucleotides) genome.

Some Retroviruses Cause Cancer and AIDS

Retroviruses have featured prominently in recent advances in the molecular understanding of cancer. Most retroviruses do not kill their host cells but remain integrated in the cellular DNA, replicating when the cell divides. Some retroviruses, classified as RNA tumor viruses, contain an oncogene that can cause the cell to grow abnormally. The first retrovirus of this type to be studied was the Rous sarcoma virus (also called avian sarcoma virus; Fig. 26-34), named for F. Peyton Rous, who studied chicken tumors now known to be caused by this virus. Since the initial discovery of oncogenes by Harold Varmus and Michael Bishop, many dozens of such genes have been found in retroviruses.

The human immunodeficiency virus (HIV), which causes acquired immune deficiency syndrome (AIDS), is a retrovirus. Identified in 1983, HIV has an RNA genome with standard retroviral genes along with several other unusual genes **(Fig. 26-35)**. Unlike many other retroviruses, HIV kills many of the cells it infects (principally T lymphocytes) rather than causing tumor formation. This gradually leads to suppression of the immune system in the host organism. The reverse transcriptase of HIV is even more error-prone than other known reverse transcriptases—10 times more so—resulting in high mutation rates in this virus. One or more errors are generally made every time the viral genome is replicated, so any two viral RNA molecules are likely to differ.

Many modern vaccines for viral infections consist of one or more coat proteins of the virus. . These proteins are not infectious on their own but stimulate the immune system torecognize and resist subsequent viral invasions. Because of thehigh error rate of the HIV reverse transcriptase, the *env* gene in this virus
(along with the rest of the genome) undergoes very rapid mutation, complicating the development of an effective vaccine. However, repeated cycles of cell invasion and replication are needed to propagate an HIV infection, so inhibition of viral enzymes offers the most effective therapy currently available. The HIV protease is targeted by a class of drugs called protease inhibitors . Reverse transcriptase is the target of someadditional drugs widely used to treat HIV-infected individuals (Box 26-2). ■

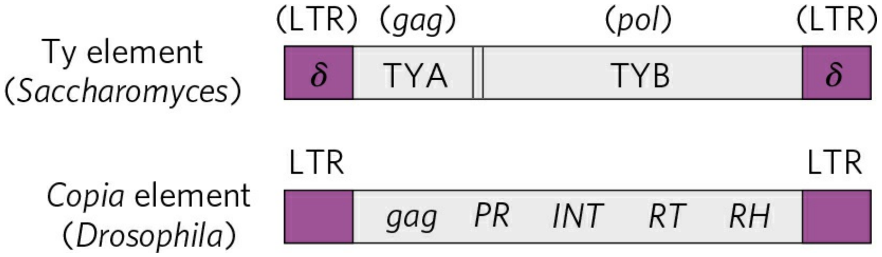


FIGURE 26-36 Eukaryotic transposons. The Ty element of the yeast *Saccharomyces* and the *copia* element of the fruit fly *Drosophila* are examples of eukaryotic retrotransposons, which often have a structure similar to retroviruses but lack the env gene. The δ sequences of the Ty element are functionally equivalent to retroviral LTRs. In the *copia* element, *INT* and *RT*

are homologous to the integrase and reverse transcriptase segments, respectively, of the *pol* gene.

Many Transposons, Retroviruses, and Introns May Have a Common Evolutionary Origin

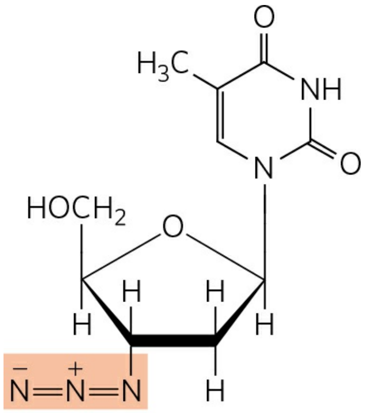
Some well-characterized eukaryotic DNA transposons from sources as diverse as yeast and fruit flies have a structure very similar to that of retroviruses; these are sometimes called retrotransposons **(Fig. 26-36)**. Retrotransposons encode an enzyme homologous to the retroviral reverse transcriptase, and their coding regions are flanked by LTR sequences. They transpose from one position to another in the cellular genome by means of an RNA intermediate, using reverse transcriptase to make a DNA copy of the RNA, followed by integration of the DNA at a new site. Most transposons in eukaryotes use this mechanism for transposition, distinguishing them from bacterial transposons, which move as DNA directly from one chromosomal location to another .

BOX 26-2 MEDICINE Fighting AIDS with Inhibitors of HIV Reverse Transcriptase

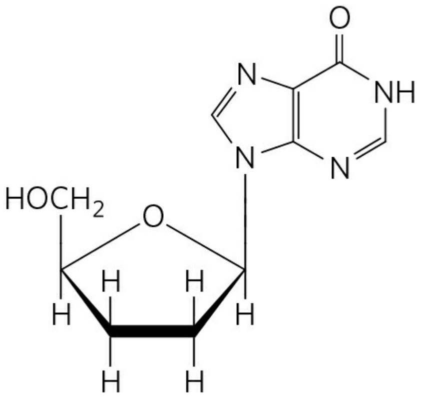
Research into the chemistry of template-dependent nucleic acid biosynthesis, combined with modern techniques of molecular biology, has elucidated the life cycle and structure of the human immunodeficiency virus, the retrovirus that causes AIDS. A few years after the isolation of HIV, this research resulted in the development of drugs capable of prolonging the lives of people infected by HIV.

The first drug to be approved for clinical use was AZT, a structural analog of deoxythymidine. AZT was first synthesized in 1964 by Jerome P. Horwitz. It failed as an anticancer drug (the purpose for which it was made), but in 1985 it was found to be a useful treatment for AIDS. AZT is taken up by T lymphocytes, immune system cells that are particularly vulnerable to HIV infection, and converted to AZT triphosphate. (AZT triphosphate taken directly would be ineffective because it cannot cross the plasma membrane.) HIV’s reverse transcriptase has a higher affinity for AZT triphosphate than for dTTP, and binding of AZT triphosphate to this enzyme competitively inhibits dTTP binding. When AZT is added to the 3′

end of the growing DNA strand, lack of a 3′ hydroxyl means that the DNA strand is terminated prematurely and viral DNA synthesis grinds to a halt.



3′-Azido-2′,3′-dideoxy-thymidine (AZT)



2′,3′-Dideoxyinosine (DDI)

AZT triphosphate is not as toxic to the T lymphocytes themselves because *cellular* DNA polymerases have a lower affinity for this compound than for dTTP. At concentrations of 1 to 5 μ_{m}, AZT affects HIV reverse transcription but not most cellular DNA replication. Unfortunately, AZT seems to be toxic to the bone marrow cells that are the progenitors of erythrocytes, and many individuals taking AZT develop anemia. AZT can increase the survival time of people with advanced AIDS by about a year, and it delays the onset of AIDS in those who are still in the early stages of HIV infection. Some other AIDS drugs, such as dideoxyinosine (DDI), have a similar mechanism of action. Newer drugs target and inactivate the HIV protease. Because of the high error rate of HIV reverse transcriptase and the resulting rapid evolution of HIV, the most effective treatments of HIV infection use a combination of drugs directed at both the protease and the reverse transcriptase.

Retrotransposons lack an *env* gene and so cannot form viral particles. They can be thought of as defective viruses, trapped in cells. Comparisons between retroviruses and eukaryotic transposons suggest that reverse transcriptase is an ancient enzyme that predates the evolution of multicellular organisms.

Many group I and group II introns are also mobile genetic elements. In addition to their self-splicing activities, they encode DNA endonucleases that promote their movement. During genetic exchanges between cells of the same species, or when DNA is introduced into a cell by parasites or by other means, these endonucleases promote insertion of the intron into an identical site in another DNA copy of a homologous gene that does not contain the intron, in a process termed **homing** **(Fig. 26-37)**. Whereas group I intron homing is DNA-based, group II intron homing occurs through an RNA intermediate. The endonucleases of the group II introns have associated reverse transcriptase activity. The proteins can form complexes with the intron RNAs themselves, after the introns are spliced from the primary transcripts. Because the homing process involves insertion of the RNA intron into DNA and reverse transcription of the intron, the movement of these introns has been called retrohoming. Over time, every copy of a particular gene in a population may acquire the intron. Much more rarely, the intron may insert itself into a new location in an unrelated gene. If this event does not kill the host cell, it can lead to the evolution and distribution of an intron in a new location. The structures and mechanisms used by mobile introns support the idea that at least some introns originated as molecular parasites whose evolutionary past can be traced to retroviruses and transposons.

THE PROCESS OF REVERSE TRANSCRIPTION

When a mature HIV-1 virion infects a susceptible target cell, interactions of the envelope glycoprotein with the coreceptors on the surface of the cell brings about a fusion of the membranes of the host cell and the virion (Wilen et al. 2011). This fusion introduces the contents of the virion into the cytoplasm of the cell, setting the stage for reverse transcription. There are complexities to the early events that accompany reverse transcription in an infected cell, not all of which are well understood, which will be considered later in this article. We will begin by discussing the mechanics of the conversion of the single-stranded RNA genome found in the virion into the linear double-stranded DNA that is the substrate for the integration process. The synthesis of this linear DNA is a reasonably well-understood process; additional details and references can be found in the books *Retroviruses* (Telesnitsky and Goff 1997) and *Reverse Transcriptase* (Skalka and Goff 1993). In orthoretroviruses, including HIV-1, reverse transcription takes place in newly infected cells. There is some debate in the literature about whether reverse transcription is initiated in producer cells. Primer tagging experiments suggest that most HIV-1 virions initiate reverse transcription in newly infected cells (Whitcomb et al. 1990); however, there are claims that a small number of nucleotides may be incorporated before the virions initiate infection of target cells (Lori et al. 1992; Trono 1992; Zhu and Cunningham 1993; Huang et al. 1997). Either way, the vast majority of the viral DNA is synthesized in newly infected cells. This is a lifestyle choice; spumaretroviruses and the more distantly related hepadna viruses carry out extensive reverse transcription in producer cells (Summers and Mason 1982; Yu et al. 1996, 1999). Although there are viral and cellular factors that assist in the process of reverse transcription (these will be discussed later) the two enzymatic activities that are necessary and sufficient to carry out reverse transcription are present in RT. These are a DNA polymerase that can copy either a RNA or a DNA template, and an RNase H that degrades RNA if, and only if, it is part of an RNA–DNA duplex.

Like many other DNA polymerases, RT needs both a primer and a template. Genomic RNA is plus-stranded (the genome and the messages are copied from the same DNA strand), and the primer for the synthesis of the first DNA strand (the minus strand) is a host tRNA whose 3′ end is base paired to a complementary sequence near the 5′ end of the viral RNA called the primer binding site (pbs). Different retroviruses use different host tRNAs as primers. HIV-1 uses Lys3. It would appear, based on in vitro experiments, that the addition of the first few nucleotides is slow and difficult. DNA synthesis speeds up considerably once the first five to six deoxyribonucleotides have been added to the 3′ end of the tRNA primer (Isel et al. 1996; Lanchy et al. 1998). In HIV-1, the pbs is approximately 180 nucleotides from the 5′ end of genomic RNA. DNA synthesis creates an RNA–DNA duplex, which is a substrate for RNase H. There are perhaps 50 RTs in an HIV-1 virion; it is unclear whether the same RT that synthesizes the DNA plays a significant role in degrading the RNA. This is not a requirement—retroviruses can replicate (at a considerably reduced efficiency) with a mixture of RTs, some of which have only polymerase activity and some that have only RNase H activity (Telesnitsky and Goff 1993; Julias et al. 2001). Moreover, in in vitro assays, little or no RNase H cleavage is detected while RT is actively synthesizing DNA; instead, cleavages occur at sites where DNA synthesis pauses (Driscoll et al. 2001; Purohit et al. 2007). Whatever the exact mechanism, RNase H degradation removes the 5′ end of the viral RNA, exposing the newly synthesized minus-strand DNA (see Fig. 1).

The ends of the viral RNA are direct repeats, called R. These repeats act as a bridge that allows the newly synthesized minus-strand DNA to be transferred to the 3′ end of the viral RNA. Retroviruses package two copies of the viral RNA genome; the first (or minus-strand) transfer can involve the R sequence at the 3′ ends of either of the two RNAs (Panganiban and Fiore 1988; Hu and Temin 1990b; van Wamel and Berkhout 1998; Yu et al. 1998). After this transfer, minus-strand synthesis can continue along the length of the genome. As DNA synthesis proceeds, so does RNase H degradation. However, there is a purine-rich sequence in the RNA genome, called the polypurine tract, or ppt, that is resistant to RNase H cleavage and serves as the primer for the initiation of the

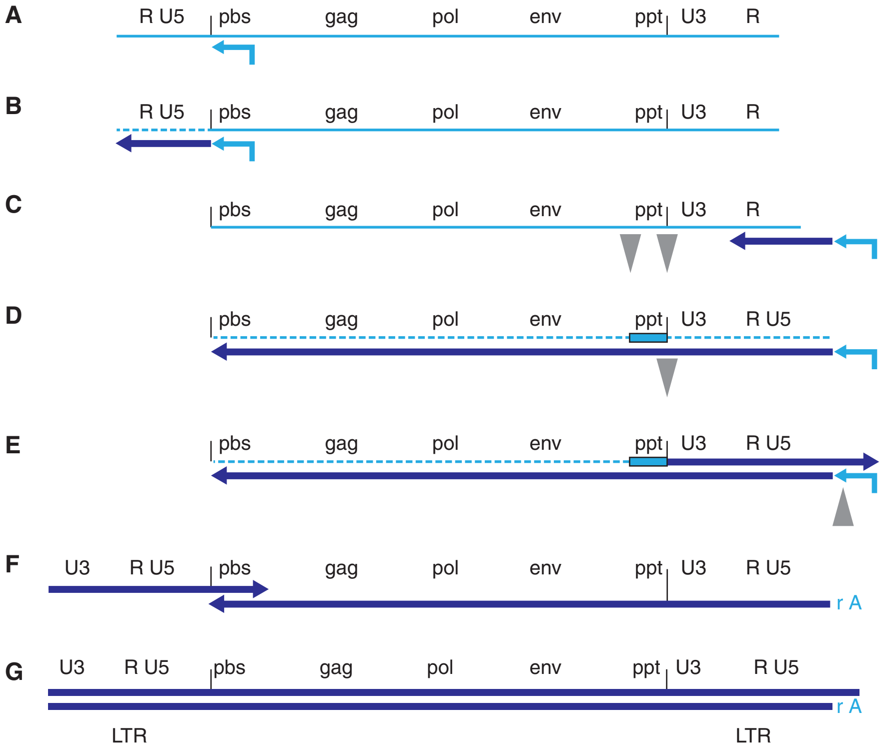


Figure 1. Conversion of the single-stranded RNA genome of a retrovirus into double-stranded DNA. (*A*) The RNA genome of a retrovirus (light blue) with a tRNA primer base paired near the 5′ end. (*B*) RT has initiated reverse transcription, generating minus-strand DNA (dark blue), and the RNase H activity of RT has degraded the RNA template (dashed line). (*C*) Minus-strand transfer has occurred between the R sequences at both ends of the genome (see text), allowing minus-strand DNA synthesis to continue (*D*), accompanied by RNA degradation. A purine-rich sequence (ppt), adjacent to U3, is resistant to RNase H cleavage and serves as the primer for the synthesis of plus-strand DNA (*E*). Plus-strand synthesis continues until the first 18 nucleotides of the tRNA are copied, allowing RNase H cleavage to remove the tRNA primer. Most retroviruses remove the entire tRNA; the RNase H of HIV-1 RT leaves the rA from the 3′ end of the tRNA attached to minus-strand DNA. Removal of the tRNA primer sets the stage for the second (plus-strand) transfer (*F*); extension of the plus and minus strands leads to the synthesis of the complete double-stranded linear viral DNA (*G*).

second (or plus) strand DNA. All retroviruses have at least one ppt. HIV-1 has two, one near the 3′ end of the RNA, the other (the central ppt) near the middle of the genome. The 3′ ppt is essential for viral replication, the central ppt probably increases the ability of the virus to complete plus-strand DNA synthesis, but is not essential (Charneau et al. 1992; Hungnes et al. 1992). When RT generates the plus-stand DNA that is initiated from the 3′ ppt, it not only copies the minus-strand DNA, but also the first 18 nucleotides of the Lys3 tRNA primer. Experiments performed with avian sarcoma-leukosis virus (ASLV) suggest that the ppt-primed plus-strand DNA synthesis stops when it encounters a modified A that RT cannot copy (Swanstrom et al. 1981). It is reasonable to expect that the same mechanism defines the portion of the HIV tRNA primer that is copied. Once the tRNA has been copied into DNA, it becomes a substrate for RNase H. Most retroviruses remove the entire tRNA; however, HIV-1 RT is the exception. It cleaves the tRNA one nucleotide from the 3′ end, leaving a single A ribonucleotide at the 5′ end of the minus strand (the specificity of RNase H cleavage is discussed at the end of this section) (Whitcomb et al. 1990; Pullen et al. 1992; Smith and Roth 1992).

In theory, minus-strand DNA synthesis can proceed along the entire length of the RNA genome; however, the genomic RNAs found in virions are often nicked. The fact that there is a second copy of the RNA genome allows minus-strand DNA synthesis to transfer to the second RNA template, thus bypassing the nick in the original template. This template switching ability contributes to efficient recombination, a topic that is considered later in this article. When minus-strand DNA synthesis nears the 5′ end of the genomic RNA, the pbs is copied, setting the stage for the second, or plus-strand transfer. The 3′ end of the plus-strand DNA contains 18 nucleotides copied from the tRNA primer, which are complementary to 18 nucleotides at the 3′ end of the minus-strand DNA that were copied from the pbs. These two complementary sequences anneal, and DNA synthesis extends both the minus and plus strands to the ends of both templates. The synthesis of plus-strand DNA does not have to be continuous; it is clear that, in ASLV, the plus strand is made in segments (Kung et al. 1981; Hsu and Taylor 1982). It has been reported that HIV-1 plus-strand DNA is also synthesized from multiple initiation sites (Miller et al. 1995; Klarmann et al. 1997; Thomas et al. 2007); however, that raises a question about the role played by the second ppt: If plus-strand DNA is made in segments, what advantage does the second ppt give HIV-1?

The reverse transcription process creates a DNA product that is longer than the RNA genome from which it is derived: both ends of the DNA contain sequences from each end of the RNA (U3 from the 3′ end and U5 from the 5′ end). Thus, each end of the viral DNA has the same sequence, U3-R-U5; these are the long terminal repeats (LTRs) that will, after integration, be the ends of the provirus. It is important to remember that the sequences at the ends of the full-length linear viral DNA are defined, on the U5 end, by the RNase H cleavage that removes the tRNA primer, and on the U3 end, by the cleavages that generate and remove the ppt primer. Despite the fact that RNase H does not have any specific sequence recognition motifs, it cleaves these substrates with single nucleotide specificity, a specificity that appears to be based on the structures of the nucleic acid substrates when they are in a complex with RT (Pullen et al. 1993; Julias et al. 2002; Rausch et al. 2002; Dash et al. 2004; Yi-Brunozzi and Le Grice 2005). The specificity of RNase H cleavage is important because the ends of the linear viral DNA are the substrates for integration. Although DNA substrates whose ends differ modestly from the consensus sequence can be used for retroviral DNA integration, the consensus sequence is the preferred substrate (Colicelli and Goff 1985, 1988; Esposito and Craigie 1998; Oh et al. 2008).