RETROVIRUSES AND RETROPOSONS

4.17.1 Introduction

Key Terms

A **retrovirus** is an RNA virus with the ability to convert its sequence into DNA by reverse transcription.

A **retroposon (retrotransposon)** is a transposon that mobilizes via an RNA form; the DNA element is transcribed into RNA, and then reverse-transcribed into DNA, which is inserted at a new site in the genome. The difference from retroviruses is that the retroposon does not have an infective (viral) form.

Transposition that involves an obligatory intermediate of RNA is unique to eukaryotes, and is provided by the ability of **retroviruses** to insert DNA copies (proviruses) of an RNA viral genome into the chromosomes of a host cell. Some eukaryotic transposons are related to retroviral proviruses in their general organization, and they transpose through RNA intermediates. As a class, these elements are called **retroposons** (or sometimes **retrotransposons**). The very simplest such elements do not themselves have transposition activity, but have sequences that are recognized as substrates for transposition by active elements. So elements that use RNA-dependent transposition range from the retroviruses themselves, able freely to infect host cells, to sequences that transpose via RNA, to those that do not themselves possess the ability to transpose. They share with all transposons the diagnostic feature of generating short direct repeats of target DNA at the site of an insertion.

Even in genomes where active transposons have not been detected, footprints of ancient transposition events are found in the form of direct target repeats flanking dispersed repetitive sequences. The features of these sequences sometimes implicate an RNA sequence as the progenitor of the genomic (DNA) sequence. This suggests that the RNA must have been converted into a duplex DNA copy that was inserted into the genome by a transposition-like event.

Like any other reproductive cycle, the cycle of a retrovirus or retroposon is continuous; it is arbitrary at which point we interrupt it to consider a "beginning." But our perspectives of these elements are biased by the forms in which we usually observe them, indicated in **Figure 17.1**. Retroviruses were first observed as infectious virus particles, capable of transmission between cells, and so the intracellular cycle (involving duplex DNA) is thought of as the means of reproducing the RNA virus. Retroposons were discovered as components of the genome; and the RNA forms have been mostly characterized for their functions as mRNAs. So we think of retroposons as genomic (duplex DNA) sequences that may transpose within a genome; they do not migrate between cells.

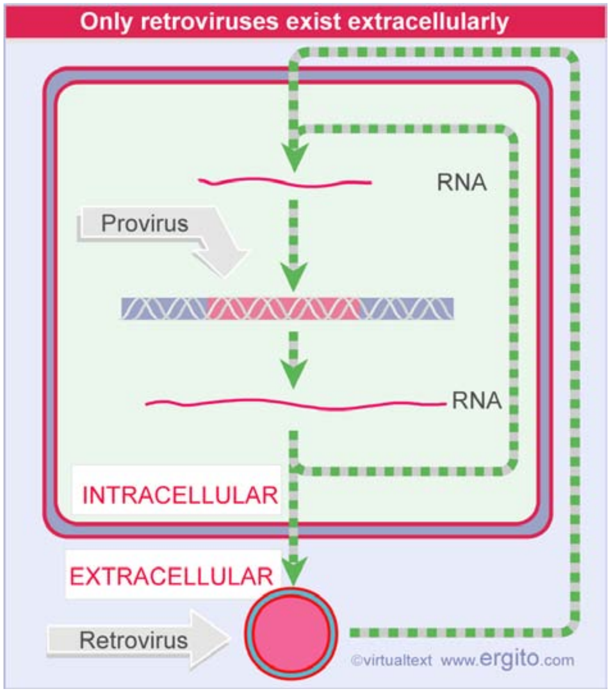

Only retroviruses exist extracellularly
RNA
Provirus
RNA
INTRACELLULAR
EXTRACELLULAR
Retrovirus
©virtualtext www.ergito.com

Figure 17.1 The reproductive cycles of retroviruses and retroposons involve alternation of reverse transcription from RNA to DNA with transcription from DNA to RNA. Only retroviruses can generate infectious particles. Retroposons are confined to an intracellular cycle.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.4.17.1

RETROVIRUSES AND RETROPOSONS

4.17.2 The retrovirus life cycle involves transposition-like events

Key Terms

Provirus is a duplex sequence of DNA integrated into a eukaryotic genome that represents the sequence of the RNA genome of a retrovirus.

Reverse transcriptase is an enzyme that uses a template of single-stranded RNA to generate a double-stranded DNA copy.

An **integrase** is an enzyme that is responsible for a site-specific recombination that inserts one molecule of DNA into another.

Key Concepts

- A retrovirus has two copies of its genome of single-stranded RNA.
- An integrated provirus is a double-stranded DNA sequence.
- A retrovirus generates a provirus by reverse transcription of the retroviral genome.

Retroviruses have genomes of single-stranded RNA that are replicated through a double-stranded DNA intermediate. The life cycle of the virus involves an obligatory stage in which the double-stranded DNA is inserted into the host genome by a transposition-like event that generates short direct repeats of target DNA (for review see 168).

The significance of this reaction extends beyond the perpetuation of the virus. Some of its consequences are that:

- A retroviral sequence that is integrated in the germline remains in the cellular genome as an endogenous **provirus**. Like a lysogenic bacteriophage, a provirus behaves as part of the genetic material of the organism.
- Cellular sequences occasionally recombine with the retroviral sequence and then are transposed with it; these sequences may be inserted into the genome as duplex sequences in new locations.
- Cellular sequences that are transposed by a retrovirus may change the properties of a cell that becomes infected with the virus.

The particulars of the retroviral life cycle are expanded in **Figure 17.2**. The crucial steps are that the viral RNA is converted into DNA, the DNA becomes integrated into the host genome, and then the DNA provirus is transcribed into RNA.

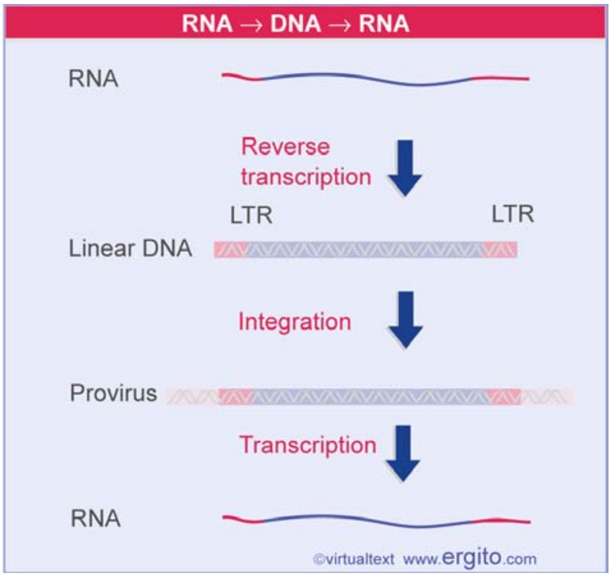


Figure 17.2 The retroviral life cycle proceeds by reverse transcribing the RNA genome into duplex DNA, which is inserted into the host genome, in order to be transcribed into RNA.

The enzyme responsible for generating the initial DNA copy of the RNA is **reverse transcriptase**. The enzyme converts the RNA into a linear duplex of DNA in the cytoplasm of the infected cell. The DNA also is converted into circular forms, but these do not appear to be involved in reproduction (573; 574).

The linear DNA makes its way to the nucleus. One or more DNA copies become integrated into the host genome. A single enzyme, called **integrase**, is responsible for integration. The provirus is transcribed by the host machinery to produce viral RNAs, which serve both as mRNAs and as genomes for packaging into virions. Integration is a normal part of the life cycle and is necessary for transcription.

Two copies of the RNA genome are packaged into each virion, making the individual virus particle effectively diploid. When a cell is simultaneously infected by two different but related viruses, it is possible to generate heterozygous virus particles carrying one genome of each type. The diploidy may be important in allowing the virus to acquire cellular sequences. The enzymes reverse transcriptase and integrase are carried with the genome in the viral particle.

Reviews

168. Varmus, H. and Brown, P. (1989). *Retroviruses*. Mobile DNA, 3-108.

References

573. Temin, H. M. and Mizutani, S. (1970). *RNA-dependent DNA polymerase in virions of Rous sarcoma virus*. Nature 226, 1211-1213.

574. Baltimore, D. (1970). *RNA-dependent DNA polymerase in virions of RNA tumor viruses*. Nature 226, 1209-1211.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.4.17.2

RETROVIRUSES AND RETROPOSONS

4.17.3 Retroviral genes codes for polyproteins

Key Concepts

- A typical retrovirus has three genes, *gag*, *pol*, *env*.
- Gag and Pol proteins are translated from a full-length transcript of the genome.
- Translation of Pol requires a frameshift by the ribosome.
- Env is translated from a separate mRNA that is generated by splicing.
- Each of the three protein products is processed by proteases to give multiple proteins.

A typical retroviral sequence contains three or four "genes," the term here identifying coding regions each of which actually gives rise to multiple proteins by processing reactions. A typical retrovirus genome with three genes is organized in the sequence *gag-pol-env* as indicated in **Figure 17.3**.

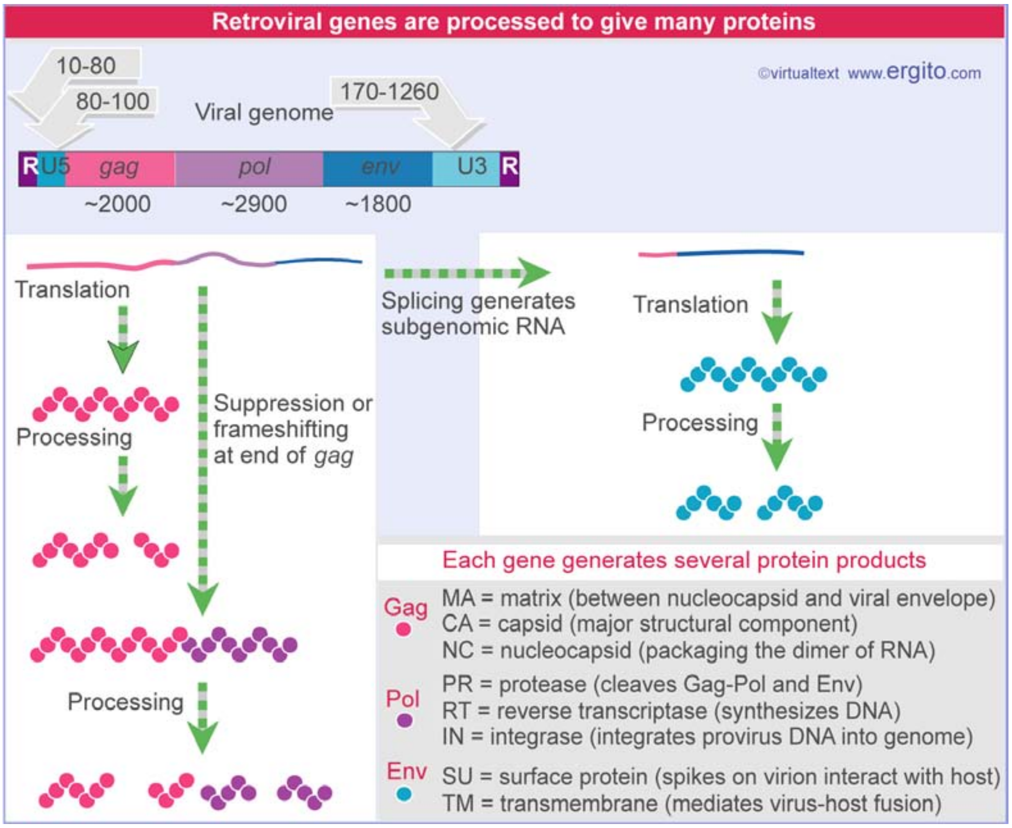

Retroviral genes are processed to give many proteins
©virtualtext www.ergito.com
10-80
80-100
Viral genome
170-1260
R
U5
gag
pol
env
U3
R
~2000
~2900
~1800
Translation
Processing
Suppression or frameshifting at end of gag
Processing
Splicing generates subgenomic RNA
Translation
Processing
Each gene generates several protein products
Gag
MA = matrix (between nucleocapsid and viral envelope)
CA = capsid (major structural component)
NC = nucleocapsid (packaging the dimer of RNA)
Pol
PR = protease (cleaves Gag-Pol and Env)
RT = reverse transcriptase (synthesizes DNA)
IN = integrase (integrates provirus DNA into genome)
Env
SU = surface protein (spikes on virion interact with host)
TM = transmembrane (mediates virus-host fusion)

Figure 17.3 The genes of the retrovirus are expressed as polyproteins that are processed into individual products.

Retroviral mRNA has a conventional structure; it is capped at the 5 ′ end and polyadenylated at the 3 ′ end. It is represented in two mRNAs. The full length mRNA is translated to give the Gag and Pol polyproteins. The Gag product is translated by reading from the initiation codon to the first termination codon. This termination codon must be bypassed to express Pol.

Different mechanisms are used in different viruses to proceed beyond the *gag* termination codon, depending on the relationship between the *gag* and *pol* reading frames. When *gag* and *pol* follow continuously, suppression by a glutamyl-tRNA that recognizes the termination codon allows a single protein to be generated. When *gag* and *pol* are in different reading frames, a ribosomal frameshift occurs to generate a single protein. Usually the readthrough is ~5% efficient, so Gag protein outnumbers Gag-Pol protein about 20-fold.

The Env polyprotein is expressed by another means: splicing generates a shorter *subgenomic* messenger that is translated into the Env product.

The *gag* gene gives rise to the protein components of the nucleoprotein core of the virion. The *pol* gene codes for functions concerned with nucleic acid synthesis and recombination. The *env* gene codes for components of the envelope of the particle, which also sequesters components from the cellular cytoplasmic membrane.

Both the Gag or Gag-Pol and the Env products are polyproteins that are cleaved by a protease to release the individual proteins that are found in mature virions. The protease activity is coded by the virus in various forms: it may be part of Gag or Pol, or sometimes takes the form of an additional independent reading frame

The production of a retroviral particle involves packaging the RNA into a core, surrounding it with capsid proteins, and pinching off a segment of membrane from the host cell. The release of infective particles by such means is shown in **Figure 17.4**. The process is reversed during infection; a virus infects a new host cell by fusing with the plasma membrane and then releasing the contents of the virion.

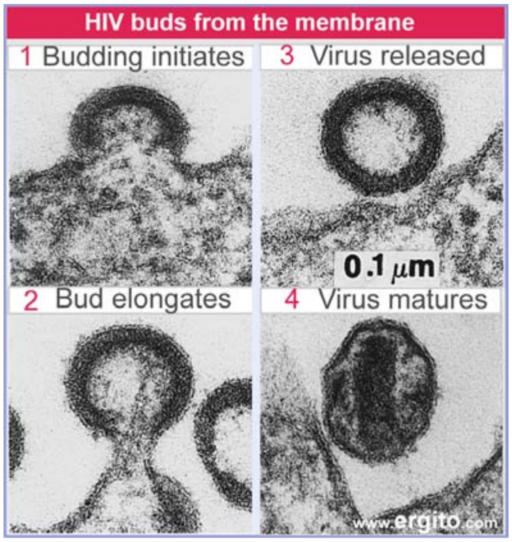


Figure 17.4 Retroviruses (HIV) bud from the plasma membrane of an infected cell. Photograph kindly provided by Matthew Gonda.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.4.17.3

RETROVIRUSES AND RETROPOSONS

4.17.4 Viral DNA is generated by reverse transcription

Key Terms

A **plus strand virus** has a single-stranded nucleic acid genome whose sequence directly codes for the protein products.

Minus strand DNA is the single-stranded DNA sequence that is complementary to the viral RNA genome of a plus strand virus.

Plus strand DNA is the strand of the duplex sequence representing a retrovirus that has the same sequence as that of the RNA.

The **R segments** are the sequences that are repeated at the ends of a retroviral RNA. They are called R-U5 and U3-R.

U5 is the repeated sequence at the 5′ end of a retroviral RNA.

U3 is the repeated sequence at the 3′ end of a retroviral RNA.

The **long terminal repeat (LTR)** is the sequence that is repeated at each end of the integrated retroviral genome.

Copy choice is a type of recombination used by RNA viruses, in which the RNA polymerase switches from one template to another during synthesis.

Key Concepts

- A short sequence (R) is repeated at each end of the viral RNA, so the 5′ and 3′ ends respectively are R-U5 and U3-R.
- Reverse transcriptase starts synthesis when a tRNA primer binds to a site 100-200 bases from the 5′ end.
- When the enzyme reaches the end, the 5′–terminal bases of RNA are degraded, exposing the 3′ end of the DNA product.
- The exposed 3′ end base pairs with the 3′ terminus of another RNA genome.
- Synthesis continues, generating a product in which the 5′ and 3′ regions are repeated, giving each end the structure U3-R-U5.
- Similar strand switching events occur when reverse transcriptase uses the DNA product to generate a complementary strand.
- Strand switching is an example of the copy choice mechanism of recombination.

Retroviruses are called **plus strand viruses**, because the viral RNA itself codes for the protein products. As its name implies, reverse transcriptase is responsible for converting the genome (plus strand RNA) into a complementary DNA strand, which is called the **minus strand DNA**. Reverse transcriptase also catalyzes subsequent stages in the production of duplex DNA. It has a DNA polymerase activity, which enables it to synthesize a duplex DNA from the single-stranded reverse transcript of the RNA. The second DNA strand in this duplex is called **plus strand DNA**. And as a necessary adjunct to this activity, the enzyme has an RNAase H activity, which can degrade the RNA part of the RNA-DNA hybrid. All retroviral reverse transcriptases share considerable similarities of amino acid sequence, and homologous sequences can be recognized in some other retroposons (see later; for review see 171).

The structures of the DNA forms of the virus are compared with the RNA in **Figure 17.5**. The viral RNA has direct repeats at its ends. These **R segments** vary in different strains of virus from 10-80 nucleotides. The sequence at the 5 ′ end of the virus is R-**U5**, and the sequence at the 3 ′ end is **U3**-R. The R segments are used during the conversion from the RNA to the DNA form to generate the more extensive direct repeats that are found in linear DNA (see **Figure 17.6** and **Figure 17.7**). The shortening of 2 bp at each end in the integrated form is a consequence of the mechanism of integration (see **Figure 17.9**).

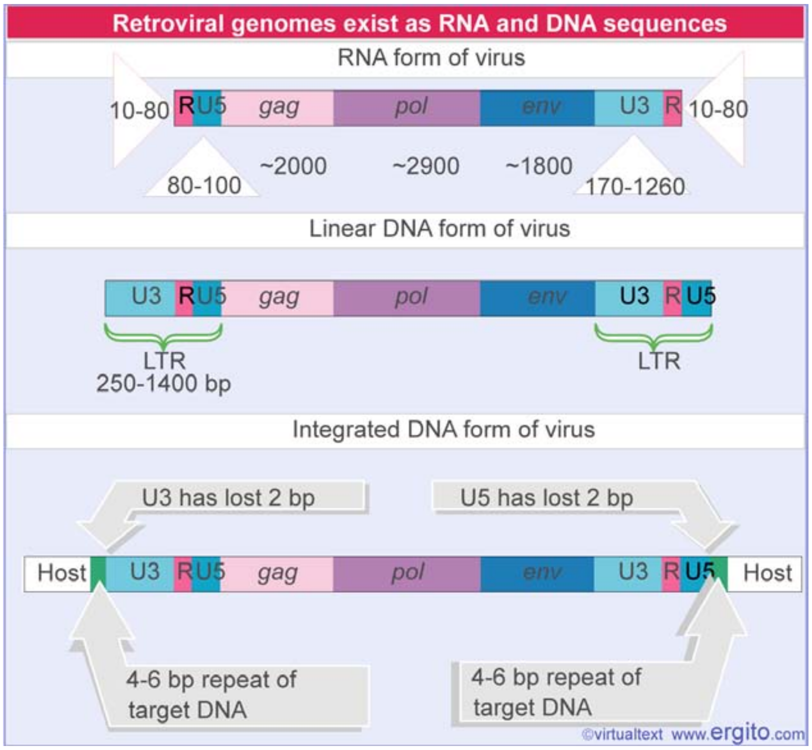


Figure 17.5 Retroviral RNA ends in direct repeats (R), the free linear DNA ends in LTRs, and the provirus ends in LTRs that are shortened by two bases each.

Like other DNA polymerases, reverse transcriptase requires a primer. The native primer is tRNA. An uncharged host tRNA is present in the virion. A sequence of 18 bases at the 3 ′ end of the tRNA is base paired to a site 100-200 bases from the 5 ′ end of one of the viral RNA molecules. The tRNA may also be base paired to another site near the 5 ′ end of the other viral RNA, thus assisting in dimer formation between the viral RNAs.

Here is a dilemma. Reverse transcriptase starts to synthesize DNA at a site only 100-200 bases downstream from the 5 ′ end. How can DNA be generated to represent the intact RNA genome? (This is an extreme variant of the general problem in replicating the ends of any linear nucleic acid; see *Molecular Biology 4.13.8 The ends of linear DNA are a problem for replication* .)

Synthesis *in vitro* proceeds to the end, generating a short DNA sequence called minus strong-stop DNA. This molecule is not found *in vivo* because synthesis continues by the reaction illustrated in **Figure 17.6**. Reverse transcriptase switches templates, carrying the nascent DNA with it to the new template. This is the first of two jumps between templates.

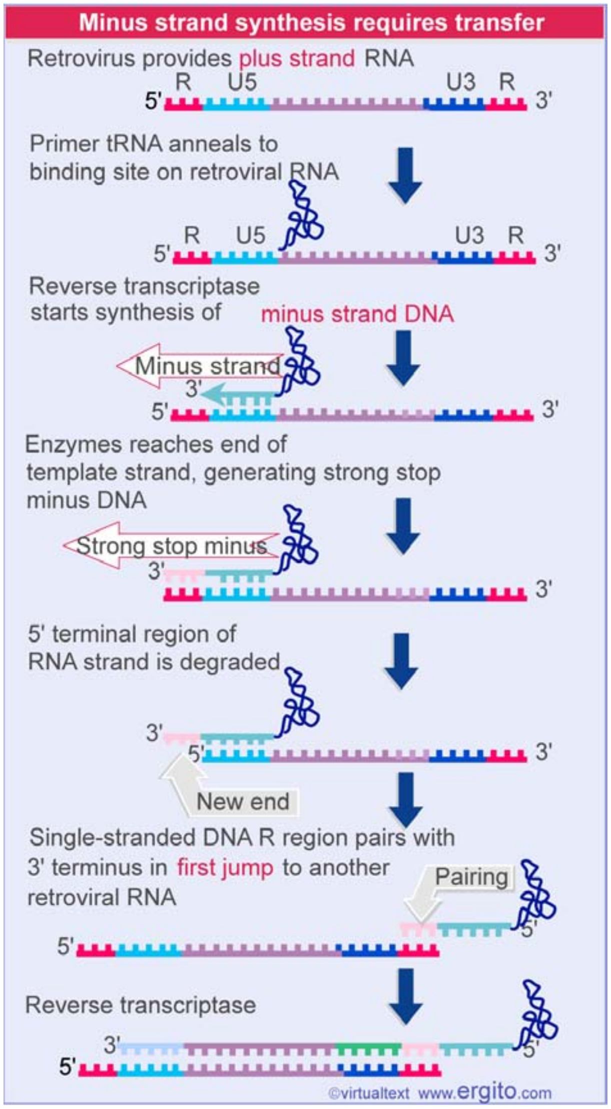

Minus strand synthesis requires transfer
Retrovirus provides plus strand RNA
R
U5
U3
R
5'
3'
Primer tRNA anneals to binding site on retroviral RNA
Reverse transcriptase starts synthesis of minus strand DNA
Minus strand
Enzymes reaches end of template strand, generating strong stop minus DNA
Strong stop minus
5' terminal region of RNA strand is degraded
New end
Single-stranded DNA R region pairs with 3' terminus in first jump to another retroviral RNA
Pairing
Reverse transcriptase
©virtualtext www.ergito.com

Figure 17.6 Minus strand DNA is generated by switching templates during reverse transcription.

In this reaction, the R region at the 5 ′ terminus of the RNA template is degraded by the RNAase H activity of reverse transcriptase. Its removal allows the R region at a 3 ′ end to base pair with the newly synthesized DNA. Then reverse transcription continues through the U3 region into the body of the RNA.

The source of the R region that pairs with the strong stop minus DNA can be either the 3 ′ end of the same RNA molecule (intramolecular pairing) or the 3 ′ end of a different RNA molecule (intermolecular pairing). The switch to a different RNA template is used in the figure because there is evidence that the sequence of the tRNA primer is not inherited in a retroposon life cycle. (If intramolecular pairing occurred, we would expect the sequence to be inherited, because it would provide the only source for the primer binding sequence in the next cycle. Intermolecular pairing allows another retroviral RNA to provide this sequence.)

The result of the switch and extension is to add a U3 segment to the 5 ′ end. The stretch of sequence U3-R-U5 is called the **long terminal repeat** (**LTR**) because a similar series of events adds a U5 segment to the 3 ′ end, giving it the same structure of U5-R-U3. Its length varies from 250-1400 bp (see **Figure 17.5**).

We now need to generate the plus strand of DNA and to generate the LTR at the other end. The reaction is shown in **Figure 17.7**. Reverse transcriptase primes synthesis of plus strand DNA from a fragment of RNA that is left after degrading the original RNA molecule. A strong stop plus strand DNA is generated when the enzyme reaches the end of the template. This DNA is then transferred to the other end of a minus strand. Probably it is released by a displacement reaction when a second round of DNA synthesis occurs from a primer fragment farther upstream (to its left in the figure). It uses the R region to pair with the 3 ′ end of a minus strand DNA. This double-stranded DNA then requires completion of both strands to generate a duplex LTR at each end.

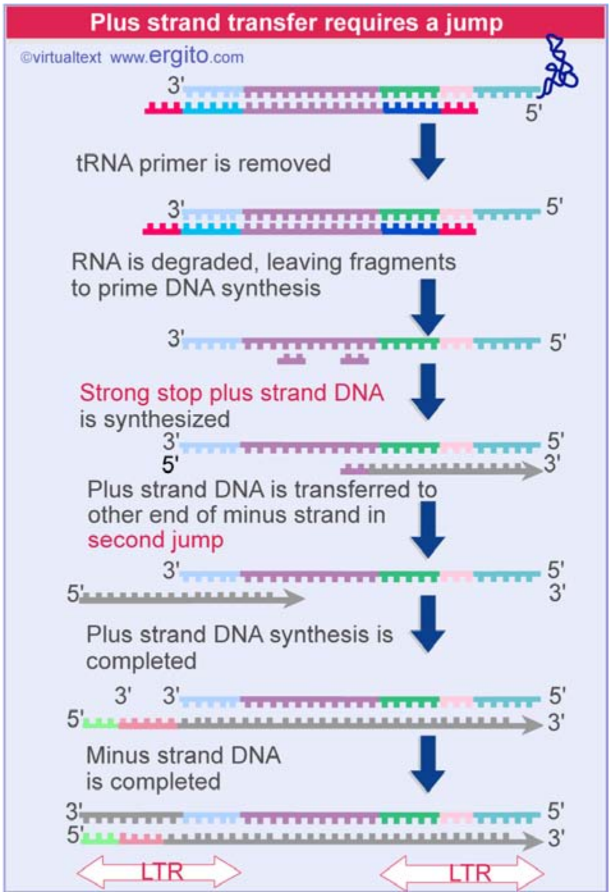

Plus strand transfer requires a jump
©virtualtext www.ergito.com
3'
5'
tRNA primer is removed
3'
5'
RNA is degraded, leaving fragments to prime DNA synthesis
3'
5'
Strong stop plus strand DNA is synthesized
3'
5'
5'
3'
Plus strand DNA is transferred to other end of minus strand in second jump
3'
5'
5'
3'
Plus strand DNA synthesis is completed
3'
3'
5'
5'
3'
Minus strand DNA is completed
3'
5'
5'
3'
LTR
LTR

Figure 17.7 Synthesis of plus strand DNA requires a second jump.

Each retroviral particle carries two RNA genomes. This makes it possible for recombination to occur during a viral life cycle. In principle this could occur during minus strand synthesis and/or during plus strand synthesis:

- The intermolecular pairing shown in **Figure 17.6** allows a recombination to occur between sequences of the two successive RNA templates when minus strand DNA is synthesized. Retroviral recombination is mostly due to strand transfer at this stage, when the nascent DNA strand is transferred from one RNA template to another during reverse transcription (for review see 2292).
- Plus strand DNA may be synthesized discontinuously, in a reaction that involves several internal initiations. Strand transfer during this reaction can also occur, but is less common.

The common feature of both events is that recombination results from a change in the template during the act of DNA synthesis. This is a general example of a mechanism for recombination called **copy choice**. For many years this was regarded as a possible mechanism for general recombination. It is unlikely to be employed by cellular systems, but is a common basis for recombination during infection by RNA viruses, including those that replicate exclusively through RNA forms, such as poliovirus (579; for review see 170).

Strand switching occurs with a certain frequency during each cycle of reverse transcription, that is, in addition to the transfer reaction that is- forced at the end of the template strand. The principle is illustrated in **Figure 17.8**, although we do not know much about the mechanism. Reverse transcription *in vivo* occurs in a ribonucleoprotein complex, in which the RNA template strand is bound to virion components, including the major protein of the capsid. In the case of HIV, addition of this protein (NCp7) to an *in vitro* system causes recombination to occur (1181). The effect is probably indirect: NCp7 affects the structure of the RNA template, which in turn affects the likelihood that reverse transcriptase will switch from one template strand to another.

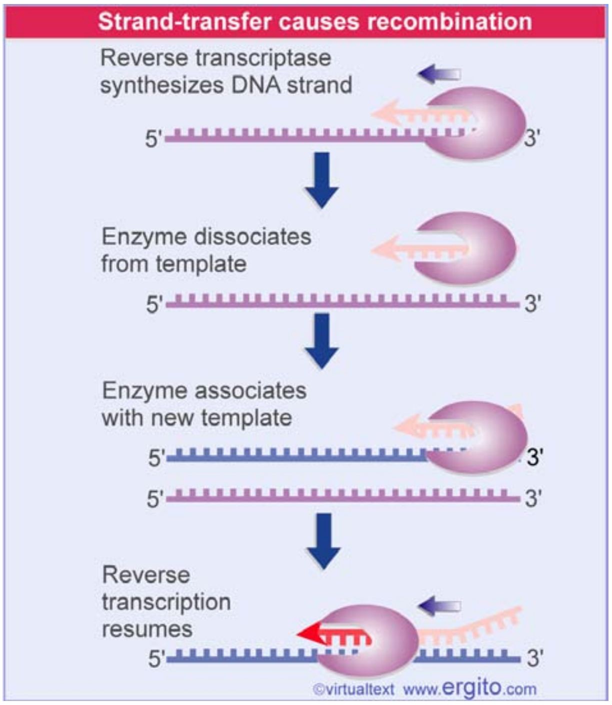

Strand-transfer causes recombination
Reverse transcriptase synthesizes DNA strand
5'
3'
Enzyme dissociates from template
5'
3'
Enzyme associates with new template
5'
3'
5'
3'
Reverse transcription resumes
5'
3'
©virtualtext www.ergito.com

Figure 17.8 Copy choice recombination occurs when reverse transcriptase releases its template and resumes DNA synthesis using a new template. Transfer between template strands probably occurs directly, but is shown here in separate steps to illustrate the process.

Last updated on 1-22-2002

Reviews

170. Lai, M. M. C. (1992). *RNA recombination in animal and plant viruses.* Microbiol. Rev. 56, 61-79.

171. Katz, R. A. and Skalka, A. M. (1994). *The retroviral enzymes.* Annu. Rev. Biochem. 63, 133-173.

2292. Negroni, M. and Buc, H. (2001). *Mechanisms of retroviral recombination.* Annu. Rev. Genet. 35, 275-302.

References

579. Hu, W. S. and Temin, H. M. (1990). *Retroviral recombination and reverse transcription*. Science 250, 1227-1233.

1181. Negroni, M. and Buc, H. (2000). *Copy-choice recombination by reverse transcriptases: reshuffling of genetic markers mediated by RNA chaperones.* Proc. Natl. Acad. Sci. USA 97, 6385-6390.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.4.17.4

RETROVIRUSES AND RETROPOSONS

4.17.5 Viral DNA integrates into the chromosome

Key Concepts

- The organization of proviral DNA in a chromosome is the same as a transposon, with the provirus flanked by short direct repeats of a sequence at the target site.
- Linear DNA is inserted directly into the host chromosome by the retroviral integrase enzyme.
- Two base pairs of DNA are lost from each end of the retroviral sequence during the integration reaction.

The organization of the integrated provirus resembles that of the linear DNA. The LTRs at each end of the provirus are identical. The 3 ′ end of U5 consists of a short inverted repeat relative to the 5 ′ end of U3, so the LTR itself ends in short inverted repeats. The integrated proviral DNA is like a transposon: the proviral sequence ends in inverted repeats and is flanked by short direct repeats of target DNA.

The provirus is generated by directly inserting a linear DNA into a target site. [In addition to linear DNA, there are circular forms of the viral sequences. One has two adjacent LTR sequences generated by joining the linear ends. The other has only one LTR [presumably generated by a recombination event and actually comprising the majority of circles. Although for a long time it appeared that the circle might be an integration intermediate (by analogy with the integration of lambda DNA), we now know that the linear form is used for integration; for review see 169).]

Integration of linear DNA is catalyzed by a single viral product, the integrase. Integrase acts on both the retroviral linear DNA and the target DNA. The reaction is illustrated in **Figure 17.9**.

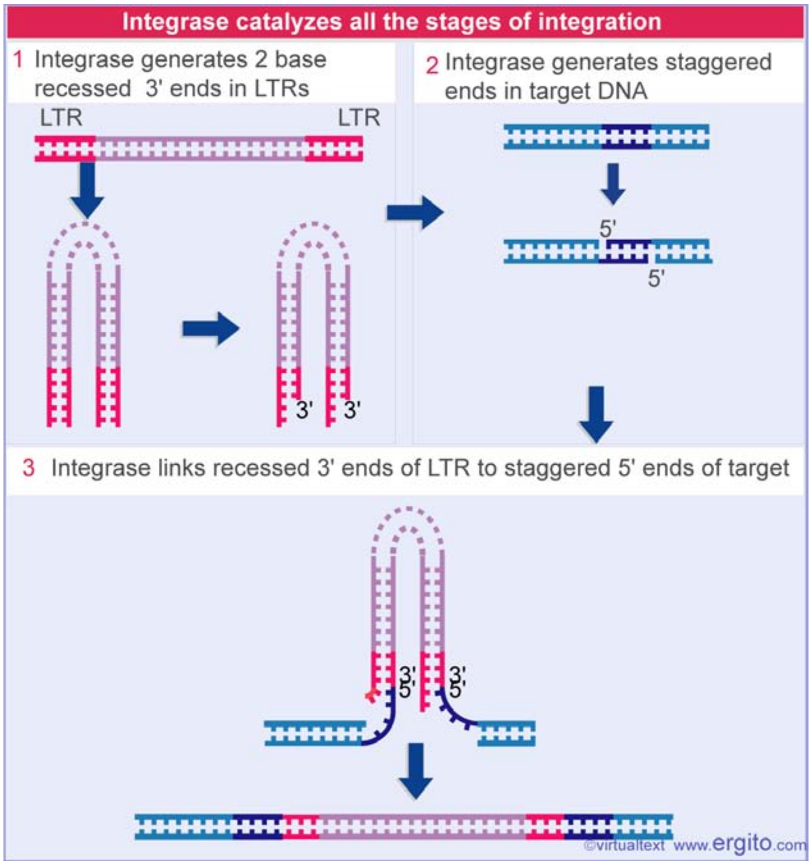

Integrase catalyzes all the stages of integration
1 Integrase generates 2 base recessed 3' ends in LTRs
LTR
LTR
3'
3'
2 Integrase generates staggered ends in target DNA
5'
5'
3 Integrase links recessed 3' ends of LTR to staggered 5' ends of target
3'
5'
3'
5'
©virtualtext www.ergito.com

Figure 17.9 Integrase is the only viral protein required for the integration reaction, in which each LTR loses 2 bp and is inserted between 4 bp repeats of target DNA.

The ends of the viral DNA are important; as is the case with transposons, mutations in the ends prevent integration. The most conserved feature is the presence of the dinucleotide sequence CA close to the end of each inverted repeat. The integrase brings the ends of the linear DNA together in a ribonucleoprotein complex, and converts the blunt ends into recessed ends by removing the bases beyond the conserved CA; usually this involves loss of 2 bases (578).

Target sites are chosen at random with respect to sequence. The integrase makes staggered cuts at a target site. In the example of **Figure 17.9**, the cuts are separated by 4 bp. The length of the target repeat depends on the particular virus; it may be 4, 5, or 6 bp. Presumably it is determined by the geometry of the reaction of integrase with target DNA.

The 5′ ends generated by the cleavage of target DNA are covalently joined to the 3′ recessed ends of the viral DNA. At this point, both termini of the viral DNA are joined by one strand to the target DNA. The single-stranded region is repaired by enzymes of the host cell, and in the course of this reaction the protruding 2 bases at each 5′ end of the viral DNA are removed. The result is that the integrated viral DNA has lost 2 bp at each LTR; this corresponds to the loss of 2 bp from the left end of the 5′ terminal U3 and loss of 2 bp from the right end of the 3′ terminal U5. There is a characteristic short direct repeat of target DNA at each end of the integrated retroviral genome.

The viral DNA integrates into the host genome at randomly selected sites. A successfully infected cell gains 1-10 copies of the provirus. (An infectious virus enters the cytoplasm, of course, but the DNA form becomes integrated into the genome in the nucleus. Retroviruses can replicate only in proliferating cells, because entry into the nucleus requires the cell to pass through mitosis, when the viral genome gains access to the nuclear material.)

The U3 region of each LTR carries a promoter. The promoter in the left LTR is responsible for initiating transcription of the provirus. Recall that the generation of proviral DNA is required to place the U3 sequence at the left LTR; so we see that the promoter is in fact generated by the conversion of the RNA into duplex DNA.

Sometimes (probably rather rarely), the promoter in the right LTR sponsors transcription of the host sequences that are adjacent to the site of integration. The LTR also carries an enhancer (a sequence that activates promoters in the vicinity) that can act on cellular as well as viral sequences. Integration of a retrovirus can be responsible for converting a host cell into a tumorigenic state when certain types of genes are activated in this way (see *Molecular Biology 6.30.6 Retroviruses activate or incorporate cellular genes*).

Can integrated proviruses be excised from the genome? Homologous recombination could take place between the LTRs of a provirus; solitary LTRs that could be relics of an excision event are present in some cellular genomes.

We have dealt so far with retroviruses in terms of the infective cycle, in which integration is necessary for the production of further copies of the RNA. However, when a viral DNA integrates in a germline cell, it becomes an inherited "endogenous provirus" of the organism. Endogenous viruses usually are not expressed, but sometimes they are activated by external events, such as infection with another virus.

Last updated on 10-3-2000

Reviews

169. Goff, S. P. (1992). *Genetics of retroviral integration.* Annu. Rev. Genet. 26, 527-544.

References

578. Craigie, R., Fujiwara, T., and Bushman, F. (1990). *The IN protein of Moloney murine leukemia virus processes the viral DNA ends and accomplishes their integrationin vitro.* Cell 62, 829-837.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.4.17.5

RETROVIRUSES AND RETROPOSONS

4.17.6 Retroviruses may transduce cellular sequences

Key Terms

A **transducing virus** carries part of the host genome in place of part of its own sequence. The best known examples are retroviruses in eukaryotes and DNA phages in *E. coli*.

A **replication-defective** virus cannot perpetuate an infective cycle because some of the necessary genes are absent (replaced by host DNA in a transducing virus) or mutated.

A **helper virus** provides functions absent from a defective virus, enabling the latter to complete the infective cycle during a mixed infection.

Transformation (Oncogenesis) of eukaryotic cells refers to their conversion to a state of unrestrained growth in culture, resembling or identical with the tumorigenic condition.

Key Concepts

- Transforming retroviruses are generated by a recombination event in which a cellular RNA sequence replaces part of the retroviral RNA.

An interesting light on the viral life cycle is cast by the occurrence of **transducing viruses**, variants that have acquired cellular sequences in the form illustrated in **Figure 17.10**. Part of the viral sequence has been replaced by the *v-onc* gene. Protein synthesis generates a Gag-v-Onc protein instead of the usual Gag, Pol, and Env proteins. The resulting virus is **replication-defective**; it cannot sustain an infective cycle by itself. However, it can be perpetuated in the company of a **helper virus** that provides the missing viral functions.

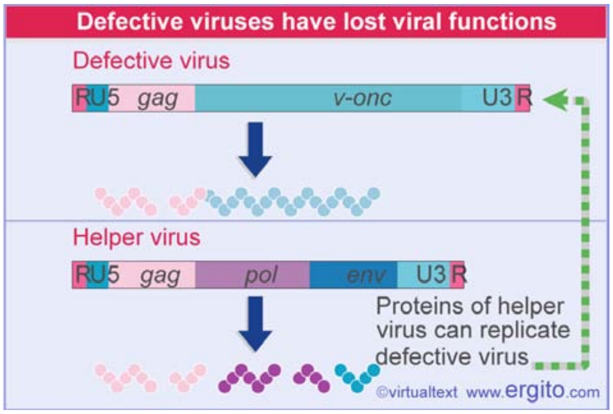

Defective viruses have lost viral functions
Defective virus
R U5 gag v-onc U3 R
Helper virus
R U5 gag pol env U3 R
Proteins of helper virus can replicate defective virus
©virtualtext www.ergito.com

Figure 17.10 Replication-defective transforming viruses have a cellular sequence substituted for part of the viral sequence. The defective virus may replicate with the assistance of a helper virus that carries the wild-type functions.

Onc is an abbreviation for **oncogenesis**, the ability to *transform* cultured cells so that the usual regulation of growth is released to allow unrestricted division. Both viral and cellular *onc* genes may be responsible for creating tumorigenic cells (see *Molecular Biology 6.30.7 Retroviral oncogenes have cellular counterparts*).

A *v-onc* gene confers upon a virus the ability to transform a certain type of host cell. Loci with homologous sequences found in the host genome are called *c-onc* genes. How are the *onc* genes acquired by the retroviruses? A revealing feature is the discrepancy in the structures of *c-onc* and *v-onc* genes. The *c-onc* genes usually are interrupted by introns, but the *v-onc* genes are uninterrupted. This suggests that the *v-onc* genes originate from spliced RNA copies of the *c-onc* genes.

A model for the formation of transforming viruses is illustrated in **Figure 17.11**. A retrovirus has integrated near a *c-onc* gene. A deletion occurs to fuse the provirus to the *c-onc* gene; then transcription generates a joint RNA, containing viral sequences at one end and cellular *onc* sequences at the other end. Splicing removes the introns in both the viral and cellular parts of the RNA. The RNA has the appropriate signals for packaging into the virion; virions will be generated if the cell also contains another, intact copy of the provirus. Then some of the diploid virus particles may contain one fused RNA and one viral RNA.

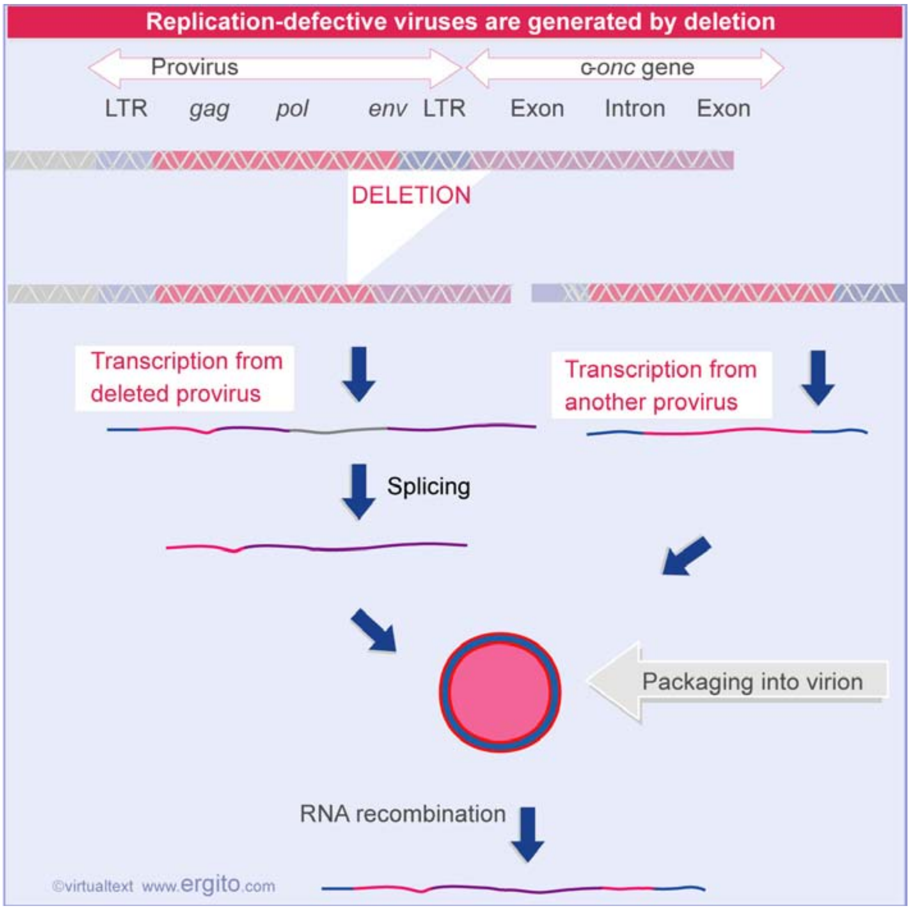

Replication-defective viruses are generated by deletion
Provirus
c-onc gene
LTR
gag
pol
env
LTR
Exon
Intron
Exon
DELETION
Transcription from deleted provirus
Transcription from another provirus
Splicing
Packaging into virion
RNA recombination
©virtualtext www.ergito.com

Figure 17.11 Replication-defective viruses may be generated through integration and deletion of a viral genome to generate a fused viral-cellular transcript that is packaged with a normal RNA genome. Nonhomologous recombination is necessary to generate the replication-defective transforming genome.

A recombination between these sequences could generate the transforming genome, in which the viral repeats are present at both ends. (Recombination occurs at a high frequency during the retroviral infective cycle, by various means. We do not know anything about its demands for homology in the substrates, but we assume that the nonhomologous reaction between a viral genome and the cellular part of the fused RNA proceeds by the same mechanisms responsible for viral recombination.)

The common features of the entire retroviral class suggest that it may be derived from a single ancestor. Primordial IS elements could have surrounded a host gene for a nucleic acid polymerase; the resulting unit would have the form LTR-pol-LTR. It might evolve into an infectious virus by acquiring more sophisticated abilities to manipulate both DNA and RNA substrates, including the incorporation of genes whose products allowed packaging of the RNA. Other functions, such as transforming genes, might be incorporated later. (There is no reason to suppose that the mechanism involved in acquisition of cellular functions is unique for *onc* genes; but viruses carrying these genes may have a selective advantage because of their stimulatory effect on cell growth.)

RETROVIRUSES AND RETROPOSONS

4.17.7 Yeast Ty elements resemble retroviruses

Key Terms

Ty stands for transposon yeast, the first transposable element to be identified in yeast.

A **retroposon (retrotransposon)** is a transposon that mobilizes via an RNA form; the DNA element is transcribed into RNA, and then reverse-transcribed into DNA, which is inserted at a new site in the genome. The difference from retroviruses is that the retroposon does not have an infective (viral) form.

Key Concepts

- Ty transposons have a similar organization to endogenous retroviruses.
- They are retroposons, with a reverse transcriptase activity, that transpose via an RNA intermediate.

Ty elements comprise a family of dispersed repetitive DNA sequences that are found at different sites in different strains of yeast. **Ty** is an abbreviation for "transposon yeast." A transposition event creates a characteristic footprint: 5 bp of target DNA are repeated on either side of the inserted *Ty* element. Ty elements are **retroposons** that transpose by the same mechanism as retroviruses. The frequency of *Ty* transposition is lower than that of most bacterial transposons, ~10^{-7}-10^{-8}.

There is considerable divergence between individual *Ty* elements. Most elements fall into one of two major classes, called *Ty1* and *Ty917*. They have the same general organization illustrated in **Figure 17.12**. Each element is 6.3 kb long; the last 330 bp at each end constitute direct repeats, called δ. Individual *Ty* elements of each type have many changes from the prototype of their class, including base pair substitutions, insertions, and deletions. There are ~30 copies of the *Ty1* type and ~6 of the *Ty917* type in a typical yeast genome. In addition, there are ~100 independent *delta* elements, called solo δs.

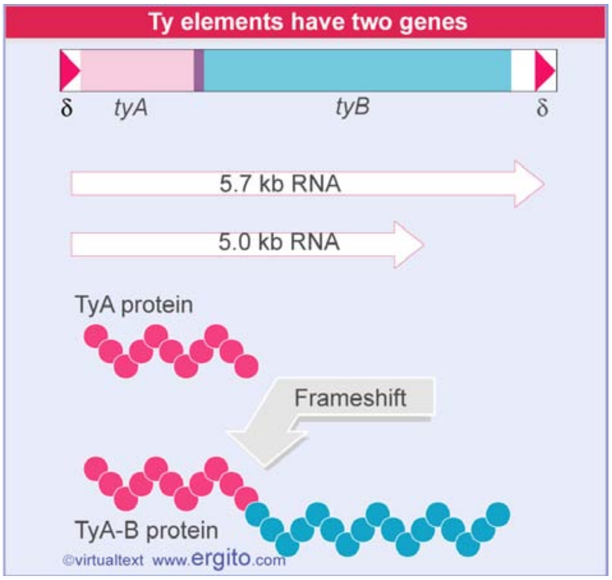


Figure 17.12 Ty elements terminate in short direct repeats and are transcribed into two overlapping RNAs. They have two reading frames, with sequences related to the retroviral *gag* and *pol* genes.

The *delta* sequences also show considerable heterogeneity, although the two repeats of an individual *Ty* element are likely to be identical or at least very closely related. The *delta* sequences associated with *Ty* elements show greater conservation of sequence than the solo *delta* elements, which suggests that recognition of the repeats is involved in transposition.

The *Ty* element is transcribed into two poly(A)$^+$ RNA species, which constitute >5% of the total mRNA of a haploid yeast cell. Both initiate within a promoter in the δ element at the left end. One terminates after 5 kb; the other terminates after 5.7 kb, within the delta sequence at the right end.

The sequence of the *Ty* element has two open reading frames, expressed in the same direction, but read in different phases and overlapping by 13 amino acids. The sequence of *TyA* suggests that it codes for a DNA-binding protein. The sequence of *TyB* contains regions that have homologies with reverse transcriptase, protease, and integrase sequences of retroviruses.

The organization and functions of *TyA* and *TyB* are analogous to the behavior of the retroviral *gag* and *pol* functions. The reading frames *TyA* and *TyB* are expressed in two forms. The TyA protein represents the *TyA* reading frame, and terminates at its end. The *TyB* reading frame, however, is expressed only as part of a joint protein, in which the TyA region is fused to the TyB region by a specific frameshift event that allows the termination codon to be bypassed (analogous to *gag-pol* translation in retroviruses).

Recombination between *Ty* elements seems to occur in bursts; when one event is detected, there is an increased probability of finding others. Gene conversion occurs between *Ty* elements at different locations, with the result that one element is "replaced" by the sequence of the other.

Ty elements can excise by homologous recombination between the directly repeated *delta* sequences. The large number of solo *delta* elements may be footprints of such events. An excision of this nature may be associated with reversion of a mutation caused by the insertion of *Ty*; the level of reversion may depend on the exact *delta* sequences left behind.

A paradox is that both *delta* elements have the same sequence, yet a promoter is active in the *delta* at one end and a terminator is active in the *delta* at the other end. (A similar feature is found in other transposable elements, including the retroviruses.)

Ty elements are classic retroposons, transposing through an RNA intermediate. An ingenious protocol used to detect this event is illustrated in **Figure 17.13**. An intron was inserted into an element to generate a unique *Ty* sequence. This sequence was placed under the control of a *GAL* promoter on a plasmid and introduced into yeast cells. Transposition results in the appearance of multiple copies of the transposon in the yeast genome; but they all lack the intron (575).

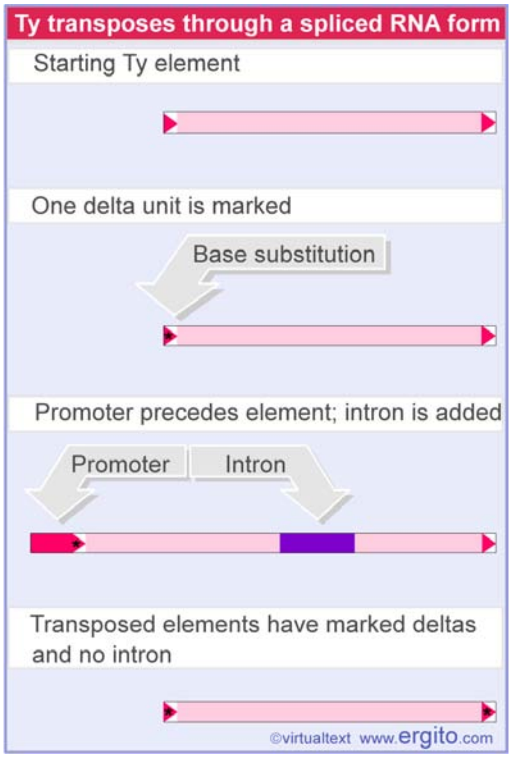

Ty transposes through a spliced RNA form
Starting Ty element
One delta unit is marked
Base substitution
Promoter precedes element; intron is added
Promoter
Intron
Transposed elements have marked deltas and no intron
©virtualtext www.ergito.com

Figure 17.13 A unique Ty element, engineered to contain an intron, transposes to give copies that lack the intron. The copies possess identical terminal repeats, generated from one of the termini of the original Ty element.

We know of only one way to remove introns: RNA splicing. This suggests that transposition occurs by the same mechanism as with retroviruses. The *Ty* element is transcribed into an RNA that is recognized by the splicing apparatus. The spliced RNA is recognized by a reverse transcriptase and regenerates a duplex DNA copy.

The analogy with retroviruses extends further. The original *Ty* element has a difference in sequence between its two *delta* elements. But the transposed elements possess identical *delta* sequences, derived from the 5 ′ delta of the original element. If we consider the *delta* sequence to be exactly like an LTR, consisting of the regions U3-R-U5, the *Ty* RNA extends from R region to R region. Just as shown for retroviruses in **Figure 17.3**, **Figure 17.4**, **Figure 17.5**, **Figure 17.6**, the complete LTR is regenerated by adding a U5 to the 3 ′ end and a U3 to the 5 ′ end.

Transposition is controlled by genes within the *Ty* element. The *GAL* promoter used to control transcription of the marked *Ty* element is inducible: it is turned on by the addition of galactose. Induction of the promoter has two effects. It is necessary to activate transposition of the marked element. And its activation also increases the frequency of transposition of the other *Ty* elements on the yeast chromosome. This implies that the products of the *Ty* element can act in *trans* on other elements (actually on their RNAs).

Although the *Ty* element does not give rise to infectious particles, virus-like particles (VLPs) accumulate within the cells in which transposition has been induced. The particles can be seen in **Figure 17.14**. They contain full-length RNA, double-stranded DNA, reverse transcriptase activity, and a TyB product with integrase activity. The TyA product is cleaved like a *gag* precursor to produce the mature core proteins of the VLP. This takes the analogy between the *Ty* transposon and the retrovirus even further. The *Ty* element behaves in short like a retrovirus that has lost its *env* gene and therefore cannot properly package its genome.

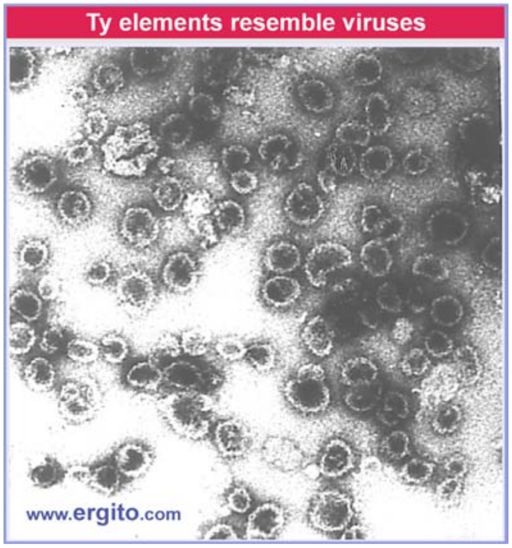


Figure 17.14 Ty elements generate virus-like particles. Photograph kindly provided by Alan Kingsman.

Only some of the *Ty* elements in any yeast genome are active: most have lost the ability to transpose (and are analogous to inert endogenous proviruses). Since these "dead" elements retain the δ repeats, however, they provide targets for transposition in response to the proteins synthesized by an active element.

References

575. Boeke, J. D. et al. (1985). *Ty elements transpose through an RNA intermediate*. Cell 40, 491-500.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.4.17.7

RETROVIRUSES AND RETROPOSONS

4.17.8 Many transposable elements reside in *D. melanogaster*

Key Concepts

- *copia* is a retroposon that is abundant in *D. melanogaster*.

The presence of transposable elements in *D. melanogaster* was first inferred from observations analogous to those that identified the first insertion sequences in *E. coli*. Unstable mutations are found that revert to wild type by deletion, or that generate deletions of the flanking material with an endpoint at the original site of mutation. They are caused by several types of transposable sequence, which are illustrated in **Figure 17.15**. They include the *copia* retroposon, the FB family, and the P elements discussed previously in *Molecular Biology 4.16.14 The role of transposable elements in hybrid dysgenesis*.

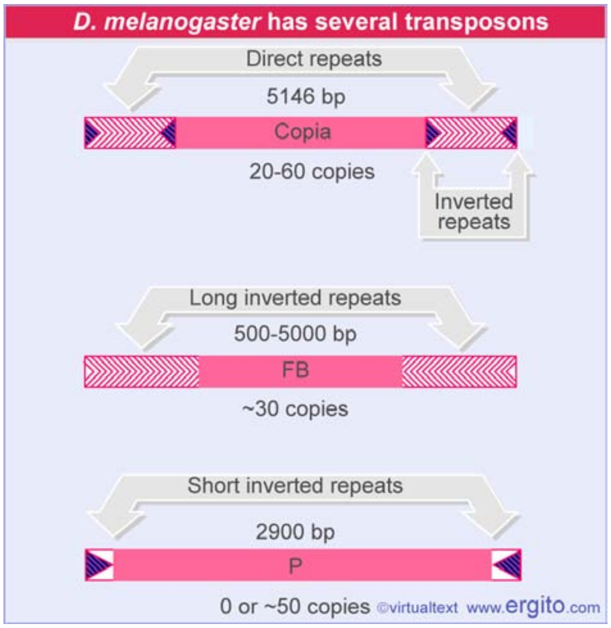


Figure 17.15 Three types of transposable element in *D. melanogaster* have different structures.

The best-characterized family of retroposons is *copia*. Its name reflects the presence of a large number of closely related sequences that code for abundant mRNAs. The *copia* family is taken as a paradigm for several other types of elements whose sequences are unrelated, but whose structure and general behavior appear to be similar.

The number of copies of the *copia* element depends on the strain of fly; usually it is 20-60. The members of the family are widely dispersed. The locations of *copia* elements show a different (although overlapping) spectrum in each strain of *D. melanogaster*.

These differences have developed over evolutionary periods. Comparisons of strains that have diverged recently (over the past 40 years or so) as the result of their propagation in the laboratory reveal few changes. We cannot estimate the rate of change, but the nature of the underlying events is indicated by the result of growing cells in culture. The number of *copia* elements per genome then increases substantially, up to 2-3 times. The additional elements represent insertions of *copia* sequences at new sites. Adaptation to culture in some unknown way transiently increases the rate of transposition to a range of 10^{-3} to 10^{-4} events per generation.

The *copia* element is ~5000 bp long, with identical direct terminal repeats of 276 bp. Each of the direct repeats itself ends in related inverted repeats. A direct repeat of 5 bp of target DNA is generated at the site of insertion. The divergence between individual members of the *copia* family is slight, <5%; variants often contain small deletions. All of these features are common to the other *copia*-like families, although their individual members display greater divergence (576).

The identity of the two direct repeats of each *copia* element implies either that they interact to permit correction events, or that both are generated from one of the direct repeats of a progenitor element during transposition. As in the similar case of *Ty* elements, this is suggestive of a relationship with retroviruses.

The *copia* elements in the genome are always intact; individual copies of the terminal repeats have not been detected (although we would expect them to be generated if recombination deleted the intervening material). *copia* elements sometimes are found in the form of free circular DNA; like retroviral DNA circles, the longer form has two terminal repeats and the shorter form has only one. Particles containing *copia* RNA have been noticed.

The *copia* sequence contains a single long reading frame of 4227 bp. There are homologies between parts of the *copia* open reading frame and the *gag* and *pol* sequences of retroviruses. A notable absence from the homologies is any relationship with retroviral *env* sequences required for the envelope of the virus, which means that *copia* is unlikely to be able to generate virus-like particles.

Transcripts of *copia* are found as abundant poly(A)$^{+}$ mRNAs, representing both full-length and part-length transcripts. The mRNAs have a common 5′ terminus, resulting from initiation in the middle of one of the terminal repeats. Several proteins are produced, probably involving events such as splicing of RNA and cleavage of polyproteins.

Although we lack direct evidence for *copia*'s mode of transposition, there are so many resemblances with retroviral organization that the conclusion seems inevitable that *copia* must have an origin related to the retroviruses. It is hard to say how many retroviral functions it possesses. We know, of course, that it transposes; but (as is the case with *Ty* elements) there is no evidence for any infectious capacity.

References

576. Mount, S. M. and Rubin, G. M. (1985). *Complete nucleotide sequence of the Drosophila transposable element copia: homology between copia and retroviral proteins*. Mol. Cell Biol. 5, 1630-1638.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.4.17.8

RETROVIRUSES AND RETROPOSONS

4.17.9 Retroposons fall into three classes

Key Terms

The **viral superfamily** comprises transposons that are related to retroviruses. They are defined by sequences that code for reverse transcriptase or integrase.

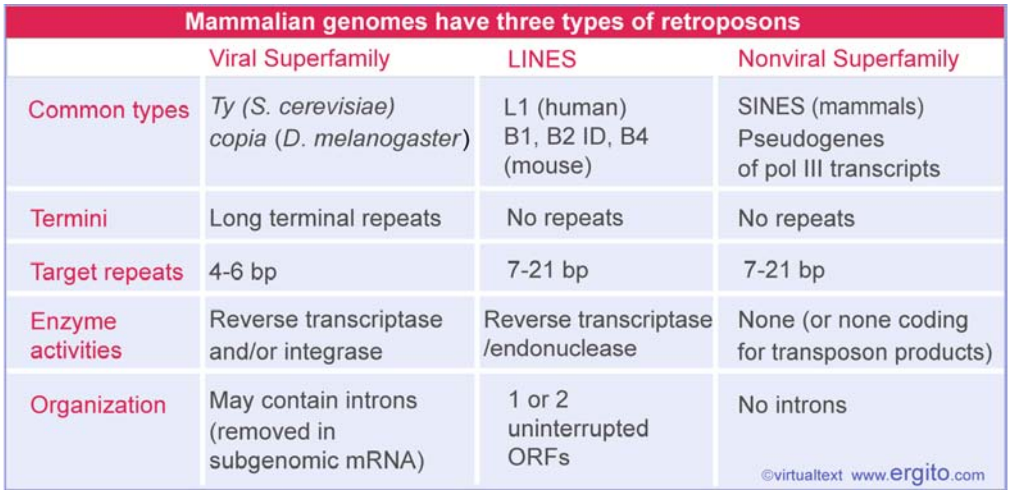
The **nonviral superfamily** of transposons originated independently of retroviruses.

Interspersed repeats were originally defined as short sequences that are common and widely distributed in the genome. They are now known to consist of transposable elements.

Key Concepts

- Retroposons of the viral superfamily are transposons that mobilize via an RNA that does not form an infectious particle.
- Some directly resemble retroviruses in their use of LTRs, but others do not have LTRs.
- Other elements can be found that were generated by an RNA-mediated transposition event, but do not themselves code for enzymes that can catalyze transposition.
- Transposons and retroposons constitute almost half of the human genome.

Retroposons are defined by their use of mechanisms for transposition that involve reverse transcription of RNA into DNA. Three classes of retroposons are distinguished in **Figure 17.16**:



Mammalian genomes have three types of retroposons

	Viral Superfamily	LINES	Nonviral Superfamily
Common types	*Ty (S. cerevisiae)* *copia* (*D. melanogaster*)	L1 (human) B1, B2 ID, B4 (mouse)	SINES (mammals) Pseudogenes of pol III transcripts
Termini	Long terminal repeats	No repeats	No repeats
Target repeats	4-6 bp	7-21 bp	7-21 bp
Enzyme activities	Reverse transcriptase and/or integrase	Reverse transcriptase /endonuclease	None (or none coding for transposon products)
Organization	May contain introns (removed in subgenomic mRNA)	1 or 2 uninterrupted ORFs	No introns

©virtualtext www.ergito.com

Figure 17.16 Retroposons can be divided into the viral superfamilies that are retrovirus-like or LINES and the nonviral superfamilies that do not have coding functions.

- Members of the **viral superfamily** code for reverse transcriptase and/or integrase activities. Like other retroposons, they reproduce like retroviruses but differ from them in not passing through an independent infectious form. They are best characterized in the *Ty* and *copia* elements of yeast and flies.

- The LINES also have reverse transcriptase activity (and may therefore be considered to comprise more distant members of the viral superfamily), but they lack LTRs and use a different mechanism from retroviruses to prime the reverse transcription reaction. They are derived from RNA polymerase II transcripts. A minority of the elements in the genome are fully functional and can transpose autonomously; others have mutations and so can only transpose as the result of the action of a *trans*-acting autonomous element.

- Members of the **nonviral superfamily** are identified by external and internal features that suggest that they originated in RNA sequences, although in these cases we can only speculate on how a DNA copy was generated (for review see 165). We assume that they were targets for a transposition event by an enzyme system coded elsewhere, that is, they are always nonautonomous. They originated in cellular transcripts. They do not code for proteins that have transposition functions. The most prominent component of this family is called SINES. They are derived from RNA polymerase III transcripts (for review see 166).

Figure 17.17 shows the organization and sequence relationships of elements that code for reverse transcriptase. Like retroviruses, the LTR-containing retroposons can be classified into groups according to the number of independent reading frames for *gag*, *pol*, and *int*, and the order of the genes. In spite of these superficial differences of organization, the common feature is the presence of reverse transcriptase and integrase activities. Typical mammalian LINES elements have two reading frames, one coding for a nucleic acid-binding protein, the other for reverse transcriptase and endonuclease activity.

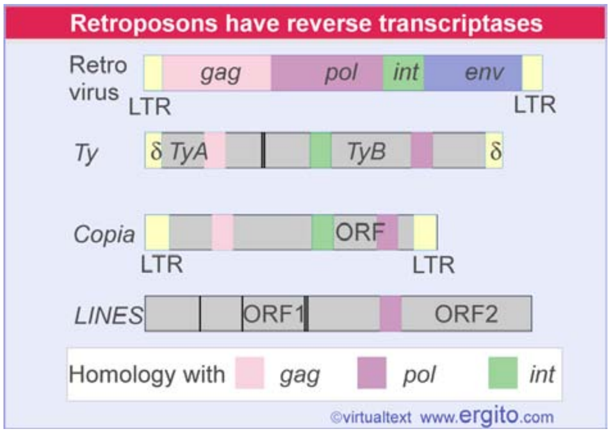


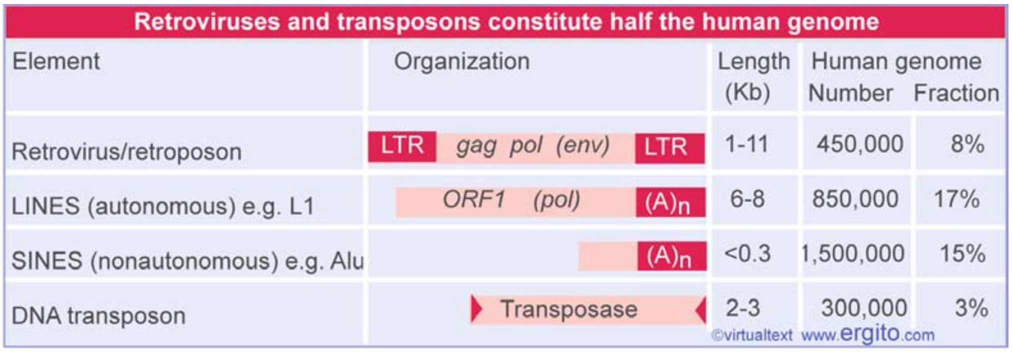
Figure 17.17 Retroposons that are closely related to retroviruses have a similar organization, but LINES share only the reverse transcriptase activity.

LTR-containing elements can vary from integrated retroviruses to retroposons that have lost the capacity to generate infectious particles. Yeast and fly genomes have the *Ty* and *copia* elements that cannot generate infectious particles. Mammalian genomes have endogenous retroviruses that, when active, can generate infectious particles. The mouse genome has several active endogenous retroviruses which are able to generate particles that propagate horizontal infections. By contrast, almost all endogenous retroviruses lost their activity some 50 million years ago in the human lineage, and the genome now has mostly inactive remnants of the endogenous retroviruses.

LINES and SINES comprise a major part of the animal genome. They were defined originally by the existence of a large number of relatively short sequences that are related to one another (comprising the moderately repetitive DNA described in *Molecular Biology 1.3.6 Eukaryotic genomes contain both nonrepetitive and repetitive DNA sequences*). The LINES comprise long interspersed sequences, and the SINES comprise short interspersed sequences. (They are described as interspersed sequences or **interspersed repeats** because of their common occurrence and widespread distribution.)

Plants contain another type of small mobile element, called MITE (for miniature inverted-repeat transposable element). Such elements terminate in inverted repeats, have a 2 or 3 bp target sequence, do not have coding sequences, and are 200-500 bp long. At least 9 such families exist in (for example) the rice genome. They are often found in the regions flanking protein-coding genes (2838; 2837). They have no relationship to SINES or LINES.

LINES and SINES comprise a significant part of the repetitive DNA of animal genomes. In many higher eukaryotic genomes, they occupy ~50% of the total DNA. **Figure 17.18** summarizes the distribution of the different types of transposons that constitute almost half of the human genome (1442). Except for the SINES, which are always nonfunctional, the other types of elements all consist of both functional elements and elements that have suffered deletions eliminating parts of the reading frames that code for the protein(s) needed for transposition. The relative proportions of these types of transposons are generally similar in the mouse genome (3203).



Retroviruses and transposons constitute half the human genome

Element	Organization	Length (Kb)	Human genome Number	Human genome Fraction
Retrovirus/retroposon	LTR *gag pol (env)* LTR	1-11	450,000	8%
LINES (autonomous) e.g. L1	*ORF1 (pol)* (A)n	6-8	850,000	17%
SINES (nonautonomous) e.g. Alu	(A)n	<0.3	1,500,000	15%
DNA transposon	▶ Transposase ◀	2-3	300,000	3%

©virtualtext www.ergito.com

Figure 17.18 Four types of transposable elements constitute almost half of the human genome.

A common LINES in mammalian genomes is called L1 (for review see 2296). The typical member is ~6,500 bp long, terminating in an A-rich tract. The two open reading frames of a full-length element are called ORF1 and ORF2. The number of full-length elements is usually small (~50), and the remainder of the copies are truncated. Transcripts can be found. As implied by its presence in repetitive DNA, the LINES family shows sequence variation among individual members. However, the members of the family within a species are relatively homogeneous compared to the variation shown between species (577; for review see 167). L1 is the only member of the LINES family that has been active in either the mouse or human lineages, and it seems to have remained highly active in the mouse, but has declined in the human lineage.

Only one SINES has been active in the human lineage; this is the common Alu element. The mouse genome has a counterpart to this element (B1), and also other SINES (B2, ID, B4) that have been active. Human Alu and mouse B1 SINES are probably derived from the 7SL RNA (see *Molecular Biology 4.17.10 The Alu family has many widely dispersed members*). The other mouse SINES appear to have originated from reverse transcripts of tRNAs. The transposition of the SINES probably results from their recognition as substrates by an active L1 element.

Last updated on 12-25-2002

Reviews

165. Weiner, A. M., Deininger, P. L., and Efstratiadis, A. (1986). *Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information*. Annu. Rev. Biochem. 55, 631-661.

166. Deininger, P. L. (1989). *SINEs: short interspersed repeated DNA elements in higher eukaryotes*. Mobile DNA, 19-636.

167. Hutchison, C. A. et al. (1989). *LINEs and related retroposons: long interspersed repeated sequences in the eukaryotic genome*. Mobile DNA, 93-617.

2296. Ostertag, E. M. and Kazazian, H. H. (2001). *Biology of mammalian L1 retrotransposons.* Annu. Rev. Genet. 35, 501-538.

References

577. Loeb, D. D. et al. (1986). *The sequence of a large L1Md element reveals a tandemly repeated 5′ end and several features found in retrotransposons*. Mol. Cell Biol. 6, 168-182.

1442. Sachidanandam, R. et al. (2001). *A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. The International SNP Map Working Group.* Nature 409, 928-933.

2837. Bureau, T. E., Ronald, P. C., and Wessler, S. R. (1996). *A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes.* Proc. Natl. Acad. Sci. USA 93, 8524-8529.

2838. Bureau, T. E. and Wessler, S. R. (1992). *Tourist: a large family of small inverted repeat elements frequently associated with maize genes.* Plant Cell 4, 1283-1294.

3203. Waterston et al. (2002). *Initial sequencing and comparative analysis of the mouse genome.* Nature 420, 520-562.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.4.17.9

RETROVIRUSES AND RETROPOSONS

4.17.10 The Alu family has many widely dispersed members

Key Terms

The **Alu family** is a set of dispersed, related sequences, each ~300 bp long, in the human genome. The individual members have Alu cleavage sites at each end (hence the name).

Key Concepts

- A major part of repetitive DNA in mammalian genomes consists of repeats of a single family organized like transposons and derived from RNA polymerase III transcripts.

The most prominent SINES comprises members of a single family. Its short length and high degree of repetition make it comparable to simple sequence (satellite) DNA, except that the individual members of the family are dispersed around the genome instead of being confined to tandem clusters. Again there is significant similarity between the members within a species compared with variation between species.

In the human genome, a large part of the moderately repetitive DNA exists as sequences of ~300 bp that are interspersed with nonrepetitive DNA. At least half of the renatured duplex material is cleaved by the restriction enzyme AluI at a single site, located 170 bp along the sequence. The cleaved sequences all are members of a single family, known as the **Alu family** after the means of its identification. There are ~300,000 members in the haploid genome (equivalent to one member per 6 kb of DNA). The individual Alu sequences are widely dispersed. A related sequence family is present in the mouse (where the 50,000 members are called the B1 family), in the Chinese hamster (where it is called the Alu-equivalent family), and in other mammals.

The individual members of the Alu family are related rather than identical. The human family seems to have originated by a 130 bp tandem duplication, with an unrelated sequence of 31 bp inserted in the right half of the dimer. The two repeats are sometimes called the "left half" and "right half" of the Alu sequence. The individual members of the Alu family have an average identity with the consensus sequence of 87%. The mouse B1 repeating unit is 130 bp long, corresponding to a monomer of the human unit. It has 70-80% homology with the human sequence.

The Alu sequence is related to 7SL RNA, a component of the signal recognition particle (see *Molecular Biology 2.8.10 The SRP interacts with the SRP receptor*). The 7SL RNA corresponds to the left half of an Alu sequence with an insertion in the middle. So the 90 5′ terminal bases of 7SL RNA are homologous to the left end of Alu, the central 160 bases of 7SL RNA have no homology to Alu, and the 40 3′ terminal bases of 7SL RNA are homologous to the right end of Alu. The 7SL RNA is coded by genes that are actively transcribed by RNA polymerase III. It is possible that these genes (or genes related to them) gave rise to the inactive Alu sequences.

The members of the Alu family resemble transposons in being flanked by short direct repeats. However, they display the curious feature that the lengths of the repeats are different for individual members of the family. Because they derive from RNA polymerase III transcripts, it is possible that individual members carry internal active promoters.

A variety of properties have been found for the Alu family, and its ubiquity has prompted many suggestions for its function, but it is not yet possible to discern its true role.

At least some members of the family can be transcribed into independent RNAs. In the Chinese hamster, some (although not all) members of the Alu-equivalent family appear to be transcribed *in vivo*. Transcription units of this sort are found in the vicinity of other transcription units.

Members of the Alu family may be included within structural gene transcription units, as seen by their presence in long nuclear RNA. The presence of multiple copies of the Alu sequence in a single nuclear molecule can generate secondary structure. In fact, the presence of Alu family members in the form of inverted repeats is responsible for most of the secondary structure found in mammalian nuclear RNA.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.4.17.10

RETROVIRUSES AND RETROPOSONS

4.17.11 Processed pseudogenes originated as substrates for transposition

Key Terms

A **processed pseudogene** is an inactive gene copy that lacks introns, contrasted with the interrupted structure of the active gene. Such genes originate by reverse transcription of mRNA and insertion of a duplex copy into the genome.

Key Concepts

- A processed pseudogene is derived from an mRNA sequence by reverse transcription.

When a sequence generated by reverse transcription of an mRNA is inserted into the genome, we can recognize its relationship to the gene from which the mRNA was transcribed. Such a sequence is called a **processed pseudogene** to reflect the fact that it was processed from RNA and is not active. The characteristic features of a processed pseudogene are compared in **Figure 17.19** with the features of the original gene and the mRNA. The figure shows all the relevant diagnostic features, only some of which are found in any individual example. Any transcript of RNA polymerase II could in principle give rise to such a pseudogene, and there are many examples, including the processed globin pseudogenes that were the first to be discovered (see *Molecular Biology 1.4.6 Pseudogenes are dead ends of evolution*).

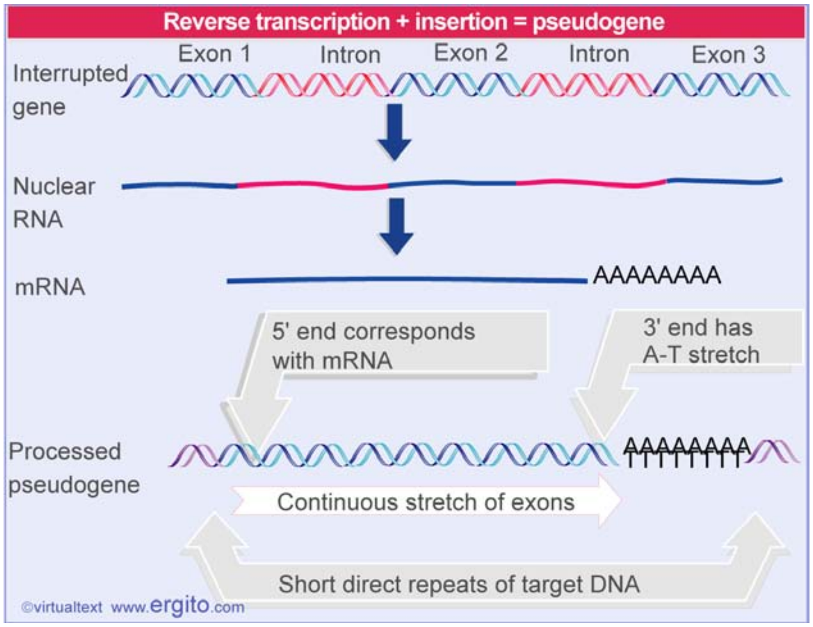

Reverse transcription + insertion = pseudogene
Exon 1
Intron
Exon 2
Intron
Exon 3
Interrupted gene
Nuclear RNA
mRNA
AAAAAAAA
5' end corresponds with mRNA
3' end has A-T stretch
Processed pseudogene
AAAAAAAA
Continuous stretch of exons
Short direct repeats of target DNA
©virtualtext www.ergito.com

Figure 17.19 Pseudogenes could arise by reverse transcription of RNA to give duplex DNAs that become integrated into the genome.

The pseudogene may start at the point equivalent to the 5 ′ terminus of the RNA, which would be expected only if the DNA had originated from the RNA. Several pseudogenes consist of precisely joined exon sequences; we know of no mechanism to recognize introns in DNA, so this feature argues for an RNA-mediated stage. The pseudogene may end in a short stretch of A·T base pairs, presumably derived from the poly(A) tail of the RNA. On either side of the pseudogene is a short direct repeat, presumed to have been generated by a transposition-like event. Processed pseudogenes reside at locations unrelated to their presumed sites of origin.

The processed pseudogenes do not carry any information that might be used to sponsor a transposition event (or to carry out the preceding reverse transcription of the RNA). This suggests that the RNA was a substrate for another system, coded by a retroposon. In fact, it seems likely that the active LINES elements provide most of the reverse transcriptase activity, and they are responsible not only for their own transposition, but also for acting on the SINES and for generating processed pseudogenes.

Last updated on 12-20-2002

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.4.17.11

RETROVIRUSES AND RETROPOSONS

4.17.12 LINES use an endonuclease to generate a priming end

Key Terms

A **processed pseudogene** is an inactive gene copy that lacks introns, contrasted with the interrupted structure of the active gene. Such genes originate by reverse transcription of mRNA and insertion of a duplex copy into the genome.

Key Concepts

- LINES do not have LTRs and require the retroposon to code for an endonuclease that generates a nick to prime reverse transcription.

LINES elements, and some others, do not terminate in the LTRs that are typical of retroviral elements. This poses the question: how is reverse transcription primed? It does not involve the typical reaction in which a tRNA primer pairs with the LTR (see **Figure 17.6**). The open reading frames in these elements lack many of the retroviral functions, such as protease or integrase domains, but typically have reverse transcriptase-like sequences and code for an endonuclease activity. In the human LINES L1, ORF1 is a DNA-binding protein and ORF2 has both reverse transcriptase and endonuclease activities; both products are required for transposition (3152; 3153).

Figure 17.20 shows how these activities support transposition. A nick is made in the DNA target site by an endonuclease activity coded by the retroposon. The RNA product of the element associates with the protein bound at the nick. The nick provides a 3 ′ –OH end that primes synthesis of cDNA on the RNA template. A second cleavage event is required to open the other strand of DNA, and the RNA/DNA hybrid is linked to the other end of the gap either at this stage or after it has been converted into a DNA duplex. A similar mechanism is used by some mobile introns (see **Figure 26.11**) (580; 581).

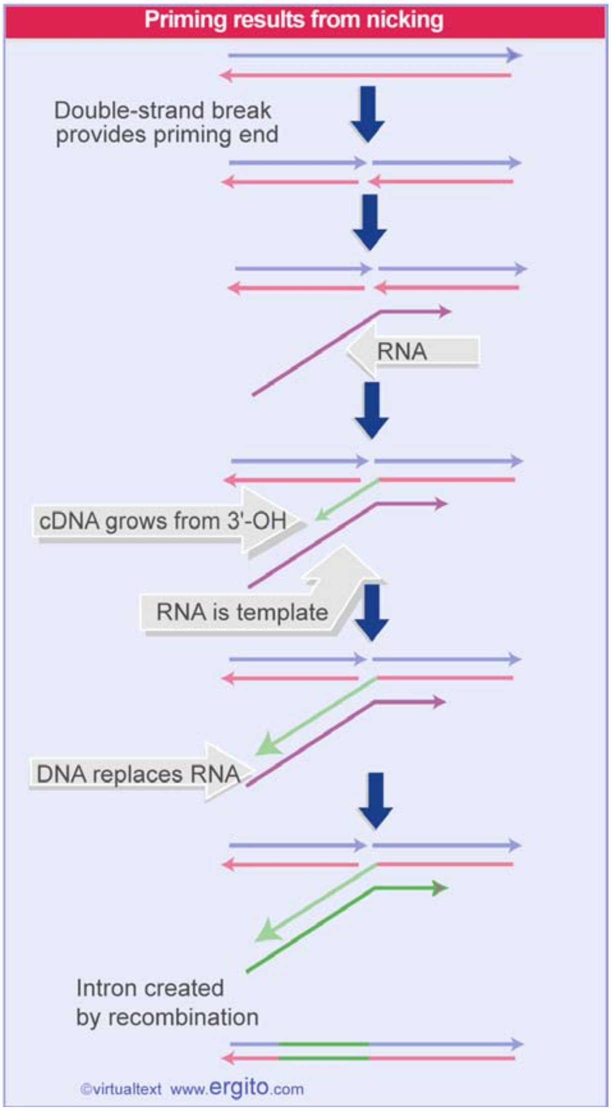

Priming results from nicking
Double-strand break provides priming end
RNA
cDNA grows from 3'-OH
RNA is template
DNA replaces RNA
Intron created by recombination
©virtualtext www.ergito.com

Figure 17.20 Retrotransposition of non-LTR elements occurs by nicking the target to provide a primer for cDNA synthesis on an RNA template. The arrowheads indicate 3 ′ ends.

When elements originate from RNA polymerase II transcripts, the genomic sequences are necessarily inactive: they lack the promoter that was upstream of the original startpoint for transcription. Because they usually possess the features of the mature transcript, they are called **processed pseudogenes**.

One of the reasons why LINES elements are so effective lies with their method of propagation. When a LINES mRNA is translated, the protein products show a *cis*-preference for binding to the mRNA from which they were translated. **Figure 17.21** shows that the ribonucleoprotein complex then moves to the nucleus, where the proteins insert a DNA copy into the genome. Often reverse transcription does not proceed fully to the end, so the copy is inactive. However, there is the potential for insertion of an active copy, because the proteins are acting on a transcript of the original active element.

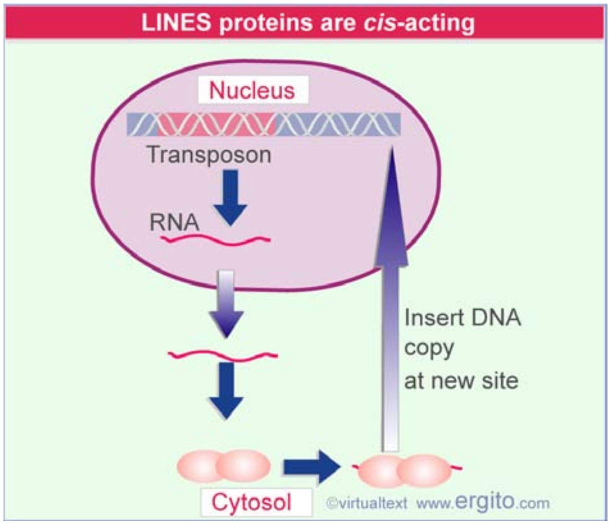

LINES proteins are *cis*-acting
Nucleus
Transposon
RNA
Insert DNA copy at new site
Cytosol
©virtualtext www.ergito.com

Figure 17.21 A LINES is transcribed into an RNA that is translated into proteins that assemble into a complex with the RNA. The complex translocates to the nucleus, where it inserts a DNA copy into the genome.

By contrast, the proteins produced by the DNA transposons must be imported into the nucleus after being synthesized in the cytoplasm, but they have no means of distinguishing full-length transposons from inactive deleted transposons. **Figure 17.22** shows that instead, they will indiscriminately recognize any element by virtue of the repeats that mark the ends, much reducing their chance of acting on a full-length as opposed to deleted element. The consequence is that inactive elements accumulate and eventually the family dies out because a transposase has such a small chance of finding a target that is a fully functional transposon.

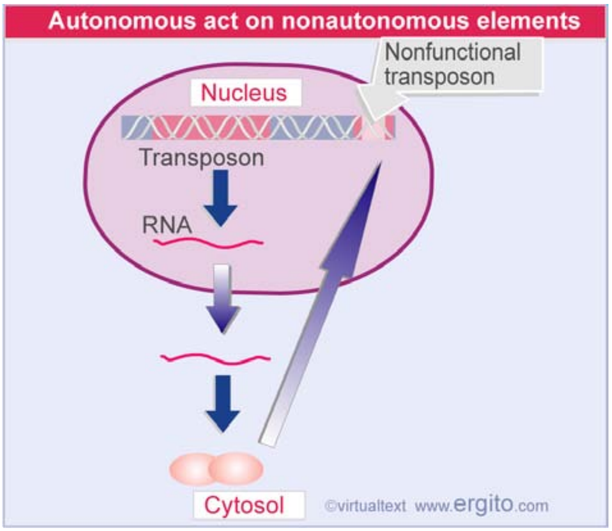

Autonomous act on nonautonomous elements
Nonfunctional transposon
Nucleus
Transposon
RNA
Cytosol
©virtualtext www.ergito.com

Figure 17.22 A transposon is transcribed into an RNA that is translated into proteins that move independently to the nucleus, where they act on any pair of inverted repeats with the same sequence as the original transposon.

Are transposition events currently occurring in these genomes or are we seeing only the footprints of ancient systems? This varies with the species. There are few currently active transposons in the human genome, but by contrast several active transposons are known in the mouse genome. This explains the fact that spontaneous mutations caused by LINES insertions occur at a rate of ~3% in mouse, but only 0.1% in man. There appear to be ~10-50 active LINES elements in the human genome. Some human diseases can be pinpointed as the result of transposition of L1 into genes, and others result from unequal crossing-over events involving repeated copies of L1 (for review see 2296). A model system in which LINES transposition occurs in tissue culture cells suggests that a transposition event can introduce several types of collateral damage as well as inserting into a new site; the damage includes chromosomal rearrangements and deletions (3150; 3151). Such events may be viewed as agents of genetic change. Neither DNA transposons nor retroviral-like retroposons seem to have been active in the human genome for 40-50 million years, but several active examples of both are found in the mouse.

Note that for transpositions to survive, they must occur in the germline. Presumably similar events occur in somatic cells, but do not survive beyond one generation.

Last updated on 8-8-2002

Reviews

2296. Ostertag, E. M. and Kazazian, H. H. (2001). *Biology of mammalian L1 retrotransposons.* Annu. Rev. Genet. 35, 501-538.

References

580. Luan, D. D. et al. (1993). *Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition*. Cell 72, 595-605.

581. Lauermann, V. and Boeke, J. D. (1994). *The primer tRNA sequence is not inherited during Ty1 retrotransposition*. Proc. Natl. Acad. Sci. USA 91, 9847-9851.

3150. Gilbert, N., Lutz-Prigge, S., and Moran, J. V. (2002). *Genomic deletions created upon LINE-1 retrotransposition.* Cell 110, 315-325.

3151. Symer, D. E., Connelly, C., Szak, S. T., Caputo, E. M., Cost, G. J., Parmigiani, G., and Boeke, J. D. (2002). *Human l1 retrotransposition is associated with genetic instability in vitro.* Cell 110, 327-338.

3152. Feng, Q., Moran, J. V., Kazazian, H. H., and Boeke, J. D. (1996). *Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition.* Cell 87, 905-916.

3153. Moran, J. V., Holmes, S. E., Naas, T. P., DeBerardinis, R. J., Boeke, J. D., and Kazazian, H. H. (1996). *High frequency retrotransposition in cultured mammalian cells.* Cell 87, 917-927.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.4.17.12

RETROVIRUSES AND RETROPOSONS

4.17.13 Summary

Reverse transcription is the unifying mechanism for reproduction of retroviruses and perpetuation of retroposons. The cycle of each type of element is in principle similar, although retroviruses are usually regarded from the perspective of the free viral (RNA) form, while retroposons are regarded from the stance of the genomic (duplex DNA) form.

Retroviruses have genomes of single-stranded RNA that are replicated through a double-stranded DNA intermediate. An individual retrovirus contains two copies of its genome. The genome contains the *gag*, *pol*, and *env* genes, which are translated into polyproteins, each of which is cleaved into smaller functional proteins. The Gag and Env components are concerned with packing RNA and generating the virion; the Pol components are concerned with nucleic acid synthesis.

Reverse transcriptase is the major component of Pol, and is responsible for synthesizing a DNA (minus strand) copy of the viral (plus strand) RNA. The DNA product is longer than the RNA template; by switching template strands, reverse transcriptase copies the 3′ sequence of the RNA to the 5′ end of the DNA, and copies the 5′ sequence of the RNA to the 3′ end of the DNA. This generates the characteristic LTRs (long terminal repeats) of the DNA. A similar switch of templates occurs when the plus strand of DNA is synthesized using the minus strand as template. Linear duplex DNA is inserted into a host genome by the integrase enzyme. Transcription of the integrated DNA from a promoter in the left LTR generates further copies of the RNA sequence.

Switches in template during nucleic acid synthesis allow recombination to occur by copy choice. During an infective cycle, a retrovirus may exchange part of its usual sequence for a cellular sequence; the resulting virus is usually replication-defective, but can be perpetuated in the course of a joint infection with a helper virus. Many of the defective viruses have gained an RNA version (*v-onc*) of a cellular gene (*c-onc*). The *onc* sequence may be any one of a number of genes whose expression in *v-onc* form causes the cell to be transformed into a tumorigenic phenotype.

The integration event generates direct target repeats (like transposons that mobilize via DNA). An inserted provirus therefore has direct terminal repeats of the LTRs, flanked by short repeats of target DNA. Mammalian and avian genomes have endogenous (inactive) proviruses with such structures. Other elements with this organization have been found in a variety of genomes, most notably in *S. cerevisiae* and *D. melanogaster*. *Ty* elements of yeast and *copia* elements of flies have coding sequences with homology to reverse transcriptase, and mobilize via an RNA form. They may generate particles resembling viruses, but do not have infectious capability. The LINES sequences of mammalian genomes are further removed from the retroviruses, but retain enough similarities to suggest a common origin. They use a different type of priming event to initiate reverse transcription, in which an endonuclease activity associated with the reverse transcriptase makes a nick that provides a 3′–OH end for priming synthesis on an RNA template. The frequency of LINES transposition is increased because its protein products are *cis*-acting; they associate with the mRNA from which they were translated to form a ribonucleoprotein complex that is transported into the nucleus.

Another class of retroposons have the hallmarks of transposition via RNA, but have no coding sequences (or at least none resembling retroviral functions). They may have originated as passengers in a retroviral-like transposition event, in which an RNA was a target for a reverse transcriptase. Processed pseudogenes arise by such events. A particularly prominent family apparently originating from a processing event is the mammalian SINES, including the human Alu family. Some snRNAs, including 7SL snRNA (a component of the SRP) are related to this family.

This content is available online at http://www.ergito.com/main.jsp?bcs=MBIO.4.17.13