Review

Reverse Transcription in the Eukaryotic Genome: Retroviruses, Pararetroviruses, Retrotransposons, and Retrotranscripts[1]

Howard M. Temin
University of Wisconsin–Madison

Recent studies indicate that >10% of the human and mouse genome appears to consist of integrated DNA copies of RNA molecules. These sequences include retroviruses, retrovirus-like DNAs, retrotransposons, and retrotranscripts and represent more than 500,000 separate integration events. The nature of the enzymes used for the reverse transcription from RNA to DNA and for integration of the DNA copies into chromosomal DNA is unknown. A major evolutionary effect of these integrations would have been mutation. Thus, present-day organisms are those that survived this mutational load.

Introduction

Reverse transcription was first discovered as a key step in the replication of certain animal viruses now called retroviruses. Although it was suggested at that time that reverse transcription is an important general biological process (Temin 1970, 1971, 1974), evidence supporting this hypothesis has only recently been obtained (Temin 1980; Temin and Engels 1984). Somewhat surprisingly, this evidence does not come from discoveries of reverse transcriptase in normal cells but from analysis of the structure of various DNAs isolated by DNA cloning. The results of reverse transcription and integration make up >10% of the human and mouse genomes, representing over 500,000 separate insertions of DNA into the chromosomes (table 1).

Reverse transcription has now been found in retroviruses and pararetroviruses (hepatitis B–like viruses). The vertebrate genome contains DNA copies of retrovirus RNA as well as (1) DNAs with organizations similar to those of retrovirus DNA (retrotransposons) and (2) DNAs with some characteristics indicating that they have been transcribed from an RNA template (retrotranscripts).

Retrovirus Life Cycle

To understand this conclusion about the important evolutionary role of reverse transcription, we must examine some details of the retrovirus life cycle (Weiss et al. 1982). At one level of description, retroviruses look and behave like other animal viruses. That is, they have a medium-sized enveloped virion, and their life cycle can be divided into virus attachment and entrance into cells, an eclipse or latent period, and a period of virus formation and release. At another level of description, the structure of the retrovirus replicative intermediate, the provirus, is like that of many cellular

1. Key words: retrovirus, reverse transcription, transposons, integration. Abbreviations: LTR = long terminal repeat, PBS = primer binding site, PPT = polypurine track, E = encapsidation sequence, R = repeat, IAP = intracisternal A-particle.

Address for correspondence and reprints: Dr. Howard M. Temin, McArdle Laboratory for Cancer Research, University of Wisconsin—Madison, Madison, Wisconsin 53706.

Mol. Biol. Evol. 2(6):455–468. 1985.

0737-4038/85/0206-0466$02.00

Table 1
Number of Copies of Reverse Transcribed DNAs per Genome

Type of Sequence	Number
Endogenous retrovirus and retrovirus-like	1,500
Retrotranscripts:	
cDNA	>1,000[a]
Small	500,000
Large	30,000

NOTE.—Data are approximate number of copies per genome of these different sequences in the mouse. Definitions and references are cited in the text.

[a] Extrapolation from present findings. The number probably is >10,000.

movable genetic elements (fig. 1) (Temin 1981, 1982; Varmus 1982); that is, the retrovirus provirus has terminal inverted repeats embedded in terminal direct repeats (two inverted repeats per direct repeat) and is integrated in many different sites in the cell genome in a small direct repeat of the cell DNA. (The direct repeats in viral DNA are called long terminal repeats or LTRs.) Thus, the viral sequences at the ends of the provirus are always constant, and the cell sequences next to the ends are different. The exact nucleotides at the ends of the provirus, TG, . . . CA (and frequently some other adjoining nucleotides) are also present in many eukaryotic cellular movable genetic elements.

The actual process of transfer of information from RNA to integrated DNA by reverse transcription in retrovirus replication is quite complex. The synthesis of the unintegrated viral DNA intermediate involves two virus-encoded primers, PBS (primer binding site) and PPT (polypurine track), as well as a small terminal direct repeat in the viral RNA (r) (fig. 1). There are also two movements of newly synthesized DNA from one template position to another. In addition, a multifunctional viral protein, called reverse transcriptase, carries out reverse transcription and DNA-directed DNA synthesis and has ribonuclease H activity. (The primer that binds to PBS is actually a cellular tRNA molecule and is not coded by the virus at all.)

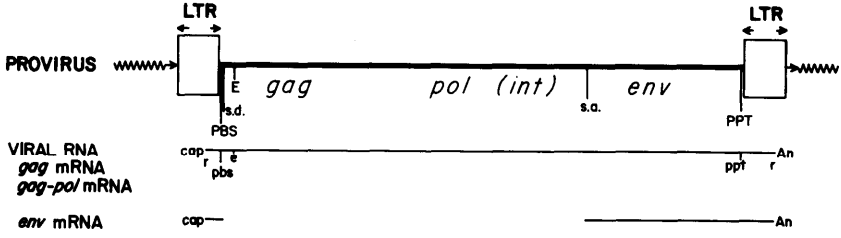

FIG. 1.—Integrated retroviral DNA and retroviral RNAs. The zigzag line represents cellular DNA; the zigzag arrow represents the direct repeat of cellular DNA formed on integration; the open boxes represent terminal direct repeats (LTR = long terminal repeat); the short arrows above the LTRs represent the inverted repeats; PBS = primary binding site; s.d. = splice donor; E = encapsidation sequence; *gag, pol,* and *env* are three genes coding for viral internal proteins, polymerase and other enzymes, and viral envelope proteins; *int* = the integrase portion of the *pol* gene; s.a. = splice acceptor; PPT = polypurine track (primer for second-strand DNA synthesis). In viral RNA, r = repeated sequence; pbs = primer binding site; e = encapsidation sequence; ppt = polypurine track; An = poly (A) track. Retrotransposons have similar LTRs but usually no s.d., s.a., or *env* sequences.

Integration of viral DNA involves several steps: the ends of linear, unintegrated viral DNA are blunt-end ligated to each other, forming a closed circle containing the approximately 20-bp viral attachment sequence required for virus integration at the junction of the ends of the linear unintegrated DNA; the closed circle is opened near the center of the attachment sequence; the central base pairs, TTAA, are lost; and a few nucleotides of cellular DNA are duplicated. One of the proteins involved in the integration, an endonuclease, is coded by the viral *int* gene, located at the 3′ end of the *pol* gene (fig. 1) (Panganiban and Temin 1984). The origin of the ligase(s), DNA polymerase, and endonuclease activities required in integration is unknown, although the lengths of the direct repeats of cellular DNA are virus specific.

Transcription of viral DNA followed by processing to give viral genomic RNA and mRNAs is performed entirely by cellular proteins acting on viral sequences. The viral control sequences include promoter, enhancer, poly (A) addition, possible terminator, and splice-donor and splice-acceptor sequences (figs. 1 and 2). In addition, there are specific viral sequences, termed E, for encapsidation or packaging of the viral genomic RNA (fig. 1). (It is relevant to the discussion below that the viral poly (A) is removed during reverse transcription.)

Viral proteins are synthesized both from full-length and from spliced mRNAs (fig. 1). In the case of C-type retroviruses, the mRNA containing the coding sequences for reverse transcriptase is translated as a read-through protein with the product of the 5′ proximal gene, *gag* (see discussion of Tyl element below).

Retrovirus Vectors

Most of the *cis*-acting sequences of retroviruses are at the ends. Thus, deletion of internal protein-coding sequences forms defective retroviruses that can still replicate in the presence of a source of viral proteins (Temin 1985). Such replication-defective viruses occur naturally (e.g., most highly oncogenic retroviruses) or can be artificially constructed (e.g., most retrovirus vectors). These vectors can consist of only the viral LTRs, PBS, PPT, and encapsidation sequences, as well as a selectable marker. Study of the replication of these vectors shows that untranslated intervening sequences, present in cellular coding genes included in the vector, are removed during virus replication as long as the viral encapsidation sequences are preserved (Temin 1985).

Such vectors can only be passaged as infectious virus in the presence of either a helper virus or a cell expressing viral proteins, i.e., a helper cell. Thus, retrovirus vectors are relevant to consideration of other passively transferred reverse transcripts (see below).

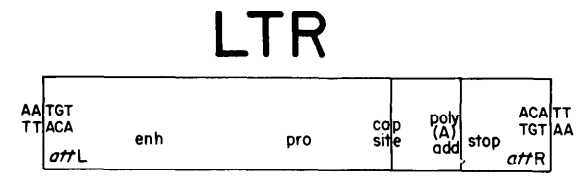

FIG. 2.—Retroviral long terminal repeat. AATGT and ACATT = the sequence of the inverted repeat in an avian retrovirus (the AA and TT are lost on integration); attL and attR = the two portions of the attachment sequence formed when they are ligated together; enh = enhancer; pro = promoter; cap site = the position of the 5′ end of viral RNA; poly (A) add = AATAA and the signal for poly (A) addition; stop = the signal for the 3′ end of transcribed viral RNA. The LTR of retrotransposons is similar, but the recognition site for transposition is not yet defined.

Endogenous Retroviruses

As was predicted from their life cycle, some retroviruses have infected the germ line and become permanent residents of the animal genome. These endogenous retroviruses are passed on to progeny organisms by Mendelian mechanisms. This origin of endogenous retroviruses from exogenous retroviruses is most clearly apparent for some mouse ecotropic retroviruses that have been recently acquired as germ-line genes in the history of some inbred mouse strains (Buckler et al. 1982). It also is suggested by the existence of some chickens with no endogenous avian leukosis viruses—whereas most chickens have several endogenous proviruses in their chromosomes—and of some mice with no mouse mammary-tumor viruses—whereas most mice have several in their chromosomes. A laboratory model for infection of the mouse germ line has been developed by Jaenisch et al. (1981).

It is important to note that such germ-line infections can be mutagenic (Harbers et al. 1984; Hutchison et al. 1984). Both spontaneous and laboratory-produced mutations are the result of retrovirus integration. The frequency of such mutations appears to be in the range of what would be expected from insertions of the provirus equally at all sites in the genome.

Pararetroviruses or Hepadnaviruses

Similarity to retroviruses is seen in pararetroviruses or hepadnaviruses, which all appear to use reverse transcription in replication but do not have an integrated replicative form (Summers and Mason 1982). These viruses include the various hepatitis B viruses and cauliflower mosaic virus (Hull and Covey 1983; Marco and Howell 1984; Miller et al. 1984; Volvitch et al. 1984). (Integrated copies of hepatitis B–virus DNA are present in some hepatocellular carcinomas, but these are defective forms, not replicative intermediates [Dejean et al. 1984]; their role in tumor formation is unclear.) Lentiviruses, a subfamily of retroviruses, may also replicate without integration (Harris et al. 1984).

Thus, we can see two possible paths of evolution from retroviruses—(1) toward more independence from the cell genome in the pararetroviruses, which no longer integrate, and (2) toward more dependence on the cell genome in the endogenous retroviruses, which usually no longer have a virion phase. That these paths represent actual evolution is supported by the nucleic-acid sequence homology between the retrovirus *pol* gene and the putative polymerase gene of the pararetroviruses and by the close sequence similarity between endogenous and exogenous retroviruses (Toh et al. 1983; Mandart et al. 1984).

Laboratory models of retrovirus replication without integration have been made with attachment site–negative or integrase-negative retroviruses and may provide a model for this evolution (Panganiban and Temin 1984). In addition, a retrovirus vector with a polyoma origin of replication may provide a model for later steps in the origin of such pararetroviruses (Berger and Bernstein 1985).

Endogenous Sequences Similar to Replication-Competent Retroviruses

There are other sequences present in the cell genome related to retroviruses by either nucleic-acid sequence homology or organization that do not have an exogenous-virus counterpart.

IAP Genes

The intracisternal A-particle (IAP) genes in mice and Syrian hamsters form a large (1,000 copies per genome) family of proviruses that code for a noninfectious A-type particle with a reverse transcriptase. These genes have been observed to transpose when forming new proviruses in lymphoid cells, although this transposition may represent somewhat aberrant retrovirus-like integration (Hawley et al. 1984). Their PBS is complementary to phenylalanine tRNA, which is different from the primer of all known exogenous retroviruses. (Exogenous retroviruses use Trp, Pro, or Lys tRNA primers.)

The origin of the IAP genes is unclear, since they have a high copy number in some species but are absent in related species (Syrian vs. Chinese hamsters). If they are endogenous retroviruses in the sense described above, the related exogenous retrovirus has disappeared and the difference in copy number relates to the extent of IAP gene activity in the germ line of different species. Alternatively, IAP sequences could be cellular movable genetic elements—that is, not evolved from an exogenous retrovirus.

VL30 Genes

VL30 genes code for an RNA that has *cis*-acting sequences recognized by mammalian C-type retrovirus proteins. VL30 genes have a structure like that of a retrovirus provirus (Rotman et al. 1984). They form a fairly large family (100 copies per genome) of related genes in mice and rats whose RNA is encapsidated and reverse transcribed, the resulting DNA being integrated by mouse leukemia-virus proteins. They have PBS's complementary to Gly, Pro, and Gln tRNAs (Itin and Keshet 1985).

So far no proteins specified by these VL30 sequences have been described. Thus, VL30 DNA may be DNA sequences that, by mutation or recombination, have evolved *cis*-acting sequences that are recognized by proteins of exogenous murine-leukemia retroviruses. Alternatively, they may be the descendants of long-extinct exogenous retroviruses. These VL30 sequences can be amplified and spread by passage in virions formed with the proteins of murine retroviruses. The wide distribution of the VL30 sequences—a distribution wider than that of any endogenous retrovirus—may indicate their success in finding the general recognition features of retrovirus proteins.

Endogenous Retrovirus-like Sequences

In addition, there are in the mouse and human genomes (the only ones thoroughly studied) numerous sequences with LTRs, PBS, and PPT sequences and, sometimes, some slight nucleic-acid sequence similarity to exogenous retroviruses. For example, the human genome contains at least four different such elements (Mager and Henthorn 1984; O'Connell et al. 1984; Steele et al. 1984). Two are reiterated approximately 100 times per genome, and one is reiterated approximately 1,000 times per genome. There is no evidence as to whether these retrovirus-like sequences are now capable of transposing. Their origins are also obscure. As has been discussed above re IAP genes, an ancient, long-extinct ancestral exogenous retrovirus can be postulated. Since one of these families has a histidine tRNA PBS, another glutamine, and another arginine, all of which are different from those in any sequenced exogenous retrovirus, the ancestral retroviruses, if they existed, must have been from extinct retrovirus species. (As noted above, a VL30 sequence also has a glutamine tRNA PBS.) Alternatively, these elements could have evolved from the cell genome and be cellular movable genetic elements.

Another family of repeated sequences (approximately 500 copies per genome in mice) resembles solitary retrovirus LTRs (Wirth et al. 1984). They form a 4-bp direct repeat of cell DNA on integration. Furthermore, these LTR-like sequences apparently can recombine with exogenous retroviruses, perhaps indicating some sequence similarity (Schmidt et al. 1984). In addition, there is evidence of other such retrovirus-like elements from DNA sequences, protein, and electron-microscopic analyses (Jerabek et al. 1984; Stumpf et al. 1984; Suni et al. 1984; Yanage and Szollosi 1984).

The frequency of these elements in the germ line (more than 1,500 copies per genome of mouse and man) and the potential mutagenic effect of each integration indicates that a severe mutagenic load would have been produced by the introduction and amplification of these elements (also see Doolittle and Sapienza 1980; Orgel and Crick 1980). Thus, there would have been strong selection for mechanisms to control this infection and/or amplification, although the nature of such controls is unclear.

Retrotransposons

Tyl Elements of Yeast

Vertebrates are not the only eukaryotes with retrovirus-like elements in their genomes. The well-characterized Tyl cellular movable genetic element of yeast has now been definitively shown to transpose through an RNA intermediate; for example, increased transcription results in increased transposition, internal introns are removed during transposition, and there is transfer of information from the 3′ parental LTR to the 5′ progeny LTR during transposition (Boeke et al. 1985). It also appears that Tyl elements use retrovirus-like mechanisms for gene expression, e.g., expression of a distal gene as a read-through fusion-product with the product of a proximal gene (like the expression of the retrovirus reverse-transcriptase gene mentioned above) (Mellor et al. 1985).

Overexpression of a modified Tyl element results in very slow growth of yeast cells. Most likely this slow growth is a result of the mutagenic effect of increased Tyl insertions, indicating a genetic load like that discussed above for germ-line integrations.

Copia-like Elements of *Drosophila*

Drosophila melanogaster contains numerous families of cellular movable genetic elements. One class of such elements has terminal inverted repeats embedded in terminal direct repeats (LTRs). On the basis of similarities of organization, nucleic-acid sequence, unintegrated intermediates, and even particles, several laboratories have suggested that these elements transpose through an RNA intermediate (Arkhipova et al. 1984; Bayev et al. 1984; Flavell 1984; Inouye et al. 1984; Saigo et al. 1984). Although no definitive evidence has been published, this hypothesis seems reasonable in terms of what has been discussed above.

There also is some suggestion of retrovirus-like cellular movable genetic elements in other organisms, including plants and protozoans (Hasan et al. 1984; Shepherd et al. 1984). However, the evidence is less direct.

The existence of retrovirus-like cellular movable genetic elements raises the same problem of origins as that of the endogenous sequences that are similar to replication-competent retroviruses. One hypothesis is that all of these elements are similar to endogenous retroviruses and arose from infection with exogenous retroviruses. The presence of *copia* particles in *Drosophila* cells seems to support this hypothesis. However, the particles may be necessary for efficient reverse transcription, not a sign of exogenous infection. (This hypothesis seems correct for intracisternal A-particles.) If

the hypothesis of an infectious origin is correct, it raises questions concerning the origin of these exogenous retroviruses and why they are now extinct. Another hypothesis is that these elements are protoviruses, i.e., elements that have evolved from the cell genome and have some virus properties but cannot form infectious virus particles. In particular, some of these elements may not have an encapsidation sequence or genes for virion structural proteins. In this respect they would be complementary to the VL30 sequences, which have encapsidation sequences but apparently code for no replicative enzymes. However, since there are endogenous retroviruses (see above) and since exogenous retroviruses must have evolved from cells at some time (unless they were the original genetic system), both hypotheses of origin of these sequences—i.e., from cellular sequences and from exogenous retroviruses—are true in some cases. The often unresolvable problem is, Which hypothesis is correct in the case of any particular element?

Retrotranscripts

A large (10%) further fraction of the vertebrate genome seems to have been formed by reverse transcription and integration (Bennett et al. 1984). This fraction includes cDNA sequences from protein-coding genes, repeated DNAs homologous to common small RNAs, and other repeated-DNA families. The evidence for a role of reverse transcription in the origin of these repeated sequences is entirely circumstantial, and in some important respects the processes are different from those used by the elements described above.

DNA is hypothesized to result from reverse transcription of RNA on the basis of (1) its similarity to an expressed gene in a different chromosomal location and (2) some or all of the following: direct repeats surrounding the DNA sequence (also characteristic of DNA transposition), 3′ poly (dA), and loss of intervening sequences (fig. 3). The loss of intervening sequences is most convincing, although it should be remembered that retroviruses maintain the intervening sequence for the *env* gene because it contains the viral encapsidation sequence. In addition, reverse transcription in retrovirus replication involves loss of the viral RNA poly (A) and formation of constant, virus-species-specific-sized direct repeats of cellular DNA on integration.

cDNA Pseudogenes

These are apparent reverse transcripts of known genes. They have been described for a large number of different genes. For example, such cDNA or processed pseu-

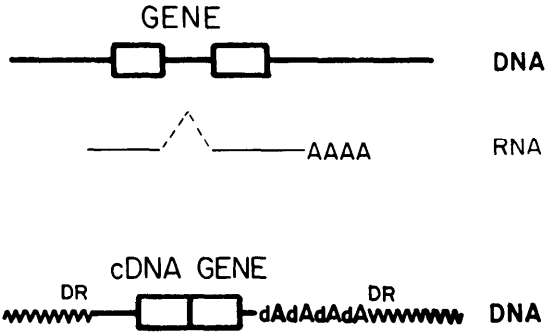

FIG. 3.—Formation of retrotranscript. The heavy line represents DNA; the open box represents the coding sequence; the light line represents RNA; the dotted line represents the intervening sequence; DR = direct repeat in surrounding chromosomal DNA; AAAA = poly (A); dAdAdAdA = poly (dA).

dogenes have recently been described for chicken calmodulin (Stein et al. 1983); mouse p53 (Zakut-Houri et al. 1983); mouse ribosomal protein L7 (Klein and Meyuhas 1984); human dihydrofolate reductase (Anagnou et al. 1984); human argininosuccinate synthetase (Freytag et al. 1984); rat cytochrome c (Scarpulla 1984); mouse ribosomal protein L30 (Wiedermann and Perry 1984); mouse, rat, and human glyceraldehyde 3-phosphate dehydrogenase (Piechaczyk et al. 1984); mouse myosin (Robert et al. 1984); mouse cytochrome c (Limbach and Wu 1985); and mouse cytokeratin endo A (Vasseur et al. 1985). They had been earlier described for some globins, immunoglobulins, and tubulins (Ueda et al. 1982). The continued discovery of such genes suggests that most protein-coding genes will be found to have a cDNA copy (Antoine and Niessing 1984). All of these cDNA copies are pseudogenes, and the amount of difference from the coding gene has been used as a measure of the time since the cDNA pseudogenes were formed.

Genes for Small RNAs

Evidence, including 3′ poly (dA) and direct repeats, has been presented that many small nuclear RNA pseudogenes—e.g., U1, U2, U3, U4, U6, and 7SL—are the result of reverse transcription (approximately 1,000 copies per genome) (Reilly et al. 1982; Bernstein et al. 1983; Ullu and Weiner 1984; Van Arsdell and Weiner 1984). Most of these pseudogenes seem to be truncated either at their 3′ ends or at their 5′ ends. Furthermore, recent evidence indicates that Alu DNA, a short sequence present in 500,000 copies per genome and dispersed throughout the human and mouse genomes (Schmid and Jelinek 1982), is also a partial reverse transcript of 7SL RNA (Ullu and Tschudi 1984).

Other Repeated DNA Sequences

These are often separated into shorter and longer sequences, but this may not be a meaningful separation (Singer 1982) (see below). The shorter sequences include the Alu sequences of man (discussed above), the related B1 sequences of mouse, the B2 sequences of mouse and the related ID sequences of rat (Milner et al. 1984), and the R sequences of mouse, which appear to be the 3′ end of the MIF-1 sequence (Bennett and Hastie 1984) (discussed below). In total these shorter sequences are present at more than 500,000 copies per mouse or human genome, about 3% of the genome. As suggested above, they may be the reverse transcripts of functional RNAs.

The longer element (>6 kbp) is called MIF-1, *Bam*HI, or L1 in mouse and *Kpn*I in man (DiGiovanni et al. 1983; Meunier-Rotival and Bernardi 1984). It is present in approximately 30,000 copies per genome and represents ~5% of the cell genome. It has a 3′ AATAAA followed by poly (dA) (which is characteristic of sequences coding for the 3′ end of mRNAs) and can be transcribed into poly (A) RNA. Different copies have a relatively constant 3′ end and a variable 5′ end consistent with truncated reverse transcription from the 3′ poly (A) (Voliva et al. 1984). There are direct repeats of 7–12 bp in different copies. These elements can transpose (Katzir et al. 1985). Longer copies may have a conserved open reading frame (Martin et al. 1984). If this open reading frame coded for a transposase or reverse transcriptase or both, these elements would be retrotransposons rather than retrotranscripts.

Mechanism of Reverse Transcription

There are several difficulties in understanding the origin of the retrotranscripts. They do not appear to have the specific primers and repeats used in retrovirus reverse

transcription—i.e., PBS, PPT, and R (fig. 1)—nor the constant termini of retroviruses and retrotransposons (figs. 1 and 2). Thus, other primers are needed for both strands of DNA synthesis, reverse-transcriptase activity is needed for DNA synthesis, and some enzyme activities are needed for integration and formation of direct repeats. It has been suggested that retrotranscripts might self-prime first-strand DNA synthesis by formation of a hairpin loop (Van Arsdell and Weiner 1984). Although this hypothesis might explain 3′ truncation, it would not explain the frequent persistence at the 3′ ends of AATAAA followed by poly (dA).

The origin of the reverse transcriptase used to synthesize retrotranscripts is also unknown. It could be coded by a retrovirus or retrotransposon, or it could be a modified cellular enzyme (Mondal and Hofschneider 1983). The reverse-transcriptase activity would either have to be present in germ cells or be brought in by infection. Evidence has been presented that both *env* mRNA in somatic cells after infection and apparently random RNAs encapsidated in a mutant retrovirus virion in disrupted virions can be reverse transcribed with a tRNA primer (Spodick et al. 1984; Taylor and Cywinski 1984).

The means of integration used by these sequences is also a problem. The retrovirus integrase seems to have DNA sequence specificity (Panganiban and Temin 1984). Furthermore, the size of the direct repeats surrounding retrovirus proviruses or retrotransposon DNA is fixed for any one species of virus or type of retrotransposon. Retrotranscripts do not contain homologues of the retrovirus attachment sequence; nor are the sizes of the direct repeats constant for any one element or gene. In addition, the direct repeats surrounding retrotranscripts are usually much larger (5–19 bp) than those so far seen for retroviruses and retrotransposons (4–6 bp). These differences could indicate an abnormal activity of the integrase from a retrovirus or retrotransposon. Alternatively, the differences could be a consequence of the use of a cellular enzyme, perhaps one with another primary activity.

Transcription of the RNA to be reverse transcribed is easier to understand. Polymerase II or III promoters and poly (A) addition signals are often present in possible "parental" sequences of retrotranscripts. Transcription in the germ line or its embryonic precursors would have to occur. Over evolutionary time, such transcription might have occurred by chance.

Evolutionary Questions

Two major evolutionary questions can be considered in addition to those discussed already: What is the evolutionary relationship of the elements in the chromosome to the viruses? What is the evolutionary role of all of these sequences and the influence on evolution of the processes that resulted in their existence?

Most of the data about these elements are from nucleic-acid sequence analysis. More data about polymorphisms in different populations and species would be useful in determining evolutionary pathways. However, the existence of all of these elements and viruses suggests that the cellular genome may be involved in constant interchange with exogenous elements; that is, there may have been repeated exchanges between endogenous and exogenous sequences. If this hypothesis is true, the first question may be undecidable. There may have been too much information transfer from chromosome to virus—as well as the reverse—to find intermediates and thus determine evolutionary history.

One evolutionary effect of these sequences and their germ-line integrations must have been mutation. Retroviruses integrate at random, and retrovirus and IAP insertion

have been shown to be mutagenic. Repeated sequences have been found near almost all genes. Both coding and controlling sequences are susceptible to interruption by these elements. In addition, the elements frequently have their own controlling sequences, which could cause further mutations. Thus, present-day organisms are descendants of those that survived this mutational load. Those organisms in which these reverse transcribed sequences caused lethal or deleterious mutations would not have survived. We cannot estimate the total number of integrations of reverse transcribed DNA because now we only see the ones that did not cause lethal or deleterious mutations.

It has also been suggested that the inserted sequences may have functional roles in genomic organization, e.g., to prevent recombination, to maintain sequences by gene conversion, or even to be control sequences (Erwin and Valentine 1984; Kress et al. 1984; Schimenti and Duncan 1984; Martin et al. 1985). They also could have been an intermediate in the formation of new genes by duplication and mutation. However, there is no evidence for this hypothesis. Integration of reverse transcripts does increase the genome size and thus reduces the probability that lethal or deleterious mutations will result from integration of other elements. However, other mechanisms that might expand the genome seem simpler.

Some of the mechanisms that organisms could have evolved to reduce the frequency of such mutations or to control the mutational effect of these insertions are (1) resistance to virus infection, (2) suppression of mutations induced by integration of reverse-transcribed DNA, (3) repression of element transcription in the germ line, and (4) restriction of integration of exogenous DNA to certain regions in the genome. However, the large number of integrations of reverse transcripts present in the human or mouse genome suggests that these integrations were a major cause of death of organisms during evolution. Rarely, perhaps, integration of reverse transcripts also could have been a source of variation during times of rapid evolution, especially in asexual or inbred species.

Acknowledgments

I thank J. Crow, J. Embretson, M. Emerman, T. Gilmore, C. Miller, and R. Temin for helpful suggestions. The work from my laboratory was supported by Public Health Service grants CA-22443 and CA-07175 from the National Institutes of Health. I am an American Cancer Society Research Professor.

LITERATURE CITED

Earlier references can be found in Weiss et al. (1982) and Temin and Engels (1984). Usually only more recent references are cited here.

ANAGNOU, N. P., S. J. O'BRIEN, T. SHIMADA, W. G. NASH, M.-J. CHEN, and A. W. NIEHUIS. 1984. Chromosomal organization of the human dihydrofolate reductase genes: dispersion, selective amplification, and a novel form of polymorphism. Proc. Natl. Acad. Sci. USA **81**: 5170–5174.

ANTOINE, M., and J. NIESSING. 1984. Intron-less globin genes in the insect *Chironomus thummi thummi*. Nature **310**:795–798.

ARKHIPOVA, I. R., T. V. GORELOVA, Y. V. ILYIN, and N. G. SCHUPPE. 1984. Reverse transcription of *Drosophila* mobile genetic element RNAs: detection of intermediate forms. Nucleic Acids Res. **12**:7533–7548.

BAYEV, A. A., JR., N. V. LYUBOMIRSKAYA, E. B. DZHUMAGALIEV, E. V. ANANIEV, I. G. AMIANTOVA, and Y. V. ILYIN. 1984. Structural organization of the transposable element

*mdg*4 from *Drosophila melanogaster* and a nucleotide sequence of its long terminal repeat. Nucleic Acids Res. **12**:3707–3723.

BENNETT, K. L., and N. D. HASTIE. 1984. Looking for relationships between the most repeated dispersed DNA sequences in the mouse: small R elements are found associated consistently with long MIF repeats. EMBO J. **3**:467–472.

BENNETT, K. L., R. E. HILL, D. F. PIETRAS, M. WOODWOTH-GUTAI, C. KANE-HAAS, J. M. HOUSTON, J. K. HEATH, and N. D. HASTIE. 1984. Most highly repeated dispersed DNA families in the mouse genome. Mol. Cell. Biol. **4**:1561–1571.

BERGER, S. A., and A. BERNSTEIN. 1985. Characterization of a retrovirus shuttle vector capable of either proviral integration or extrachromosomal replication in mouse cells. Mol. Cell. Biol. **5**:305–312.

BERNSTEIN, L. B., S. M. MOUNT, and A. M. WEINER. 1983. Pseudogenes for human small nuclear RNA U3 appear to arise by integration of self-primed reverse transcripts of the RNA into new chromosomal sites. Cell **32**:461–472.

BOEKE, J. D., D. J. GARFINKEL, C. A. STYLES, and G. R. FINK. 1985. Ty elements transpose through an RNA intermediate. Cell **40**:491–500.

BUCKLER, C. E., S. P. STAAL, W. P. ROWE, and M. A. MARTIN. 1982. Variation in the number of copies and in the genomic organization of ecotropic murine leukemia virus proviral sequences in sublines of AKR mice. J. Virol. **43**:629–640.

DEJEAN, A., P. SONIGO, S. WAIN-HOBSON, and P. TIOLLAIS. 1984. Specific hepatitis B virus integration in hepatocellular carcinoma DNA through a viral 11 base pair direct repeat. Proc. Natl. Acad. Sci. USA **81**:5350–5354.

DIGIOVANNI, L., S. R. HAYNES, R. MISRA, and W. R. JELINEK. 1983. *Kpn* I family of long-dispersed repeated DNA sequences of man: evidence for entry into genomic DNA of DNA copies of poly(A)-terminated *Kpn* I RNAs. Proc. Natl. Acad. Sci. USA **80**:6533–6537.

DOOLITTLE, W. F., and C. SAPIENZA. 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature **284**:601–603.

ERWIN, D. H., and J. M. VALENTINE. 1984. "Hopeful monsters," transposons, and Metazoan radiation. Proc. Natl. Acad. Sci. USA **81**:5482–5483.

FLAVELL, A. J. 1984. Role of reverse transcription in the generation of extrachromosomal *copia* mobile genetic elements. Nature **310**:514–516.

FREYTAG, S. O., H.-G. O'BOCK, A. L. BEAUDET, and W. E. O'BRIEN. 1984. Molecular structures of human arginosuccinate synthetase pseudogenes. J. Biol. Chem. **259**:3160–3166.

HARBERS, K., M. KUEHN, H. DELIUS, and R. JAENISCH. 1984. Insertion of retrovirus into the first intron of (alpha)1(I) collagen gene leads to embryonic lethal mutation in mice. Proc. Natl. Acad. Sci. USA **81**:1504–1508.

HARRIS, J. D., H. BLUM, J. SCOTT, B. TRAYNOR, P. VENTURA, and A. HAASE. 1984. Slow virus visna: reproduction in vitro of virus from extrachromosomal DNA. Proc. Natl. Acad. Sci. USA **81**:7212–7215.

HASAN, G., M. J. TURNER, and J. S. CORDINGLEY. 1984. Complete nucleotide sequence of an unusual mobile element from *Trypanosoma brucei*. Cell **37**:333–341.

HAWLEY, R. G., M. J. SHULMAN, and N. HOZUMI. 1984. Transposition of two different intracisternal A particle elements into an immunoglobulin Kappa-chain gene. Mol. Cell. Biol. **4**: 2565–2572.

HULL, R., and S. N. COVEY. 1983. Does cauliflower mosaic virus replicate by reverse transcription? Trends Biochem. Sci. **8**:119–121.

HUTCHISON, K. W., N. G. COPELAND, and N. A. JENKINS. 1984. Dilute-coat-color locus of mice: nucleotide sequence analysis of the d^{+2J} and d^{+Ha} reverent alleles. Mol. Cell. Biol. **4**: 2899–2904.

INOUYE, S., S. YUKI, and K. SAIGO. 1984. Sequence-specific insertion of the *Drosophila* transposable genetic element 17.6. Nature **310**:332–333.

ITIN, A., and E. KESHET. 1985. Primer binding sites corresponding to several tRNA species are present in DNAs of different members of the same retrovirus-like gene family (VL30). J. Virol. **54**:236–239.

JAENISCH, R., D. JAEHNER, P. NOBIS, I. SIMON, J. LOEHLER, K. HARBERS, and D. GROTKOPP. 1981. Chromosomal position and activation of retroviral genomes inserted into the germ line of mice. Cell **24**:519–529.

JERABECK, L. B., R. C. MELLORS, K. B. ELKON, and J. W. MELLORS. 1984. Detection and immunochemical characterization of a primate type C retrovirus–related p30 protein in normal human placentas. Proc. Natl. Acad. Sci. USA **81**:6501–6506.

KATZIR, N., G. RECHAVI, J. B. COHEN, T. UNGER, F. SIMONI, S. SEGAL, D. COHEN, and D. GIVOL. 1985. "Retroposon" insertion into the cellular oncogene c-*myc* in canine transmissible venereal tumor. Proc. Natl. Acad. Sci. USA **82**:1054–1058.

KLEIN, A., and O. MEYUHAS. 1984. A multigene family of intron lacking and containing genes, encoding mouse ribosomal protein L7. Nucleic Acids Res. **12**:3763–3776.

KRESS, M., Y. BARRA, J. G. SEIDMAN, G. KHOURY, and G. JAY. 1984. Functional insertion of an Alu type 2 (B2 SINE) repetitive sequence in murine class I genes. Science **226**: 974–977.

LIMBACH, K. J., and R. WU. 1985. Characterization of a mouse somatic cytochrome c gene and three cytochrome c pseudogenes. Nucleic Acids Res. **13**:617–630.

MAGER, D. L., and P. S. HENTHORN. 1984. Identification of a retrovirus-like repetitive element in human DNA. Proc. Natl. Acad. Sci. USA **81**:7510–7514.

MANDART, E., A. KAY, and F. GALIBERT. 1984. Nucleotide sequence of a cloned duck hepatitis B virus genome: comparison with woodchuck and human hepatitis B virus sequences. J. Virol. **49**:782–792.

MARCO, Y., and S. H. HOWELL. 1984. Intracellular forms of viral DNA consistent with a model of reverse transcriptional replication of the cauliflower mosaic virus genome. Nucleic Acids Res. **12**:1517–1528.

MARTIN, S. L., C. F. VOLIVA, F. H. BURTON, M. H. EDGELL, and C. A. HUTCHISON III. 1984. A large interspersed repeat found in mouse DNA contains a long open reading frame that evolves as if it encodes a protein. Proc. Natl. Acad. Sci. USA **81**:2308–2312.

MARTIN, S. L., C. F. VOLIVA, S. C. HARDIES, M. H. EDGELL, C. A. HUTCHISON III. 1985. Tempo and mode of concerted evolution in the L1 repeat family of mice. Mol. Biol. Evol. **2**:127–140.

MELLOR, J., S. M. FULTON, M. J. DOBSON, W. WILSON, S. M. KINGSMAN, and A. J. KINGSMAN. 1985. A retrovirus-like strategy for expression of a fusion protein encoded by yeast transposon Ty 1. Nature **313**:243–246.

MEUNIER-ROTIVAL, M., and G. BERNARDI. 1984. The Bam repeats of the mouse genome belong in several superfamilies the longest of which is over 9 kb in size. Nucleic Acids Res. **12**: 1593–1608.

MILLER, R. H., P. L. MARION, and W. S. ROBINSON. 1984. Hepatitis B viral DNA-RNA hybrid molecules in particles from infected liver are converted to viral DNA molecules during an endogenous DNA polymerase reaction. Virology **139**:64–72.

MILNER, R. J., F. E. BLOOM, C. LAI, R. A. LERNER, and J. G. SUTCLIFFE. 1984. Brain-specific genes have identifier sequences in their introns. Proc. Natl. Acad. Sci. USA **81**:713–717.

MONDAL, H., and P. H. HOFSCHNEIDER. 1983. Demonstration of free reverse transcriptase in the nuclei of embryonic tissues of the Japanese quail. Biochem. Biophys. Res. Comm. **116**: 303–311.

O'CONNELL, C., S. O'BRIEN, W. G. NASH, and M. COHEN. 1984. ERV3, a full-length human endogenous provirus: chromosomal localization and evolutionary relationships. Virology **138**:225–235.

ORGEL, L. E., and F. H. C. CRICK. 1980. Selfish DNA: the ultimate parasite. Nature **284**:604–607.

PANGANIBAN, A. T., and H. M. TEMIN. 1984. The retrovirus *pol* gene encodes a product required for DNA integration: identification of a retrovirus *int* locus. Proc. Natl. Acad. Sci. USA **81**: 7885–7889.

PIECHACZYK, M., J. M. BLANCHARD, S. RIAAD-EL SABOUTY, C. DANI, L. MARTY, and P.

JEANTEUR. 1984. Unusual abundance of vertebrate 3-phosphate dehydrogenase pseudogenes. Nature **312**:469–471.

REILLY, J. G., R. OGDEN, and J. J. ROSSI. 1982. Isolation of a mouse pseudo tRNA gene encoding CCA—a possible example of reverse flow of genetic information. Nature **300**:287–289.

ROBERT, B., P. DAUBAS, M.-A. AKIMENKO, A. COHEN, I. GARNER, J.-L. GUENET, and M. BUCKINGHAM. 1984. A single locus in the mouse encodes both myosin light chains 1 and 3, a second locus corresponds to a related pseudogene. Cell **39**:129–140.

ROTMAN, G., A. ITIN, and E. KESHET. 1984. 'Solo' large terminal repeats (LTR) of an endogenous retrovirus–like gene family (VL30) in the mouse genome. Nucleic Acids Res. **12**:2273–2282.

SAIGO, K., W. KUGIMIYA, Y. MATSUO, S. INOUYE, K. YOSHIOKA, and S. YUKI. 1984. Identification of the coding sequence for a reverse transcriptase–like enzyme in a transposable genetic element in *Drosophila melanogaster*. Nature **312**:659–661.

SCARPULLA, R. C. 1984. Processed pseudogenes for rat cytochrome C are preferentially derived from one of three alternate mRNAs. Mol. Cell. Biol. **4**:2279–2288.

SCHIMENTI, J. C., and C. H. DUNCAN. 1984. Ruminant globin gene structures suggest an evolutionary role for Alu-type repeats. Nucleic Acids Res. **12**:1641–1655.

SCHMID, C. W., and W. R. JELINEK. 1982. The Alu family of dispersed repetitive sequences. Science **216**:1065–1070.

SCHMIDT, M., K. GLOGGER, T. WIRTH, and I. HORAK. 1984. Evidence that a major class of mouse endogenous long terminal repeats (LTRs) resulted from recombination between exogenous retroviral LTRs and similar LTR-like elements (LTR-IS). Proc. Natl. Acad. Sci. USA **81**:6696–6700.

SHEPHERD, N. S., Z. SCHWARZ-SOMMER, J. B. VEL SPALVE, M. GUPTA, U. WIENAND, and H. SAEDLER. 1984. Similarity of the *Cin*1 repetitive family of *Zea mays* to eukaryotic transposable elements. Nature **307**:185–187.

SINGER, M. F. 1982. SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. Cell **28**:433–434.

SPODICK, D. A., L. H. SOE, and P. ROY-BURMAN. 1984. Genetic analysis of the feline RD-114 retrovirus-related endogenous elements. Virus Res. **1**:543–555.

STEELE, P. E., A. B. RABSON, T. BRYAN, and M. A. MARTIN. 1984. Distinctive termini characterize two families of human endogenous retroviral sequences. Science **225**:943–947.

STEIN, J. P., R. P. MUNJAAL, L. LAGACE, E. C. LAI, B. W. O'MALLEY, and A. R. MEANS. 1983. Tissue-specific expression of a chicken calmodulin pseudogene lacking intervening sequences. Proc. Natl. Acad. Sci. USA **80**:6485–6489.

STUMPF, W. E., C. P. HODGSON, M.-J. TSAI, and B. W. O'MALLEY. 1984. Genomic structure and possible retroviral origin of the chicken CR1 repetitive DNA sequence family. Proc. Natl. Acad. Sci. USA **81**:6667–6671.

SUMMERS, J., and W. S. MASON. 1982. Replication of the genome of a hepatitis B–like virus by reverse transcription of an RNA intermediate. Cell **29**:403–415.

SUNI, J., A. NARVANEN, T. WAHLSTROM, M. AHO, R. PAKKANEN, A. VAHERI, T. COPELAND, M. COHEN, and S. OROSZLAN. 1984. Human placental syncytiotrophoblastic Mr 75,000 polypeptide defined by antibodies to a synthetic peptide based on a cloned human endogenous retroviral sequence. Proc. Natl. Acad. Sci. USA **81**:6197–6201.

TAYLOR, J. M., and A. CYWINSKI. 1984. A defective retrovirus particle (SE21Q1b) packages and reverse transcribes cellular RNA, utilizing tRNA-like primers. J. Virol. **51**:267–271.

TEMIN, H. M. 1970. Malignant transformation of cells by viruses. Perspect. Biol. Med. **14**: 11–26.

———. 1971. The protovirus hypothesis. J. Natl. Cancer Inst. **46**:III–VII.

———. 1974. On the origin of RNA tumor viruses. Annu. Rev. Genet. **8**:155–177.

———. 1980. Origin of retroviruses from cellular moveable genetic elements. Cell **21**:599–600.

———. 1981. Structure, variation, and synthesis of retrovirus long terminal repeat. Cell **27**: 1–3.

———. 1982. Function of the retrovirus long terminal repeat. Cell **28**:3–5.

———. 1985. Developments in molecular virology: cloning of retrovirus DNA in bacteria and cloning of other DNA in retroviruses. Pp. 3–14 *in* Y. BECKER, ed. Recombinant DNA research and virus. Martinus Nijhoff, Boston.

TEMIN, H. M., and W. ENGELS. 1984. Movable genetic elements and evolution. Pp. 173–201 *in* J. W. POLLARD, ed. Evolutionary prospects in the 1980s. Wiley, Chichester, England.

TOH, H., H. HAYASHIDA, and T. MIYATA. 1983. Sequence homology between retroviral reverse transcriptase and putative polymerases of hepatitis B virus and cauliflower mosaic virus. Nature **305**:827–829.

UEDA, S., S. NAKAI, Y. NISHIDA, H. HISAJIMA, and T. HONJO. 1982. Long terminal repeat–like elements flank a human immunoglobulin epsilon pseudogene that lacks introns. EMBO J. **1**:1539–1544.

ULLU, E., and C. TSCHUDI. 1984. Alu sequences are processed 7SL RNA genes. Nature **312**: 171–172.

ULLU, E., and A. M. WEINER. 1984. Human genes and pseudogenes for the 7SL RNA component of signal recognition particle. EMBO J. **3**:3303–3310.

VAN ARSDELL, S. W., and A. M. WEINER. 1984. Pseudogenes for human U2 small nuclear RNA do not have a fixed site of 3′ truncation. Nucleic Acids Res. **12**:1463–1471.

VARMUS, H. E. 1982. Form and function of retroviral proviruses. Science **216**:812–820.

VASSEUR, M., P. DUPREY, P. BRULET, and F. JACOB. 1985. One gene and one pseudogene for cytokeratin endo A. Proc. Natl. Acad. Sci. USA **82**:1155–1159.

VOLIVA, C. F., S. L. MARTIN, C. A. HUTCHISON III, and M. H. EDGELL. 1984. Dispersal process associated with the L1 family of interspersed repetitive DNA sequences. J. Mol. Biol. **178**: 795–813.

VOLVITCH, M., N. MODJTAHEDI, P. YOT, and G. BRUN. 1984. RNA-dependent DNA polymerase activity in cauliflower mosaic virus–infected plant leaves. EMBO J. **3**:309–314.

WEISS, R., N. TEICH, H. E. VARMUS, and J. COFFIN, eds. 1982. RNA tumor viruses: the molecular biology of tumor viruses, 2d ed. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

WIEDERMANN, L. M., and R. P. PERRY. 1984. Characterization of the expressed and several processed pseudogenes for the mouse ribosomal protein L30 gene family. Mol. Cell. Biol. **4**: 2518–2528.

WIRTH, T., M. SCHMIDT, T. BAUMRUKER, and I. HORAK. 1984. Evidence for mobility of a new family of mouse middle repetitive DNA elements (LTR-IS). Nucleic Acids Res. **12**: 3603–3610.

YANAGE, Y. Y., and D. SZOLLOSI. 1984. Virus-like particles and related expressions in mammalian oocytes and preimplantation stage embryos. Pp. 218–234 *in* J. VAN BLERKOM and P. MOTTA, eds. Ultrastructure of reproduction. Martinus Nijhoff, Boston.

ZAKUT-HOURI, R., M. OREN, B. BIENZ, V. LAVIE, S. HAZUM, and D. GIVOL. 1983. A single gene and a pseudogene for the cellular tumor antigen p53. Nature **306**:594–597.

WALTER M. FITCH, reviewing editor

Received April 18, 1985; revision received June 26, 1985.