Retroviral DNA Transposition: Themes and Variations

ANNA MARIE SKALKA
Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, PA 19111, United States

ABSTRACT Retroviruses and LTR retrotransposons are transposable elements that encapsidate the RNAs that are intermediates in the transposition of DNA copies of their genomes (proviruses), from one cell (or one locus) to another. Mechanistic similarities in DNA transposase enzymes and retroviral/retrotransposon integrases underscore the close evolutionary relationship among these elements.

The retroviruses are very ancient infectious agents, presumed to have evolved from Ty3/Gypsy LTR retrotransposons (1), and DNA copies of their sequences can be found embedded in the genomes of most, if not all, members of the tree of life. All retroviruses share a specific gene arrangement and similar replication strategies. However, given their ancestries and occupation of diverse evolutionary niches, it should not be surprising that unique sequences have been acquired in some retroviral genomes and that the details of the mechanism by which their transposition is accomplished can vary.

While every step in the retrovirus lifecycle is, in some sense, relevant to transposition, this Chapter focuses mainly on the early phase of retroviral replication, during which viral DNA is synthesized and integrated into its host genome. Some of the initial studies that set the stage for current understanding are highlighted, as well as more recent findings obtained through use of an ever-expanding technological toolbox including genomics, proteomics, and siRNA screening. Persistence in the area of structural biology has provided new insight into conserved mechanisms as well as variations in detail among retroviruses, which can also be instructive.

PROVIRAL GENE ORGANIZATION AND EXPRESSION

Two unique features of retroviruses are their genome organization and strategies for gene expression. While the gene order of retroviruses is conserved, the synthesis of retroviral proteins can be controlled by different mechanisms. Variations include the ways in which RNA splicing is used to produce mRNAs from a single long transcript, a portion of which must remain unspliced to serve as the viral genome. With these general principles in mind, an overview of the retrovirus family follows, along with a description of how the gene organization of its members relates to gene expression and the viral entry process. Detailed descriptions of the molecular aspects of these processes can be found in the comprehensive Retroviruses (2) an overview chapter in Fields Virology (3), or a number of recent reviews that focus mainly on the early steps in the reproductive cycle of human immunodeficiency virus type-1 (HIV-1) (4–11).

The family Retroviridae (Fig. 1) includes two subfamilies. The subfamily Oncoretrovirinae comprises six genera, five of which are designated alpha- to epsilon-retroviruses, and the sixth comprises the lentiviruses. The subfamily Spumavirinae includes the “unconventional” spumaviruses (12). The organization of all retroviral proviruses follows the general pattern illustrated at the top of Figure 2. Open reading frames are flanked by long terminal repeats (LTRs) of varying lengths, which include sequences that direct transcription by the host cell RNA polymerase II. The gene order starts with those encoding structural proteins (gag) and enzymes (pro, pol), followed by a gene (env) for an envelope.
FIGURE 1 A phylogenetic tree based on the collection of RT-IN sequences for all retrotranscribing viruses in the NCBI taxonomy and NCBI RefSeq databases. Full-length Pol or Gag-Pol sequences for all viruses were downloaded and truncated based on their alignment with the RT-IN region of HIV-1. ClustalX algorithm with neighbor joining clustering was then used for tree reconstruction. The tree in the illustration is an artistic representation, based on the results. Red boxes identify viral species discussed in this overview: for alpharetroviruses, RSV is Rous sarcoma virus and ALV avian leukosis virus; betaretrovirus MMTV is mouse mammary tumor virus; gammaretrovirus MMLV is Moloney murine leukemia virus; deltaretrovirus HTLV-1 is human T-lymphotropic virus 1; epsilonretrovirus WDSV is walleye dermal sarcoma virus; lentivirus HIV-1 is human immunodeficiency virus type 1; and spumaretrovirus PFV is the prototype foamy virus. The analysis and Figure were kindly provided by Dr. Vladimir Belyi, Rutgers Cancer Institute of New Jersey. doi:10.1128/microbiolspec.MDNA3-0005-2014.f1
FIGURE 2  The organization of proviruses in each of the retroviral genera. A generic proviral map is included at the top. Representative genomes have been aligned to allow comparisons, and are not to scale. Viral species are identified in the Figure 1 legend. Origins of the major transcripts of ALV, MLV, and PFV are represented by green arrows below the maps. Translational frameshifts are indicated by descending arrows, and read-throughs by vertical arrows in the gene coding regions. doi:10.1128/microbiolspec.MDNA3-0005-2014.f2
protein that is inserted into the membrane that surrounds the viral capsid, and which determines host cell receptor specificity. Retroviruses with “simple” genomes (i.e. alpharetroviruses and gammaretroviruses) have only these essential genes. The genomes of the other retroviruses include additional sequences that specify auxiliary proteins, most encoded in alternative reading frames that begin on either side of env.

As illustrated for a representative alpharetrovirus, gammaretrovirus, and spumavirus (Fig. 2), one long transcript, which includes the R and U5 regions from the upstream LTR and the U3 and R regions from the downstream LTR, serves as gag mRNA for all retroviruses. Env protein is always translated from a spliced mRNA, with sub-optimal splicing signals regulating the amounts of gag and env mRNAs (and the corresponding major structural and envelope polyproteins) that are appropriate for particle production. For all but the spumaviruses, pol gene expression is regulated at the level of translation. Gag-Pol polyproteins are produced from the full-length gag transcripts by occasional ribosomal frameshifting or codon read-through at the gag-pro or pro-pol borders. This mechanism limits the amount of Gag-Pol to approximately 5 to 10% of Gag alone, a ratio that is optimal for packaging into progeny particles. More recent studies have revealed an exception to this paradigm: the Pol protein of spumaviruses is translated from a spliced (pro, pol) mRNA (Fig. 2), so that the Gag to Pol ratio is regulated by splicing rather than by ribosomal frameshifting. Multiply spliced mRNAs, characteristic of retroviruses with complex genomes, typically encode the auxiliary proteins that are important for control of viral gene expression, replication, and pathogenesis of these viruses.

The retroviral mRNAs include cis-acting signals that facilitate nuclear exit and translation. These mRNAs are translated either by cytoplasmic ribosomes (gag and gag-pro-pol) or those attached to the rough endoplasmic reticulum (env) (Fig. 3, Late Phase). For the assembly of each virus particle, these viral polyproteins and two copies of full-length viral RNA accumulate at the plasma membrane, as illustrated for the alpharetrovirus, avian leukosis virus (ALV), or in the cytoplasm, for betaretroviruses and spumaviruses. Progeny virus particles are released by budding at sites where the processed Env proteins (surface, SU, and trans membrane, TM, components) have accumulated, either at the plasma membrane or into intracellular vesicles, via co-option of the host cell’s intracellular vesicle trafficking machinery. With all but the spumaviruses, mature, infectious virions are generated, during or following budding, by cleavage of the Gag and Gag-Pol polyprotein precursors by the viral protease, which produces the mature viral structural proteins (matrix, MA, capsid, CA, nucleocapsid, NC, and protease, PR) and active viral enzymes (reverse transcriptase, RT, and integrase, IN). This proteolytic processing reaction is associated with condensation of the capsid structures.

The assembly and maturation pathway of all retroviruses positions the processed Pol products, RT and IN, within the viral capsids in preparation for the early steps in infection. Exit from a particle-producing cell and subsequent entry into the cytoplasm of a naive cell is a major distinction between the retroviruses and the intracellular cycling, LTR retrotransposons. With the latter, particles accumulate in the cytoplasm of the producing cell and DNA copies are inserted at new sites when a pre-integration complex enters the nucleus of the same cell (see Chapter 42 Sandmeyer). Retroviral particles attach to specific receptors on host cells via interaction with their envelope proteins (Fig. 3, Early Phase). Capsid entry into the cytoplasm is mediated via membrane fusion, which is triggered by major rearrangements in the viral envelope protein. Such fusion takes place either at the cell surface or following particle uptake into endosomes, depending on the virus and, in some cases, the target cell type.

SYNTHESIS OF RETROVIRAL DNA

Although the biochemistry of the RT enzymes is fairly well understood (see Chapter 46 Hughes), and routinely exploited in biotechnology, details of the retroviral reverse transcription process that takes place within infected cells have not been fully elucidated. The current state of the field is a dichotomy: seminal and atomic level studies of isolated RT enzymes, versus the continuing exploration of RT complexes in infected cells, which may include viral and host proteins that participate in the reaction. The latter studies have been driven forward by the use of new, post genomics-era technologies.

The Reverse Transcriptase Complex

The steps following virus capsid entry (Fig. 3, Early Phase) are poorly understood, even for the most intensely studied retrovirus, the lentivirus HIV-1. The capsids of HIV-1 and the gammaretrovirus murine leukemia virus (MLV) are known to be incompletely disassembled following entry and to remain associated with the subviral, “reverse transcription complex” (RTC). The RTC contains two copies of viral RNA coated with the viral nucleocapsid (NC) protein, a specific tRNA primer,
endogenous reaction in which both RNA- and DNA-templated viral DNA synthesis occurs (20–22). Studies with cultured cells infected with HIV-1 or the alpharetrovirus, avian sarcoma and leucosis virus (ASLV) have demonstrated that synthesis of viral DNAs can be detected within 2–3 hours following virus entry into the cytoplasm (23, 24). However, the structures of the DNA products isolated from such cells vary among retroviruses, indicating that virus-specific differences in the details or timing of the process exist. While RNA-templated minus-strand DNA of the beta- and gammaretroviruses, mouse mammary tumor virus (MMTV) and MLV, are mostly full length, both have large gaps in their plus-strand DNAs. Plus-strands of ASLV are also discontinuous, but include many overlapping “flaps”, which are produced by RT-catalyzed “strand displacement synthesis” as the 5’ end of a previously synthesized plus-strand is removed from the minus-strand template by 3’ the leading edge of a newly synthesized strand. The plus-strand of HIV-1 contains a central flap produced by strand displacement synthesis through a central PPT, which encodes as an additional primer for plus-strand synthesis in this virus. Studies with PICs isolated from cytoplasmic extracts of HIV-1 infected cells have suggested that discontinuities in DNA may persist even after integration (25).

RT-mediated Recombination

The high rate of genetic recombination, a hallmark of retrovirus replication, is facilitated by the presence of two RNA templates within the RTC. Although only one viral DNA molecule is normally produced by each infecting virion, recombination can occur if RT switches from one RNA template to a homologous sequence on the invading plus-strand is joined to a plus-strand of the sequence on the minus-strand DNA synthesized from the first template strand. While this “copy choice” mechanism is probably the most frequent (26), recombination may also occur among DNA products copied from both RNA templates. In the latter model, the 5’-end of a DNA plus-strand that is displaced from the minus-strand DNA synthesized from one RNA template (a DNA “flap”), binds to a complementary sequence on the minus-strand DNA synthesized from the second RNA template. A recombinant is formed when the invading plus-strand is joined to a plus-strand of the second template (27). The ability to form a complete provirus from two damaged RNA templates was probably strongly selected during evolution. If the two RNA templates in an RTC are identical, recombination events are invisible. However, the incorporation of two distinct RNA templates in a single virus particle can lead to the
production of new combinations of sequences. Combinations of viral and cellular sequences may also arise in RTCs from proviral transcripts that include both viral and downstream host sequences that are acquired via abnormal RNA splicing, transcription read-through, or following deletion of downstream proviral sequences. Such transcripts will be incorporated into progeny viral particles by virtue of their viral components. Although rare, such events can give rise to oncogene transducing proviruses that are almost always replication defective, but can have profound effects on host cell biology and evolution. Exactly how the multiple, highly-ordered reactions necessary for producing a single double stranded viral DNA copy from two RNA templates are catalyzed in infected cells is an area of ongoing research.

RT Composition and Organization

While similar reverse transcription reactions are catalyzed by all retroviruses, the structures of the responsible RTs can vary considerably. These findings raise intriguing evolutionary and biochemical questions. For example, retrovirus-specific differences in proteolytic processing of the Gag-Pol polyprotein precursors can lead to the inclusion of additional RT sequences or processing of the Gag-Pol polyprotein precursors can yield Gag-Pro and Gag-Pro-Pol fusions that differ in amino acid sequence downstream of the shift site such that distinct unshifted C-terminal PR and shifted N-terminal RT sequences are encoded from the same nucleotide sequence. The first 27 amino acids of the MMTV RT are derived via frameshift from sequences near the end of the pro gene, that encodes PR, and a 26 amino acid N-terminal stretch in the RT of the deltaretrovirus, bovine leukemis virus (BLV) is acquired by a similar frameshift in pro. The most extreme of these fusions is found in the prototype foamy virus (PFV) where the entire protease region (PR) is attached to the N-terminus of RT. This latter arrangement is somewhat of a puzzle, as PR functions as a dimer but RT as a monomer, and it has been proposed that the PFV PR domain may be dimerized transiently by binding to viral RNA. Biochemical studies indicate that the C-terminal 42 amino acids of the PR domain are important for PFV RT activity and stability.

Biochemistry of Reverse Transcription

A detailed, biochemical analysis of the reverse transcription is provided by Hughes in Chapter 46. This section provides a general description of the process, noting the similarities and differences observed among the various retroviral genera. An outline of the distinct stages in the multi-step process of reverse transcription was deduced mainly from study of the structure and function of purified, bacterial produced enzymes.

FIGURE 3 The single cell reproductive cycle of the alpharetrovirus, ALV. The virus life cycle is divided into an early phase that includes steps from virus infection to establishment of the provirus, and a late phase that includes expression of the provirus and formation of progeny virions. Adapted from: Principles of Virology, 3rd edition Vol. I. 2009.

DNA synthesis is primed by a tRNA that is annealed to the primer binding site (pbs) near the 5'-end of the viral RNA template and extends to the 5'-end of the RNA including a sequence (R), which is repeated at the 3'-end of the viral RNA (Fig. 5A and B). This first template exchange allows continued synthesis of minus-strand DNA and digestion of the viral RNA template by RNase H. Synthesis of plus-strand DNA is primed by a purine rich, RNase H-resistant fragment of RNA (ppt), followed by a second strand exchange at the primer binding site in minus-strand DNA (PBS) (Fig. 5C). The ends of the completed linear DNA product are produced by strand displacement synthesis in which sequences corresponding to the LTRs are copied at each end (Fig. 5D). A single RT molecule contains all of the enzymatic activities required to catalyze each of these steps, but the number of molecules in the RTC that actually take part in viral DNA synthesis in vivo is unknown. Genetic and structural studies suggest that the polymerase and RNase H catalytic sites do not act simultaneously; conformational changes are required to optimize each function (32). In vitro studies of purified HIV-1 RT have revealed the amazing dynamic capabilities of the enzyme, which is able to bind to the primer/substrate in a position poised for catalysis in one direction and to flip 180 degrees to initiate synthesis from the ppt primer. RT flipping and sliding can occur without disengaging from the template (33, 34). Such large-scale molecular contortions may be required to accomplish the various steps in reverse transcription.

Comparison of the biochemical activities of RTs from the various retroviral families has revealed similarities but also notable differences, even among those of closely related viruses, such as the lentiviruses HIV-1 and HIV-2 (28, 35). In some cases, the differences in catalytic activity, or sensitivity to inhibitors, can be explained by particular variations in specific amino acid side chains. The fingers-palm-thumb domains characteristic of all nucleic acid polymerases are also present in all retrovirus RT structures, but amino acid differences in the palm modules can account for the fact that non-nucleoside inhibitors (NNRTIs) that bind to this module and inactivate HIV-1 RT are ineffective against HIV-2 RT.
A. **Initiation of (–) strand DNA synthesis**

The 5′ end of the viral RNA genome is degraded by the RNase H activity of RT as the (–) strand DNA is synthesized.

B. **First template exchange**

The RNA genome continues to be degraded as (–) strand DNA is synthesized

C. **(+)-strand DNA synthesis**

The pbs sequence is copied twice:
- once from the RNA genome
- once from the tRNA primer

RNAse H endonuclease activity of RT removes both primer RNAs

DNA ends are juxtaposed by annealing at complementary PBS sequences

D. **Strand Displacement Synthesis**

Linear Product

(+)-strand synthesis continues from (–)-strong stop DNA

Sequences bordered by the (–)- and (+)-strand start sites (pbs and ppt) are copied twice

**FIGURE 5** Major steps in retroviral reverse transcription. For simplicity, reverse transcription from a single RNA template is shown, and potential (+)-strand synthesis from viral RNA fragments other than the ppt are omitted. RNA is represented by green lines, with key regions identified in lower case: pbs, tRNA primer binding site; u5, unique 5′-end sequence; r, repeated sequence; u3, unique 3′-end sequence, ppt, polypurine tract. Light blue lines represent (–) strand DNA, and dark blue, (+)-strand DNA: key regions are identified in uppercase. A modified base in the tRNA primer (C.) blocks further reverse transcription of the tRNA. Adapted from: Principles of Virology, 3rd edition Vol. I. 2009. S.J Flint, L.W. Enquist V.R Racaniello, and A.M. Skalka, ASM Press Washington DC, Figures 7.3–7.6. doi:10.1128/microbiolspec.MDNA3-0005-2014.f5
Despite their structural and biochemical differences, recent X-ray crystallographic analyses of enzymes bound to cognate DNA-RNA hybrids reveal a strikingly similar topological arrangement of the modular structure among RTs as distinct as the heterodimer of HIV-1, the monomer of MLV-related XMRV, and the homodimer of the yeast retrotransposon Ty3 (Fig. 6). The fingers-palm-thumb modules, and the positions of the nucleic acids, are almost superimposable in the catalytic subunits of these enzymes, even though the dimer architectures of HIV-1 and Ty3 RTs are distinct (36). The RNase H topologies are also similar but in the case of Ty3 the catalytic polymerase and RNase H domains are derived from separate monomers. Equally striking is the similarity in overall topology of the non-catalytic subunits of HIV-1 and Ty3 RTs, in which the RNase H domain of the catalytic polymerase subunit of Ty3 occupies the same position as the connection domain in p51 of HIV-1. The conserved architectures are most evident in features that are critical to the catalytic reactions common to these enzymes. The differences in sequence and quaternary architectures that exist among specific RTs must therefore relate to other functions or interactions that are important in the replication of these viruses and retrotransposons, most of which remain to be discovered.

FIGURE 6 Comparison of the structures of three RTs. Top row: The DNA polymerase domains of lentiviral (HIV-1 p66, left), gammaretroviral (XMRV, center) and LTR-retrotransposon (Ty3 subunit A, right) RTs. Fingers, palm and thumb subdomains are designated F, P and T, respectively. Positions of the RNA (magenta) and DNA strands (teal) of the bound RNA/DNA hybrids are shown. Bottom row: Architectures of the non-catalytic subunits of the dimeric RTs: HIV-1 p51 (left) and Ty3 subunit B (right). Both subunits contain F, P and T subdomains in analogous positions; the p51 connection and Ty3 (subunit B) RNase H domain, denoted C and R respectively, are also in analogous positions. Superposition of the asymmetric p66/p51 HIV-1 RT heterodimer and the symmetric Ty3 (A)/(B) homodimer is shown in the center. HIV RT subunits are in orange and grey, and Ty3 subunits in green and yellow. The illustration was prepared by Drs. Jason Rausch, and Stuart Le Grice, NCI-Frederick, and Dr. Marcin Nowotny, International Institute of Molecular and Cell Biology, Poland. Structure details for HIV-1 RT are in (32, 121), for XMRV RT in (122, 123), and for Ty3 RT in (36). doi:10.1128/microbiolspec.MDNA3-0005-2014.f6
FROM REVERSE TRANSCRIPTION TO A PRE-INTEGRATION COMPLEX

The nucleoprotein dynamics that lead to the production of viral DNA poised for integration are incompletely understood. Among the key questions are: How do RT and IN function coordinately? How do IN-DNA complexes move into the nucleus? Which host proteins participate in, or block these steps? Our current understanding represents an amalgam of historical and recent findings.

It seems likely that structural changes accompany the transition from an RTC to a PIC in an infected cell. Differences in the representation of proteins associated with these complexes have been observed (37), but the details of molecular arrangements within them remain obscure. As the DNA product of RT is the substrate for IN, functional coupling between the two proteins at early stages in the retrovirus life cycle may be expected, and several lines of evidence support this idea.

RT-Integrase (IN) Interactions

The earliest biochemical studies of retroviral RTs focused on proteins purified from alpharetroviruses, because large quantities of particles could be prepared from the plasma of chickens infected with avian myeloblastosis virus or from cultured cells infected with the Rous ASLV. The presence of IN at the C-terminus of the larger, β subunit of RTs of these ASLVs (Fig. 4), prompted initial speculation that RT might be able to catalyze both viral DNA synthesis and its integration. Indeed, the first site-selective nicking of viral DNA end sequences was actually detected with this RT (38). Although it was later shown that the isolated ASLV and MLV IN proteins could catalyze the integration reaction in vitro (39, 40), subsequent genetic and biochemical studies have suggested that retroviral RT and IN proteins may, indeed, interact in functionally relevant ways. For example, physical association between the RT and IN proteins of MLV and HIV-1 has been detected by a variety of assays in vitro (28, 41–46). Investigations with purified HIV-1 RT indicate that IN can stimulate both the initiation and elongation activities of the RT by enhancing its processivity in vitro (47). Site-directed mutagenesis studies have identified residues in the C-terminal domain (CTD) of HIV-1 IN that are critical for RT binding (48). Experiments in yeast cells indicate that the IN protein of Ty3 is required for the initiation of reverse transcription during retrotransposition (49). However, because RT and IN interactions may occur in the context of polyproteins as well as in the processed forms, the biological relevance of particular contacts can be difficult to determine.

Genetic analyses have shown that HIV-1 IN mutants are pleiotropic; some mutations (called Class I mutants) block integration specifically, for example by affecting residues in the active site or residues important for substrate DNA binding. But others (Class II mutants), which can be mapped throughout the IN coding region, affect additional stages of the life cycle including reverse transcription, and virus particle assembly and release (50). A significant, as yet unexplained, effect on capsid morphogenesis is suggested by the fact that viral RNA is excluded from the capsid in particles formed by some HIV-1 IN mutants, and in particles that lack IN protein entirely (51). Some Ty3 IN mutants appear to be similar to these HIV class II mutants, as processing defects that might be linked to assembly were observed (39). Despite the potential complexities, complementation assays in which the reverse transcription defects of certain HIV-1 Class II IN mutants were corrected by inclusion of a Class I mutant as a Vpr-IN fusion in trans, have confirmed the importance of IN-RT interactions at early stages in HIV-1 replication (43, 52, 53). The impact of such interactions for other retroviruses remains obscure.

First Steps of IN Catalysis

Retroviral IN proteins catalyze two sequential reactions, which are separable both biochemically and temporally (Fig. 7, left). In the first reaction, IN removes nucleotides (usually two) that follow the conserved dinucleotide CA at the 3′-ends of the double-stranded viral DNA termini produced by RT. Analysis of cells infected with MLV, showed that “processed,” recessed viral DNA ends can be detected in the cytoplasm shortly after blunt ends are formed at viral DNA termini (54). The second step catalyzed by IN, concerted joining of the two processed CAOH-3′-ends of viral DNA to host DNA, can be detected by providing a target DNA (e.g. from bacteriophage lambda) to PICs isolated from the cytoplasm of both MLV and HIV-1 infected cells. Because this reaction was reported to be inefficient with ASLV PICs, unless deoxyribonucleotide triphosphates are added, it is possible that for some retroviruses synthesis of blunt viral DNA ends, and processing by IN, may not be completed until the PIC enters the nucleus (55).

The activity and function of retroviral IN proteins may also be affected by post-translational modifications (e.g. ubiquitination, SUMOylation, acetylation, phosphorylation) that can occur in the cytoplasm or nuclei of infected cells. The pathways for the placement and removal of such modifications in cellular proteins are complex. The interplay among them, and the net downstream
effects, regulate critical events in cell biology including gene expression, DNA repair, and the cell cycle. While a variety of potential of post-translational modification sites can be identified at various locations in retroviral IN proteins, studies of the potential effects of such modifications have focused primarily on HIV-1 IN. This protein is reported to be modifiable by ubiquitination of residues in the CTD (Fig. 8) and the preceding linker (K211, 215, 219, 273), acetylation in the CTD (K258, 264, 266, 273), SUMOylation in all three domains (K46, 136, 244), and by phosphorylation of a serine in the central, catalytic core domain (S57) (reviewed in 56). Such modifications could affect not only IN stability or catalytic activity, but also interactions with other proteins at various stages of the retroviral life cycle.

HIV-1 IN is subject to degradation by the ubiquitin-proteosome pathway, and it has been suggested that proteasome degradation following interaction of IN with the von Hippel-Lindau binding protein (VBP1) and ubiquitination by the Cul2/VHL ligase, may be required for cellular components of post-integration repair and gene transcription to gain access to the provirus after its integration (57). Two histone acetyl transferases, p300 and GCN5, have been shown to bind to HIV-1 IN, in vitro and in infected cells, and acetylation of lysines in the CTD is reported to enhance integration by increasing DNA binding affinity in vitro (58, 59). In another study, HIV-1 SUMOylation site mutants were shown to be defective in a step after reverse transcription, but before integration (60), and it was suggested

FIGURE 7 Reactions catalyzed by retroviral integrases. Left: Reactions at the retroviral DNA ends produced by RT in infected cells. The processing reaction takes place in the cytoplasm as soon as DNA synthesis is completed at the termini. Following nuclear entry the two processed viral DNA ends are joined to host DNA in concerted cleavage and ligation reactions at staggered positions in the target site. Repair of the resulting gaps is catalyzed by host enzymes. The integrated provirus is flanked by short repeats (indicated by vertical lines) of the host DNA, with length determine by the distance between the staggered cuts made by IN. Right: Reactions as assayed in vitro using duplex oligonucleotides containing viral DNA end sequences and target DNAs. Red stars indicate radioactive or fluorescent labels that can be used for distinguishing reactions (A) and identifying recombinant molecules (B–D) following electrophoresis in denaturing or non-denaturing (D) gels. doi:10.1128/microbiolspec.MDNA3-0005-2014.f7
that this modification may enhance IN binding to cellular proteins that contain SUMO interaction motifs (SIMs), and are required for efficient viral replication. However, unraveling the functions of any of these modifications in the HIV-1 life cycle will require further study.

FIGURE 8 Domain organization of IN proteins from different retroviruses. Maps for the organization of IN proteins are shown with amino acid numbers that delineate the start and end of each domain. The lengths of linkers that connect the domains are also indicated below the lines between domains. The domain models below are from the crystal structures of the HIV-1 N-terminal domain (NTD), catalytic core domain (CCD), and C-terminal domain (CTD), PDB codes 1K6Y, 1BIU, 1EX4, respectively. The Zn$^{2+}$ ion in the NTD is shown as a blue sphere, and the Mg$^{2+}$ ion in the active site of the CCD, as a green sphere. Domains in proteins for which there is no experimentally determined structure from crystallography are shown in muted colors in the maps above. The domain pictures were generated using Chimera software (UCSF) and the figure kindly provided by Dr. M.D. Andrake, Fox Chase Cancer Center, Philadelphia, PA. doi:10.1128/microbiolspec.MDNA3-0005-2014.f8
Nuclear Entry
In non-dividing cells, the nuclear membrane separates the retroviral PIC from its host DNA target. For MLV, this is a nearly impenetrable barrier. Integration of MLV DNA is only efficient in cycling cells, in which the nuclear membrane breaks down during mitosis and re-assembles before G1. Tethering to mitotic chromosomes by the MLV gag-encoded protein p12 in the PIC, allows its selective accumulation in newly formed nuclei (61–63). In contrast, HIV-1 and ASLV integration can occur in non-cycling cells, although the process is more efficient with the former (64–66). Indeed for HIV-1, nuclear transport entry is likely to be the major mechanism of PIC access to target host DNA, even in dividing cells. Considerable effort has been expended in numerous laboratories to identify the viral and cellular components that are critical for nuclear entry of the HIV-1 PIC. Results have been contradictory and sometimes controversial, perhaps because there are compensating modes of transport into the nucleus for this virus. Nevertheless, the current consensus is that interaction of HIV-1 CA, and perhaps other viral components, with nuclear pore proteins help to position the PIC for transport (reviewed in 6); (see also Chapter 45, Craigie & Bushman). As the cytoplasmic HIV-1 PIC exceeds the size of the nuclear pore, one attractive model is that, as with herpesviruses and adenoviruses (67), attachment to the pore facilitates (partial) disassembly of the complex allowing transport of a streamlined PIC into the nuclear compartment. A transferable nuclear localization signal has been identified in ASLV IN, and studies with permeabilized cells indicate that nuclear import of this protein relies on one or more of the cellular components that mediate transport of the linker histone H1 (68). Whether IN is the major viral determinant of nuclear transport for the ASLV PIC remains to be resolved. Nevertheless, while it is clear that ASLV DNA can be integrated in non-dividing cells, it is often misstated in the literature that the lentiviruses are “unique” in this property. In fact, very little is known about the manner in which PICs other than those of ASLV, MLV, and HIV-1, gain access to their cellular DNA targets. Moreover, as early steps in retrovirus replication can be restricted in non-dividing cells by factors other than nuclear access (reviewed in 69), cell-cycle dependence cannot be used as a surrogate criterion for transit through the nuclear pore.

THE MECHANICS OF INTEGRATION
Molecular cloning and characterization of MLV and ASLV proviruses in the 1980s identified the essential hallmarks of the integration reaction (Fig. 7), and revealed their similarity with DNA transposable elements of bacteria and eukaryotic cells (70–73). Like these transposable elements, proviral DNAs all contain the dinucleotide the CA at their 3′-ends. The finding that these dinucleotides are sub-terminal in unintegrated viral DNA indicated that two base pairs must be lost from either end upon integration. As with DNA transposable elements, the ends of the provirus are flanked by direct repeats of host DNA that were formed during integration. These duplications are characteristic of the virus (i.e. 4-6 base pairs). The genesis of these features became clear in subsequent analyses of the viral DNAs formed in infected cells, and of the activities of purified IN proteins.

Target Site Selection
Retroviral DNA arrives in the nuclear compartment with processed ends tightly bound to IN and poised for joining to a target DNA. This IN-viral DNA complex has been called an “intasome.” While integration can occur at many sites in host cell DNA, it is now clear that site selection is not random. LTR retrotransposons are known to be integrated at very specific target sites. For example, the Ty1 and Ty3 retrotransposons of the budding yeast Saccharomyces cerevisiae are inserted preferentially upstream of genes transcribed by RNA polymerase III (e.g. tRNA and 5S ribosomal genes), and the Ty5 retrotransposon into regions of heterochromatin at telomeres and the silent mating-type loci. The availability of sequence data for human, mouse, and avian genomes, and use of next generation sequencing methods have now made it possible to determine integration site preferences for representatives of most retroviral genera, which also vary in characteristic ways (74). The gammaretroviruses and spumaretroviruses tested all show a marked preference for integration in, or near, transcription start sites and CpG islands, which are enriched in the promoter regions of highly-expressed genes. The lentiviruses show strong preference for integration within active transcription units (but not start sites or CpG islands), and the alpharetroviruses and deltaretroviruses exhibit only weak preference for transcription units and CpG islands (but not transcription start sites). Finally, the betaretroviruses tested show no significant deviation from random insertion in integration site preference. These differences may be explained, in part, by dissimilar requirements for chromatin structure or accessibility, and/or by IN-specific cellular protein attachments that tether the intasome to chromatin. For HIV-1, interaction of the chromatin-targeting, lens
epithelium-derived growth factor (LEDGF/p75) with the catalytic core dimer interface of IN is clearly a major determinant of integration site selection (for review see 75). Recent studies with gammaretroviruses MLV and FeLV have identified bromodomain and extraterminal domain (BET) proteins (Brd 2, -3, -4) as virus-specific cellular tethers. In these cases, the conserved ET domains bind to a region of IN that includes portions of the catalytic core and C-terminal domains, and direct the intasomes to transcription start sites (76, 77); (see Chapter 45 Craigie & Bushman). This interaction may also affect IN structure in some beneficial way, as the binding of an isolated ET domain was shown to increase the catalytic activity of MLV IN in vitro, and its overexpression in cultured cells stimulated MLV integration.

Comparison of the integration sites of MLV and two other gammaretrovirus provirus in various cell types showed strong associations with binding sites for a particular transcription factor (STAT1) and specific covalent modifications of histone H3, consistent with the idea that chromosomai features per se, may also affect integration site selection (78). Recent in vitro studies with chromatized target DNA indicate that nucleosome density or stability can influence target selection differently for ASLV, MLV, HIV-1, and PFV intasomes in the absence of tethering (79). DNA sequence can also influence integration site selection. Weak, but characteristic palindromic consensus sequences have been identified at the integration target sites for members of the different retroviral families (for review see 74). Moreover, the preference for a specific consensus sequence appears to be distinct from IN tethering, as HIV-1 infection of cells that lack LEDGF show the same palindromic preference as cells that posses the targeting protein, despite the fact that that transcription units are no longer preferred and integration efficiency is reduced substantially in LEDGF deficient cells (80). These findings suggest that nucleosome context and sequence preferences are dictated by distinct features of the intasomes.

IN Domain Composition and Multimerization
Retroviral IN proteins contain three characteristic structural domains (Fig. 8): an N-terminal domain (NTD) that includes a helix-turn-helix fold, which is characterized by conserved HHCC motif that binds a single Zn ion; a larger catalytic core domain (CCD or core) that includes a conserved D, D(35)E motif, which chelates the two divalent metal ions (Mg$^{2+}$ or Mn$^{2+}$) required for catalysis; and a C-terminal domain (CTD) with an SH3-like fold. However, as with RT, some IN proteins (e.g. MLV and the spumavirus prototype foamy virus, PFV) include additional domains. Furthermore, linkers of different length and sequence composition separate the three major IN domains in proteins from different retroviruses. These distinct features and variations in the non-conserved residues contribute to the properties that distinguish IN proteins from different retroviruses. For example, the bacterially-produced IN proteins from particular retroviruses vary in their specific activities and solubility (81). The proteins can also differ in multimerization properties (82, 83). Under similar conditions at concentration of ~2 mg/ml (~ 60 μM) ASLV IN is a dimer, HIV-1 IN is a tetramer, while PFV IN is a monomer even at twice this concentration (84). Assuming that no other macromolecules are present, a conservative calculation for the intracapsid concentration of ASLV IN in a virus particle is ~150 μM (82), a concentration at which a dimer should predominate. However, until more is known about the internal organization of retroviral capsids, RTCs, and PICs, the biological significance of differences in the solution properties of retroviral IN proteins is impossible to gauge.

Biochemistry of Integration
Despite uncertainties regarding PIC composition and organization, much crucial information has been garnered from in vitro studies with reconstructed systems comprising purified IN proteins, cognate viral DNA substrates, and model targets. An early, major breakthrough in these efforts was the development of simple assays that used short duplex oligonucleotides to monitor both the processing of viral DNA end sequences and their joining to a target (85) (Fig. 7, right A-C). Biochemical, genetic, and complementation studies subsequently revealed that a single active site catalyzes both steps, and that IN proteins of ASLV and HIV-1 function as multimers. Later investigations, and the development of simple concerted integration assays (Fig. 7 right D), identified a tetramer as the minimal functional unit for ASLV and HIV-1 IN (86, 87). The use of model target DNAs assembled in nucleosome arrays revealed differences in the tolerance of the ASLV and HIV-1 integration reactions to chromatin compaction by histone H1 (88). More recent use of a similar target has suggested a role for cellular chromatin remodeling proteins (SWI/SNF) in HIV-1 integration (89).

Determination of the crystal structures of the isolated CCDs of ASV IN and HIV-1 IN (90, 91) revealed superimposable architectures, and their relationship to other polynucleotidyl transferases (Fig. 8). Subsequent
solution of the isolated flanking terminal domain structures, and of two-domain IN fragments, provided important insights into IN function. However, the more recent solution of crystal structures of a tetramer of PFV IN in complex with viral DNA end sequences, and with a target DNA fragment, were major advances in the field (92–94) (see also Chapter ___ Engelman & Cherapanov). The PFV structures not only provided valuable templates from which to model IN-DNA complexes of other retroviruses (95), but also revealed the molecular details of how active site inhibitors block integration. In addition to being extraordinarily illuminating, the PFV structures also raised some interesting new questions. For example, although an IN tetramer is required for integration, only two monomers, comprising an “inner” dimer (Fig. 9A), appear to participate in catalysis. As predicted from earlier studies (96) there is a reciprocal (trans) arrangement in the inner dimer of the crystal structure: Each viral DNA end is bound by the two terminal domains of one monomer (with tips frayed (97, 98) via interaction of the 5′-ends with the CTD), but processed in the CCD of the second monomer. Removal of the two nucleotides at the 3′-ends in the processing reaction allows binding of a bent target DNA, and the concerted joining of the 3′-ends of viral DNA to both strands of the target, five base pairs apart. Two other monomers are bound to either end of the inner dimer by CCD-CCD interactions, which define the “outer” dimer interfaces, but only the CCDs of these outer monomers were resolved in the crystals and these subunits do not appear to contribute to the reaction in any way. Subsequent studies verified that the unexpected arrangement

![FIGURE 9 Solution models for the PFV IN tetramer in an intasome and the ASV IN reaching dimer. A. The PFV IN tetramer in an intasome, with DNA omitted to draw attention to IN subunit organization of the inner dimer. Color codes for IN domains in the inner dimer are as in Figure 8, with one subunit in dim pastels for ease of distinction. All three domains of the outer monomers are in yellow. Coordinates for the PFV intasome solution structure were kindly provided by Dr. Kushol Gupta. B. The left side shows the solution structure of the ASV IN reaching dimer structure in the absence of DNA, color coded as in the PFV IN inner dimer. In this structure the CTDs interact with each other, in a “closed” configurations. The right side shows a hypothetical “open” configuration of the reaching dimer formed by rotation of the domain linkers. This open configuration resembles that of the PFV inner IN dimer and could bind viral DNA ends in a similar manner. Pictures were generated using Chimera software (UCSF) and the figure kindly provided by Drs. R. Bojja and M.D. Andrake, Fox Chase Cancer Center, Philadelphia, PA. Structural details are found in (99, 100). doi:10.1128/microbiolspec.MDNA3-0005-2014.f9]
of the PFV IN tetramer in the crystal structure is not due to crystal packing constraints. A combination of biophysical approaches, including small angle X-ray and neutron scattering (SAXS/SANS), revealed a similar structure for the PFV IN-DNA complex in solution (99). In the solution structure, the NTDs and CTDs domains of the outer monomers are extended outwards on either side of the complex (Fig. 9A). It has been speculated that these outer subunits may contribute to tetramer stability, or that they may be required for appropriate assembly of the tetramer in the presence of DNA. Further study will be required to answer these interesting questions for PFV and other retroviral IN proteins.

Biophysical studies of the ASLV IN dimer in the absence of DNA, which employed a combination of SAXS, chemical crosslinking and mass spectrometry, revealed another unexpected structure, named a “reaching dimer” (Fig. 9B). As with the inner dimer of the PFV IN-DNA complex, the CCDs of the ASLV IN reaching dimer lie at opposite poles, and the structure is stabilized by interactions of the NTD of one monomer with CCD and CTD of the second monomer (100). However, in contrast to the PFV IN inner dimer, the ASLV IN reaching dimer is also stabilized by CTD-CTD interactions, which must disassociate to bind viral DNA ends. A similar reaching dimer was identified for HIV-1 IN, but with this protein, dimers that are stabilized by CCD-CCD interactions were also detected (Fig. 10) (84).

The latter finding is consistent with preliminary calculations for free energy of formation, which predict similar stabilities for the HIV-1 IN reaching dimer and CCD-CCD interfaces. Based on SAXS data, a model was proposed for the HIV-1 IN tetramer in the absence of DNA, in which the CCD-CCD interfaces lie at opposite poles (Fig. 10). As with the reaching dimers there is no way to accommodate viral DNA in this tetramer, and major conformational changes would be required to form an intasome. Biochemical studies indicate that HIV-1 IN multimer dissociation is required for function (101), and that the HIV-1-DNA complex may also be formed by dissociated monomers (102), as is the case for the PFV intasome. However, given the wide variation in self-association properties among IN proteins, and the possible contributions of other components in RTCs and PICs, it is not yet possible to delineate the mode of intasome assembly for either the HIV-1 or ASLV IN.

Post-Integration Events
Retroviral DNA integration produces a double strand break in the genome of its target cell, in which host DNA ends are held together by single strand attachments to the viral DNA (Fig 7). Interruption of existing chromatin conformation and composition by the insertion of a large stretch of newly synthesized viral DNA comprises a major assault on the genomic integrity of the cell. While the 3′-end of viral DNA are linked covalently to host DNA substrates were derived by HADDOCK data-driven modeling of the HIV-1 IN monomer, dimer, and tetramer in solution, based on Small Angle X-Ray Scattering and protein cross-linking data (84). It is not yet known which of these multimeric forms are competent for viral DNA binding in the formation of an HIV-1 intasome. Figures were generated using Chimera software (UCSF) and kindly provided by Drs. R. Bojja and M.D. Andrake, Fox Chase Cancer Center, Philadelphia, PA. doi:10.1128/microbiolspec.MDNA3-0005-2014.f10
DNA, the 5′-ends are free and adjacent to a short stretches of single-stranded host DNA. In MLV-infected cells, repair of this gap and covalent joining of the 5′-ends of the integrated viral DNA to host DNA could be detected within an hour after 3′-end attachment (103). Subsequent studies with MLV, ASLV, and HIV-1 infected cells have indicated that such post-integration repair requires signaling through the two major DNA damage sensing kinases ATM, ATR, and the function of components in the non-homologous-end joining (NHEJ) pathway (24, 104–106). As it has been reported that the cellular Ku protein, a component of the NHEJ complex, is bound to the HIV-1 PIC (107), and that the repair protein Rad18 interacts with HIV-1 IN (108), it is possible that binding of such proteins to PIC components may facilitate post-integration repair for some viruses. This idea is supported by suppressive effects of the Rad52 protein, a component of the homologous recombination repair pathway, on HIV-1 DNA integration (109). Decrease in the amount of Rad52 protein enhanced HIV-1 DNA integration, and its overexpression reduced integration, but other components of this repair pathway had no effect. From these and other observations, it was proposed that the RAD52 protein may bind to viral DNA ends in the PIC in a way that blocks loading or recruitment of Ku or other proteins that are required for efficient reverse transcription or integration.

A selected screen with siRNAs for DNA repair pathway genes, revealed that knockdown of several enzymes in the short patch, base excision repair (BER) pathway reduces the efficiency of HIV-1 infection (110). Knockdown of other DNA repair proteins or NHEJ components had no effect in these assays, but a negative result is inconclusive because small amounts of proteins that remain after knockdown may be sufficient for function. Subsequent studies with murine cells carrying deletions of specific genes in the BER pathway (Ogg1, Myh, Nei1 and Pol6) showed that these proteins are required for efficient integration of the lentiviruses HIV-1 and feline immunodeficiency virus, FIV, but not the gammaretrovirus MLV (111). Exactly how these proteins affect integration efficiency, and the reason for the observed differences among retroviruses, remain to be elucidated.

Finally, the expression of proviral genes depends on integration in numerous ways. The local chromatin environment may have an influence by promoting (euchromatin) or silencing (heterochromatin) proviral gene expression, and it is generally believed that features of neighboring chromatin can “spread” into the integrated retroviral DNA along with the acquisition of histones. Alternatively, or in addition, autonomous chromatin domains may be established on the provirus, either rapidly or over many cell divisions (112, 113). One example of an autonomous mechanism is silencing of the MLV provirus in mouse stem cells. In this case the proviral PBS acts as a cis-acting sequence to promote silencing by binding a stem cell–specific complex, comprising a DNA-binding Zn-finger protein (ZFP809), the co-repressor TRIM28, and the ErbB3-binding protein 1 (EBP1) (114, 115). Another example is the human Daxx protein, identified as an interactor with ASLV IN. Daxx binds to the ASLV IN in the PIC and recruits repressive histone deacetylase (HDAC1) and DNA methyl transferases to proviruses in mammalian cells (116). A similar role for Daxx has been suggested for HIV-1 (117). Deposition of histone H3, and repressive histone marks, can be detected on the ASLV provirus as early as 12 hours post-infection, suggesting that its chromatization occurs quite rapidly. Epigenetic repression is also prominent in the establishment of HIV-1 latency, and evidence suggests that transcription factor binding and DNA methylation may cooperate to maintain HIV latency (118, 119). As the establishment and maintenance of a silent provirus may be advantageous under some conditions, mechanisms that promote this response are likely to be widespread in the different retroviral genera.

POSTSCRIPT

Early studies of retroviral DNA transposition centered on the avian and murine retroviruses viruses following infection of chicken and mouse cells. These were valuable experimental systems that established some of the main features of the retroviral life cycle. Furthermore, as these viruses were known to cause cancer in their host species, biochemical and genetic studies with these viruses did much to move that field along. The focus of retrovirus research changed dramatically in the 1980’s with the identification of retroviral species that infect humans, particularly the AIDS virus HIV-1. Because of the critical need to understand as much as possible about its replication and pathogenesis, a substantial increase in financial resources, and the continuing pressure to develop anti-viral drugs and a vaccine, the vast majority of retrovirologists today are HIV-1 virologists. Yet a world of potential variety and nuance is suggested by consideration of the relatively little we know about some of the other, non-lentiviral genera, and by consideration of the striking contrasts in lifestyle among others: from the epsilonretrovirus WDSV, which must promote on-ogenesis to propagate (120), to the unconventional spumaviruses, in which reverse-transcription begins before
Retroviral DNA Transposition: Themes and Variations

virus particle release (12). Further elucidation of the similarities and variations by which different retroviruses accomplish transposition will surely enhance our knowledge of both the viruses and the cells that they infect.

ACKNOWLEDGMENTS

It is a pleasure to acknowledge the expertise and generosity of Drs. Vladimir Belyi, Stuart Le Grice, and Fox Chase colleagues Dr. Mark Andrake, and Ravi Shankar Bojja, in compiling and analyzing information, and preparing several figures. Ms. Karen Trush, Fox Chase Cancer Center Special Services, also provided valuable assistance with several illustrations. Drs. Le Grice, Andrake, and Richard Katz read early drafts of the text and offered many helpful suggestions. The assistance of Ms. Marie Estes in preparation of the manuscript and bibliography is also most gratefully acknowledged.

This work was supported by National Institutes of Health grants AI40385, CA71515, CA006927, and the WW Smith Charitable Trust.

REFERENCES


Retroviral DNA Transposition: Themes and Variations


