Cell, Vol. 115, 135–138, October 17, 2003, Copyright ©2003 by Cell Press

Targeting Survival: Integration Site Selection by Retroviruses and LTR-Retrotransposons

Minireview

Frederic D. Bushman*
Infectious Disease Laboratory
The Salk Institute
10010 North Torrey Pines Rd.
La Jolla, California 92037

Replication of retroviruses and retrotransposons depends on selecting a favorable chromosomal site for integration of their genomic DNA. Different retroelements meet this challenge by targeting distinctive chromosomal regions. Despite these differences, recent data hints at a common targeting mechanism—tethering of integration complexes to proteins bound at favorable sites.

Retroviruses are distinguished from other viruses by two characteristic steps in the viral life cycle—reverse transcription, which results in the formation of a double-stranded DNA copy of the viral RNA genome, and integration, which results in the covalent attachment of the viral DNA to host cell DNA (for reviews, see Bushman, 2001 and Coffin et al., 1997). The choice of integration target sites has a decisive influence on retroviral growth. Integration is not sequence specific, so many chromosomal sites can host integration events. This creates a hazard of integration at an unfavorable location. For example, integration at a site unsuitable for high-level transcription may obstruct production of progeny virions (Jordan et al., 2003). Pressures on target site selection are even more extreme for the related retrotransposons of yeasts (Ty elements in *Saccharomyces cerevisiae* and Tf elements in *Schizosaccharomyes pombe*). These elements replicate by cycles of transcription, reverse transcription, and integration—as with retroviruses—but all within a single cell. The yeast retrotransposons must integrate their DNA into a densely packed genome without suicidal disruption of a gene necessary for survival of the host cell. A series of recent reports discloses the extent to which these elements have evolved sophisticated targeting specificities that befit their replication styles.

Studies of integration targeting are also topical due to recent setbacks in human gene therapy. Retroviral vectors are commonly used for delivery of therapeutic sequences in patients. However, integration of retroviral sequences near protooncogenes has long been known to be capable of activating their expression, contributing to tumorigenesis in animal models (Coffin et al., 1997). Unfortunately, insertional activation has also recently been seen in two patients undergoing retrovirus-based gene therapy (Check, 2002), focusing further attention on integration site selection.

This review first covers elegant studies of integration targeting by the yeast retrotransposons, then recent genome-wide surveys of integration by retroviruses in the human genome. Unexpectedly, the studies of integration targeting by retroviruses now show possible parallels with the yeast elements. The review ends by considering some of the selective forces acting on integrating systems that may explain their observed targeting strategies.

Yeast Models for Integration Targeting

The enzymes involved in reverse transcription (RT) and integration (IN) are similar between yeast retrotransposons and retroviruses, and mechanistic studies in vitro show many parallels. The structures of the termini of retrovirus and yeast retrotransposon genomes are also similar, consisting of long terminal repeats (LTRs) containing sequences important for transcription, reverse transcription, and integration. The yeast elements differ from retroviruses in that they lack an envelope gene and so lack an extracellular phase in their replication—all steps take place inside a single cell, making the Ty and Tf elements retrotransposons instead of retroviruses.

This life-style poses special problems. The *Schizosaccharomyes pombe* genome is 60% gene-coding regions and the *Saccharomyces cerevisiae* genome is fully 70% coding. Thus, an element that integrates randomly into its host cell is at risk of committing suicide by insertional inactivation of host genes. Furthermore, yeasts spend part of their life in the haploid state, so during this phase any damaged gene would be present in only one copy.

Probably for this reason, integration by the yeast retrotransposons is tightly targeted (Sandmeyer, 2003). The Ty1 and Ty3 elements integrate upstream of PolIII transcribed genes, regions of the genome that can tolerate insertions without adverse effect. The Ty3 element has particularly impressive selectivity, integrating only in DNA encoding the 5′ end of PolIII transcripts. Ty1, in contrast, integrates within a window of about 750 bp upstream of PolIII transcribed genes (Boeke and Devine, 1998). For the case of Ty3, the mechanism of integration appears to involve local tethering of the integration complex to the TFIIIB component of the PolIII transcription apparatus (Kirchner et al., 1995). The mechanism of targeting by Ty1 is not yet clarified.

The *Schizosaccharomyes pombe* Tf1 element shows a different targeting strategy, integrating preferentially upstream of PolII transcribed genes (Singleton and Levin, 2002). Though different from the Tys, the Tf1 strategy achieves the same result—integration into a benign genomic region.

The Ty5 element displays still another integration specificity, with 95% of new insertions taking place in heterochromatin at telomeres or the silent mating loci (*HML* and *HMR*). These sites also seem to tolerate insertions without adverse effects on the host cell. A striking series of recent experiments from Voytas and coworkers have suggested a detailed mechanism, also based on tethering, to explain Ty5 targeting. In cells lacking the heterochromatin protein silent information regulator 4 (Sir4p), Ty5 integration was no longer targeted, implicating Sir4p in specifying integration sites (Zhu et al., 1999). Binding assays showed that the IN enzyme encoded by Ty5 bound Sir4p, and this finding allowed mapping of the binding determinants in each. The Ty5 IN targeting domain could be mapped to a 6 amino acid sequence

*Correspondence: bushman@salk.edu

near the IN C terminus and the Sir4p binding domain was mapped to the carboxyl-terminal half of the protein.

An incisive series of experiments followed, demonstrating that the observed binding accounted for targeted integration (Zhu et al., 2003). The Sir4p carboxyl terminus was fused to the DNA binding domain of lexA, a bacterial repressor. Cells harboring the fusion protein and artificially introduced lexA binding sites showed a new hotspot for Ty5 integration at the lexA sites. Targeting was not all-or-nothing—only 14% of new integration events were at the lexA sites—but the experiment did strongly support the tethering mechanism.

These results set the stage for targeting integration by entirely new protein-protein interactions. The short targeting domain in Ty5 IN was replaced with two other short peptides that bind tightly to known proteins, either a 13 amino-acid motif from Rad9 that mediates binding to the two forkhead-associated (FHA) domains of Rad53p, or a 12 amino acid proline-rich motif from human NpwBP that binds the WW domain of Npw38. The ligand binding regions of Rad53p and Npw38 were each fused to lexA and targeting by Ty5 assessed. Impressively, new integration hotspots (10%–15% of events) were seen at lexA operators when the appropriate binding partners were present. These beautiful experiments provide the strongest evidence yet that a tethering mechanism can account for retroelement targeting in vivo.

Genome-Wide Studies of Retroviral Integration

Genome-wide surveys of integration targeting by retroviruses (Schroder et al., 2002; Wu et al., 2003) have hinted that there may be more parallels between retroviral and yeast retrotransposon integration than previously appreciated. The methods used in the two large-scale surveys reported so far were quite similar. Both studies used mostly retroviral vectors instead of retroviruses, since vectors are convenient to handle and model gene transfer during gene therapy. Retroviral vectors use the viral RT and IN enzymes to carry out early replication steps, and so are expected to show the same integration targeting as authentic viruses.

In the first of these studies, 524 sites of integration by HIV-1 or an HIV-based vector were cloned and analyzed in human SupT1 cells, a T cell line modeling the main cellular target of HIV infection in vivo. The distribution of these sites relative to chromosomal features suggests that 69% of integration events were in transcription units. At present, about one-third of the human genome appears to be transcribed—thus HIV strongly favors integration into transcription units. No obvious bias was seen for the positions of HIV integration sites within the transcription units themselves.

Transcriptional profiling of the SupT1 target cells revealed that the genes targeted for integration were more active than the average of all genes analyzed on the chip. Evidently not just genes, but active genes, are favored targets. Further transcriptional profiling studies showed that the genes that hosted integration events were particularly active after infection with the HIV-based vector.

A more recent paper from Burgess and coworkers compared integration targeting by HIV and murine leukemia virus (MLV) in the human genome (Wu et al., 2003). For HIV, they characterized 379 sites made by integration of an HIV vector in Hela cells (cervical epithelial cells) or HIV-1 in H9 cells (a T cell line). In both cases, integration was found to be favored in transcription units. Comparison with transcriptional profiling data again supported the idea that active genes were favored.

For MLV, analysis of a whopping 903 sites of integration in Hela cells yielded a strikingly different result. Only 34% of integration events were in transcription units, significantly less than for HIV. The most surprising finding came when Burgess and coworkers checked whether promoters were favored for integration. To approximate the positions of promoters, they examined a window to either side of the start points of transcription, which are now mapped for many genes. They found that fully 16.8% of integration events took place in a two kilobase window centered on the transcription start, but the positive effect was quite localized, dropping off by about 5 kb in either direction. Integration frequency was also assessed at CpG islands, which are commonly associated with promoters in humans, revealing that 17% of integration events were within 1 kb of CpG islands, eight times higher than for random sites. A similar study of HIV sites showed no such bias in favor of promoters.

While the detailed mechanism of retroviral integration targeting is unknown, the above studies can be comfortably accommodated in tethering models, as with the yeast retrotransposons. An early model for retroviral targeting suggested that open chromatin near active genes was favored, explaining favored integration at DNase I hypersensitive sites (Coffin et al., 1997). Differential access to targets may well contribute to integration specificity, but such a model alone cannot account for different target preferences by MLV and HIV in the same cells. Instead, some form of the tethering model seems attractive. In one version of such a model, MLV might bind to transcription factors or modified histones bound at or near the 5′ ends of genes and so carry out integration locally. HIV might similarly interact with positive factors bound within transcription units.

In support of such models, studies in vitro have documented that artificial tethering can support selective targeting by engineered retroviral IN enzymes. In these experiments, retroviral IN proteins were fused to sequence-specific DNA binding domains, and these fusions were shown to direct preferential integration in vitro at target sites containing the appropriate DNA recognition sequences (Bushman, 2002 and references therein).

While the differences between MLV and HIV seem to point to a tethering model, additional mechanisms may well contribute to integration site selection. For example, variation in the intranuclear position of whole chromosomes may affect which chromosomes are favored targets in different cell types. Another possibility is suggested by studies of integration at a model gene in avian cells, where it was found that very high level transcription actually disfavored integration by an avian retrovirus (ALV) (Weidhaas et al., 2000). It is not known whether integration by all retroviruses is disfavored by very high levels of transcription, or whether ALV is unique in this respect. The primary sequence of the target DNA can also have a detectable influence on target site selection, though the effect is weak and probably not a major contributor in vivo (Coffin et al., 1997).

None of these models are mutually exclusive, and it

Table 1. Target Specificities of Some Integrating Elements

Element	Element type	Host Organism	Salient Features of Integration Targeting	Reference
Ty1	LTR retrotransposon	*Saccharomyces cerevisiae*	750 bp window upstream of Pol III genes	Boeke and Devine (1998)
Ty3	LTR retrotransposon	*Saccharomyces cerevisiae*	Transcription start of Pol III genes	Kirchner et al. (1995)
Ty5	LTR retrotransposon	*Saccharomyces cerevisiae*	Heterochromatin at telomeres and silent mating type loci	Zhu et al. (1999)
Tf1	LTR retrotransposon	*Schizosaccharomyes pombe*	Upstream regions of Pol II transcribed genes	Singleton and Levine (2002)
HIV-1	Lentivirus	*Homo sapiens*	Active genes	Schroder et al. (2002)
MLV	Retrovirus	*Mus musculus*	Promoters of active genes	Wu et al. (2003)
LINE	Non-LTR retrotransposon	*Homo sapiens, others*	Targeted widely, rearranges integration site	Gilbert et al., Symer et al. (2002)
Sleeping Beauty	DNA transposon	*Homo sapiens, others*	Targeted widely, avoids Alu repeats	Vigdal et al. (2002)
Adeno-associated virus	Parvovirus	*Homo sapiens*	Active genes	Nakai et al. (2003)b

seems likely that more than one mechanism will contribute. If tethering is indeed involved in targeting retroviral integration, then the field is faced with a new challenge—identifying the chromosomal ligands for the retroviral integration machinery. Several cellular DNA binding proteins have been described that bind integration complexes and/or facilitate integration in vitro, including BAF, HMGa1, Ini-1, Ku, and LEDGF (Sandmeyer, 2003; Coffin et al., 1997; Bushman, 2001). These proteins are candidates for binding partners influencing target site selection as well.

Targeting to Survive

The integration targeting strategies of retroelements are well-suited to promoting their evolutionary persistence (Table 1). The yeast LTR-retrotransposons inhabit a very gene-dense host and so target integration outside genes. HIV has an opposite strategy, targeting integration to transcription units. Most cells infected by HIV persist only for a couple of days before they are eliminated, either by the cytopathic effect of infection or by immune clearance. Thus, the HIV provirus will maximize the production of progeny virions by producing the largest number possible in the short time available, and high-level transcription may be facilitated by integration in transcriptionally active regions (Jordan et al., 2003). For MLV, there is less information available on the dynamics of infection, making the influence of integration site selection harder to assess. It may well be that integration in promoters also favors active proviral transcription, but in this case it is also possible that the upstream regions are relatively benign targets, suitable for long term residence in the murine genome. The Tf1 retrotransposon appears to have a targeting specificity resembling that of MLV, so studies in the Tf1 system may help clarify the MLV targeting mechanism.

Another set of forces appears to account for the chromosomal positions of human endogenous retroviruses (HERVs), evolutionarily ancient insertions in the primate lineage, which account for about 8% of the human genome sequence. The HERVs accumulated outside of genes, the opposite of HIV (Smit, 1999). The integration targeting specificities for the HERVs are unknown, so they might have favored initial integration in intergenic regions. However, in this case, it seems likely that selection after integration played an important role. The minority of HERV sequences within genes are oriented opposite to the direction of host gene transcription fully 80% of the time. The reverse orientation of HERVs means that the element-encoded signals for RNA processing (splicing, cleavage, and polyadenylation) do not disrupt expression of the host gene, rendering the inserted HERV sequences relatively benign. Thus, integrated HERVs have apparently been selected at the cellular level after integration to minimize genetic damage to the host genome. This observation supports the idea that insertional inactivation of genes has been selected against in the lineage leading to modern humans.

Genome-wide surveys as described above for retroviruses have also been carried out for three other elements that integrate in the human genome, illustrating the diversity of targeting strategies and genetic consequences. Long interspersed nuclear elements (LINES) are non-LTR retrotransposons, which comprise fully 20% of our genomes. LINES integrate by nicking the target DNA, then using the nick to prime reverse transcription (reviewed in Kazazian and Goodier, 2002). Genome-wide surveys of LINE integration revealed a high rate of rearrangement of the integration target sequences, emphasizing the likely roles of these elements in genome remodeling over time (Gilbert et al., 2002; Kazazian and Goodier, 2002; Symer et al., 2002). As yet there are too few sites analyzed to assess detailed biases in targeting, though so far there is not strong evidence in favor of integration in or near active genes.

Adeno-associated virus (AAV) is a DNA virus that integrates in human cells at a single location on chromosome 19, but curiously, integration of AAV-based vectors results in more widely distributed integration. AAV-vector integration causes frequent rearrangements at integration sites, as is often seen with integration of naked DNA, suggesting (together with other data) a prominent role for cellular DNA repair pathways in the integration mechanism. The genetic damage accompanying integration is an issue for use of this virus as a vector in gene therapy. AAV was recently found to favor integration in active genes in a liver gene therapy model (Nakai et al., 2003).

The DNA transposon *Sleeping Beauty* showed a distinctive pattern of preferred sites in the human genome. *Sleeping Beauty* is a cut-and-paste DNA transposon that excises its DNA and integrates using a transposase enzyme related to the retroviral INs. *Sleeping Beauty* integrates relatively indiscriminately, but oddly avoids Alu elements. Possibly this is because *Sleeping Beauty* strongly favors a particular sequence at integration target sites, and the recognition sequence may be rare in Alus (Vigdal et al., 2002).

The relationship between integration targeting by these three elements and their evolutionary persistence is only beginning to be explored. Possible contributions of tethering to the targeting mechanisms are as yet uninvestigated. However, genome-wide surveys like those above are beginning to pose the questions for the next round of mechanistic studies.

In summary, studies of yeast retrotransposons strongly support a tethering mechanism for integration site selection, and studies of MLV and HIV are beginning to point to a tethering mechanism as well. Considering the field more broadly, these studies and other genome-wide surveys are revealing much about the selective pressures directing target site selection by integrating elements. Given that eukaryotic genomes are largely composed of sequences derived from integrating parasites (at least 40% in humans), these studies can tell us much about how our own genomes were formed.

Selected Reading

Boeke, J.D., and Devine, S.E. (1998). Cell *93*, 1087–1089.

Bushman, F.D. (2001). Lateral DNA Transfer: Mechanisms and Consequences (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press).

Bushman, F.D. (2002). Mol. Ther. *6*, 570–571.

Check, E. (2002). Nature *420*, 116–118.

Coffin, J.M., Hughes, S.H., and Varmus, H.E. (1997). Retroviruses (Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press).

Gilbert, N., Lutz-Prigge, S., and Moran, J.V. (2002). Cell *110*, 315–327.

Jordan, A., Bisgrove, D., and Verdin, E. (2003). EMBO J. *22*, 1868–1877.

Kazazian, H.H.J., and Goodier, J.L. (2002). Cell *110*, 277–280.

Kirchner, J., Connolly, C.M., and Sandmeyer, S.B. (1995). Science *267*, 1488–1491.

Nakai, H., Montini, E., Fuess, S., Storm, T.A., Grompe, M., and Kay, M.A. (2003). Nat. Genet. *34*, 297–302.

Sandmeyer, S. (2003). Proc. Natl. Acad. Sci. USA *100*, 5586–5588.

Schroder, A., Shinn, P., Chen, H., Berry, C., Ecker, J.R., and Bushman, F.D. (2002). Cell *110*, 521–529.

Singleton, T.L., and Levin, H.L. (2002). Eukaryot. Cell *1*, 44–55.

Smit, A.F. (1999). Curr. Opin. Genet. Dev. *9*, 657–663.

Symer, D.E., Connelly, C., Szak, S.T., Caputo, E.M., Cost, G.J., Parmigiani, G., and Boeke, J.D. (2002). Cell *3*, 327–338.

Vigdal, T.J., Kaufman, C.D., Izsvak, Z., Voytas, D.F., and Ivics, Z. (2002). J. Mol. Biol. *323*, 441–452.

Weidhaas, J.B., Angelichio, E.L., Fenner, S., and Coffin, J.M. (2000). J. Virol. *74*, 8382–8389.

Wu, X., Li, Y., Crise, B., and Burgess, S.M. (2003). Science *300*, 1749–1751.

Zhu, Y., Dai, J., Fuerst, P.G., and Voytas, D.F. (2003). Proc. Natl. Acad. Sci. USA *100*, 5891–5895.

Zhu, Y., Zou, S., Wright, D.A., and Voytas, D.F. (1999). Genes Dev. *13*, 2738–2749.