CMLS, Cell. Mol. Life Sci. 61 (2004) 2588–2596
1420-682X/04/202588-09
DOI 10.1007/s00018-004-4206-9
© Birkhäuser Verlag, Basel, 2004

CMLS Cellular and Molecular Life Sciences

Review

Integration target site selection for retroviruses and transposable elements

X. Wu [a] and S. M. Burgess [b,*]

[a] Laboratory of Molecular Technology, Scientific Application International Inc., National Cancer Institute at Frederick, 915 Tollhouse Ave., Frederick, Maryland 21701 (USA)
[b] Genome Technology Branch, National Human Genome Research Institute, 50 South Drive, Room 5537, Bethesda, Maryland 20892 (USA)

Received 13 May 2004; received after revision 21 June 2004; accepted 6 July 2004

Abstract. When a retrovirus infects a cell, its RNA genome is reverse transcribed into a double-stranded DNA, which is then permanently integrated into the host chromosome. Integration is one of the essential steps in the retroviral life cycle. Many transposable elements also move around and integrate into the host genome as part of their life cycle, some through RNA intermediates and some through 'cut-and-paste' mechanisms. Integration of retroviruses and transposable elements into 'sensitive areas' of the genome can cause irreparable damage. On the other hand, because of their ability to integrate permanently, and the relatively efficient rates of transgenesis, retroviruses and transposable elements are widely used as gene delivery tools in basic research and gene therapy trials. Recent events in gene therapy treatments for X-linked severe combined immunity deficiencies (X-SCID) have highlighted both the promise and some of the risks involved with utilizing retroviruses. Nine of 11 children were successfully treated for X-SCID using a retrovirus carrying the gene mutated in this disease. However, later two of these children developed leukemias because of retroviral integrations in the putative oncogene LMO2 [1]. A third child has also been demonstrated to have an integration in LMO2, but is as of yet nonsymptomatic [2]. It is a bit difficult to explain the high frequency of integrations into the same gene using a random model of retroviral integration, and there has been evidence for decades that retroviral integrations may not be random. But the data were somewhat limited in their power to determine the precise nature of the integration biases. The completion of the human genome sequence coupled with sensitive polymerase chain reaction techniques and an ever-decreasing cost of sequencing has given a powerful new tool to the study of integration site selection. In this review, we describe the findings from several recent global surveys of target site selection by retroviruses and transposable elements, and discuss the possible ramifications of these findings to both mechanisms of action and to the use of these elements as gene therapy vectors.

Key words. MLV; HIV-1; insertion; transposable element; co-factors; genomic bias.

Early in vitro studies on target site selection

Many early studies of target site selection for retroviral integration employed in vitro systems and found that most DNA sequences can act as integration target sites, though some weak biases were observed for primary sequences [3–13]. Other factors that affected integration site selection included nucleosomal structure and DNA binding proteins [3, 6, 7, 14]. Proviral integration favors the DNA major groove, particularly regions that are kinked on nucleosomes. The integrations generally do not insert in the minor groove [6, 7]. DNA binding proteins

* Corresponding author.

can either block integrations by steric hindrance or promote integration by inducing sharp bends into the DNA molecule [3, 14]. Although in vitro studies have helped to solve many of the questions about the biochemical mechanisms, they do not reflect how the cellular environment might influence in vivo target site selection.

Early in vivo studies on target site selection

The mechanism by which retroviruses select their sites of integration has been a topic of study for virologists for several decades. Early studies faced many technical challenges that severely limited the ability of researchers to make precise observations. In most cases, it was necessary to get a clonal population of cells containing a single integration in order to identify the integration junction site. Often the integration sites were cloned from retrovirus-induced tumors or cell lines. Tumor formation itself requires that the retrovirus express viral oncogenes or activate cellular genes with growth advantage and positional effects on gene expression that could introduce bias for integration sites. Even given these shortcomings, analysis of a small number of in vivo integration sites in cultured cells, two groups reported that Moloney murine leukemia virus (MLV) retroviruses preferentially integrate close to DNase I-hypersensitive sites [15, 16]. Similarly, Scherdin et al. reported that transcription active regions and CpG islands are preferred sites for MLV integration [17]. However, the conclusions were drawn from an inaccurate assumption that less than 20% of the genome was transcribed and from crude assays for detecting transcriptionally active regions and CpG islands, such as hybridization to nuclear run-on transcripts and restriction enzyme digestion with enzymes that cut preferentially in CpG-rich islands. A study on human immunodeficiency virus-type 1 (HIV-1) integration [18] found that 4 out of 8 integration sites landed near long interspersed nucleotide element-1 (LINE-1) elements, and the authors concluded that HIV-1 prefers to integrate into or near LINE-1 elements in human genome. A later study based on 29 integration sites led to the conclusion that HIV-1 might prefer to integrate into or near Alu elements [19]. Both conclusions were based on small sample size and no accurate information about LINE-1 or Alu distribution. The small sample number and inaccurate assumptions about the whole genome undermined the conclusion drawn from these early studies.

Many other in vivo studies focused on defined chromosomal sites and used polymerase chain reaction (PCR) assays to survey integration site distribution in that region. The study of avian leucosis virus integration in the Turkey embryo fibroblast (TEF) cells using this method led to the conclusion that most or all regions of the TEF genome are accessible to avian leucosis virus (ALV) retroviral integration [20]. The author also noted that specific sites within certain regions are hot spots for integration compared to other sites in the same region and concluded that local structural features play a major role in integration specificity. Using a similar method, integration sites of ALV into an artificial minigene cassette were analyzed for the effect of transcription activity [21]. Contrary to the conclusions from earlier MLV studies, which suggested that transcriptionally active regions are the favored sites for integration, it was found that increasing transcriptional activity of the minigene cassette resulted in a decrease of integration frequency in the region. Conclusions at the time were that the data from this study were incompatible with previous observations, and assumptions about retroviral integration would need to be reevaluated. However, the primary concern for this type of study is how well the specific regions under investigation represent the whole genome, so the debate continued. More recent evidence suggests that both conclusions could be right.

Recent genome-wide surveys of integration sites

HIV-1 and MLV

Only with the recent completion of the human genome and rapid progress of other genome projects has it become possible to paint global pictures of in vivo target site selection for retroviruses and transposable elements. Additionally, the advancement of cloning technology for retroviral junction sequences has made it possible to get a large number of junctions between retroviral LTR sequences (and other insertional elements), and the adjacent chromosomal DNA. Given even a very small length of this genomic sequence (often as little as 20 bp) a precise location in the human (or other) genome can be determined. The most commonly used methods to clone retroviral integration site junction sequences include linker-mediated PCR and inverse PCR. These methods are sensitive enough to allow cloning of hundreds to thousands of integration sites from large unselected pools of infected cells, allowing for an unbiased study of the integration preference profile.

Using just such a technique, Schroder et al. reported a large scale survey of in vivo integration sites for HIV-1 in the human genome [22]. In this study, 524 integration sites of wildtype HIV-1 or a HIV-based vector were cloned from human SupT1 cells. As a control, in vitro integrations into the naked human DNA were generated using HIV pre-integration complexes (PICs). One hundred eleven such integration sites were cloned and sequenced. These integration sites were mapped on the human genome and analyzed for their relative position to various chromosomal features. The analysis showed that 69% of the integrations in SupT1 cells reside in transcription units, essentially between the start and stop signals of the

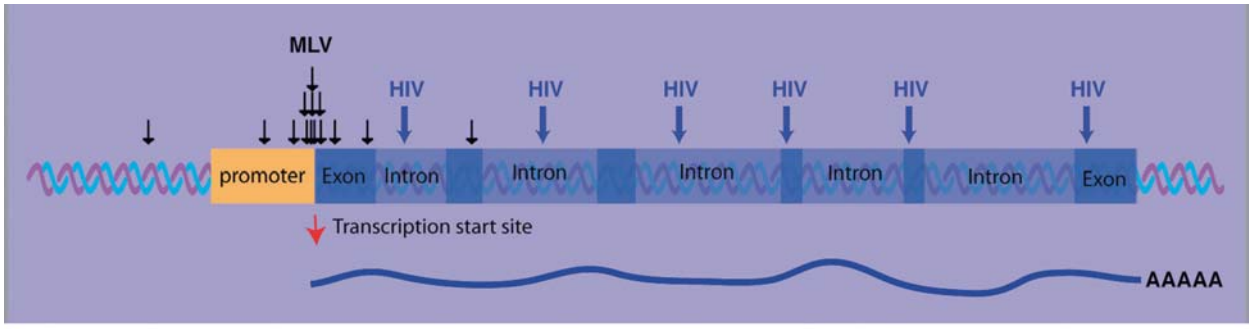

Figure 1. Integration preferences for MLV and HIV-1. MLV prefers to integrate in regions surrounding the transcription start site of genes, and that preference drops off with increasing distance from that location in either direction. HIV-1 has a very strong preference for landing inside genes, but there is no precise location in the gene that is preferred. Integration preference for HIV-1 drops off dramatically once outside the gene boundary.

genes (see fig. 1). It is estimated that about 33% of the human genome is transcribed. Thus their data suggested that HIV-1 is quite non-random in its site selection. It strongly prefers transcribed regions for integration. Further analysis showed that targeted genes are more actively transcribed. In contrast, only 35% of the in vitro integration sites in the naked DNA were in transcription units, suggesting that in vitro integration into naked DNA behaves in an apparently random fashion. This makes it likely there are factors in the cellular milieu that are strongly influencing integration choice.

In a similar study comparing MLV integrations and HIV-1 integrations in human HeLa cells, we mapped 903 MLV integrations and 379 HIV integrations in the human genome [23]. The distribution of HIV-1 integrations in our study was essentially identical to that of Schroder et al. The way we performed our analyses was slightly different in that we limited the definition of gene to only those in RefSeq [24, 25], so our 'gene hit' numbers were lower (58 vs. 69%), but analyzing the data sets in parallel demonstrates that they are indistinguishable at this level of analysis. One notable difference between our set and the Schroder data is that we did not observe any integrations into a hotspot they identified on chromosome 11q13. We used a different cell line in our studies, and perhaps that hotspot was a specific feature of the SupT1 cells or the lack of a hotspot is a feature of HeLa cells. Further data would be needed to determine the cause of the observed difference. In our studies we also showed that ~34% of the MLV integrations were in genes. Although this is significantly different than randomly generated integrations (22%), it is also significantly lower than what we found with HIV-1 (58%). Further analysis of the distribution showed MLV has a strong preference for the region surrounding the transcription start site, while HIV has no such preference (see fig. 1). Corroborating this was the observation that MLV preferred to integrate near CpG islands, which are commonly associated with the transcription start sites of genes. HIV showed no such preference for CpG islands. These results clearly demonstrated that although both MLV and HIV prefer actively transcribed regions, they have clear differences in the fine details of integration site selection. One important outcome of this study was to demonstrate that influences to integration preferences are going to be specific for each retrovirus, and biases cannot be assumed based on the preferences of other viruses. Thus it is possible, even likely, that previous studies that appeared to be contradictory, could in fact both be correct, and it simply reflects actual differences in the particular viruses being studied. Recent experiments mapping the preference of avian sarcoma-leukosis virus (ASLV) emphasizes this point. When this virus is used to infect human cells, the virus demonstrates neither preference for active genes nor preference for the 5′ ends of genes [R. Bushman, personal communication]. It will be interesting to determine whether the cellular co-factors in human cells are incompatible with the viral proteins, or whether ASLV will display the same preferences in avian genomes.

In a search for cancer-causing genes, thousands of MLV integration sites in mouse tumors were cloned and mapped to the genome [26–29]. Hundreds of common integration sites (CISs) from at least two classes of tumor samples were cataloged in the Mouse Retroviral Tagged Cancer Gene Database (RTCGD, http://rtcgd.ncifcrf.gov/) [30]. These integration events represent sites that should cause growth advantages, transforming cells into tumor cells. Thus these integrations represent a selected population and are likely to be highly biased. It is very intriguing to compare the integration profile of this selected set with unselected integration events. Most integrations within the CISs located upstream of a known or putative cancer gene, with a smaller number located downstream of a cancer gene. Surprisingly, the MLV integrations in the CISs, showed strong orientation preferences, which were not observed in the random integration survey. If the integration was upstream of a CIS gene, it was most likely oriented in the opposite direction of the gene. Most of the downstream integrations showed the retrovirus oriented in the same direction as the gene. It is likely that the ori-

entation preference was a result of the tumor selection process. One likely explanation is the orientation of the LTRs may be important for the function of the enhancer/promoter sequences embedded in the retroviral LTRs. These enhancer/promoter sequences are probably involved in misexpressing oncogenes, which would then yield a growth advantage for tumor cells. Certain orientations may be more efficient for this misexpression.

HTLV-1

In another large-scale survey, 218 integration sites were cloned from patients infected with human T-cell leukemia virus type 1 (HTLV-1) [31]. No preference for transcribed regions or non-transcribed regions was observed, compared to a control set of random human sequences. The authors also concluded that HTLV preferred AT-rich regions by comparing the AT content of the junction sequences (57%) with that of the control set (51%). However, the AT content of the junction sequences (57%) described by the authors showed no significant difference from the human genome average (59%), so it is unclear whether there is an actual preference for AT-rich regions. The analysis of data in this study was done prior to the availability of the human genome, and thus the conclusions presented need to be validated and refined given the new expanded genome information.

AAV

Adeno-associated virus (AAV) is a DNA virus that has a biphasic life cycle. In the presence of helper virus such as adenovirus or herpesvirus, AAV enters a productive life cycle. In the absence of helper virus, AAV integrates into the human genome and establishes latency. Wild-type AAV highly prefers a single locus at chromosome 19q13 [32, 33]. This site-specific integration depends on AAV replication (Rep) proteins and a cis-sequence element on chr19 that shares homology with the AAV ITR [34, 35]. This site specificity made AAV an exciting candidate for gene therapy. However, integration of AAV-based vectors (AAVr, recombinant) modified for gene therapy, which lack Rep proteins, showed loss of the site specificity and integration sites distributed widely in the genome. The integration mechanisms for AAV-based vectors are not clear, but cellular DNA repair pathway proteins are suspected to play a major role. A recent survey of 29 integration sites showed that the AAV-based vector preferred to integrate into transcribed regions in the genome [36]. The sites of integration are not as 'clean' as those for retroviruses, with most integrations causing small deletions in the host genome. A much larger survey based on hundreds of integration sites confirms that AAVr prefers genes as target but more interestingly shows that AAVr highly prefers transcription start sites or CpG islands, and a few new hotspots were observed [H. Nakai, personal communication]. Even though the apparent mechanism for integration is quite different than for retroviruses, the preference for actively transcribed regions by AAVr bears a strong resemblance to MLV. This raises the possibility that there are cellular co-factors that are common between the two viruses. We will return to the issue of co-factors later in the review.

Transposable elements

Yeast Ty and Tf elements

Transposable elements or transposons are widely distributed throughout the eukaryotes and prokaryotes. Most transposons in eukaryotic cells are retrotransposons and move via RNA intermediates rather than DNA intermediates. The mechanism of transposition of these elements is indistinguishable from the replication of retroviruses, which have provided the prototype system for studying retroviral integration.

Ty elements of the baker's yeast *Saccharomyces cerevisiae* are some of the best understood of the retrotransposons. Ty1 and Ty3 elements are quite specific in their choice of site for integration [37]. Ty1 prefers the upstream sequences of transfer RNA (tRNA) or other PolIII transcribed genes [38]. Ty3 also target upstream sequences of tRNAs but to a more precise location which is 1–4 bp from the transcription start site [39]. The precise targeting is achieved by interaction between the Ty3 preintegration complex (PIC) and the Pol III transcription factor (TF) IIIB/TFIIIC [40]. Ty5 specifically targets different regions than Ty1 or Ty3, the transcriptionally 'silent' regions of the yeast genome, such as telomeres or the silent mating loci HM. The target site selection by Ty5 is mediated by binding of Ty integrase to transcription silencing protein Sir4p, which binds DNA in the silent regions [41, 42]. The theory behind the strong site preferences for the Ty elements is that by targeting a location in the yeast genome that is 'safe', meaning it has a low probability of killing the cell, the Ty element has a stronger chance of survival over time. Unlike retroviruses, retrotransposons do not exit the cell and infect other cells, but reinfect intracellularly. Thus there are no mechanisms to prevent superinfection of the cells. By only landing in the safer locations, you are insuring the long-term health of the host. Regardless of the reason for the site biases of the Ty elements, the common theme underlying the mechanisms is the binding of the Ty integrases to a host DNA binding protein, which then specifies the location of the integration.

The Tf1 retrotransposon in the fission yeast *Schizzosaccharomyces pombe* has also shown a targeting bias for its transposition. Ninety-one integrations total identified in two different labs [43, 44] have shown a preference for a region 100–400 bp upstream of the translation start site

of PolII genes. No interacting proteins have yet been defined for this targeting of integration.

P elements

The P transposable element in *Drosophila* encodes a protein transposase, which can catalyze the 'hopping' of the transposon DNA in the host genome in a 'cut-and-paste" fashion. Early study showed that P elements insert nonrandomly [45]. Some integrations were found in precise locations in multiple mutants, and euchromatic sites were preferred over heterochromatic sites [46]. The *Drosophila* genome project has generated a large number of mutants using P element insertional mutagenesis. The distribution of the target sites showed a marked preference for the 5′ end of gene [47]. However, all these early integration sites were from collections selected for phenotype and thus may be biased for specific types of integration events. Recently, a survey reported the cloning of 2266 unselected P element integration sites using the inverse PCR technique [48]. The authors focused on the analysis of the structural features of the integration sites and found that the target sites showed a palindromic structure. We aligned the reported junction site sequences to the *Drosophila* genome. Distribution of the integration sites showed that most insertions occurred within a few hundred bases of the transcription start site of genes [unpublished]. Remarkably, even though the transposition mechanism is only distantly related to the retroviral integration mechanism, the integration preference for P elements is very similar to the preference for MLV. Although CpG islands do not apply to *Drosophila*, integrations showed a preference for sites with high GC content. This raises the possibility that insertional elements could potentially be classified by their global integration preferences, and that P elements and MLV may share common factors that influence site selection. Another interesting concept is that retroviruses and transposable elements may potentially be classified by their global genomic preferences instead of, or in conjunction with, the apparent evolutionary relatedness of the integrase proteins.

Tc/mariner transposons

Tc/mariner transposons are a large family of DNA transposons with a widespread presence in many species [49]. Martin et al. reported 1088 integration sites of Tc1, Tc3 and a related Tc5 in the *Caenorhabditis elegans* genome [50]. No overall chromosomal feature bias for any of the three transposons was observed. A uniform distribution of integration sites was found along each chromosome. Tc1 showed some preference for certain chromosomes, with chromosome V having twice the density of integrations as chromosome III. In the previously described HIV-1 integration study, it was observed that HIV-1 integration sites distribution correlated with the gene density on the chromosome [22]. However, the difference of integration density observed for Tc1 in *C. elegans* cannot be explained by gene density on each chromosome, as GC content (36%) and gene density (1 gene/5 kb) for *C. elegans* are fairly uniform throughout all six chromosomes [51]. About 20% of the integrations are located in protein coding sequences and 30% in introns, compared to an expected value of 26% in protein coding region and 14% in introns if distribution were random. These results show that although Tc transposons have very little preference for coding regions, they highly prefers introns. One possible explanation for this comes from the knowledge that the Tc elements have a primary sequence target consensus sequence of TA (Tc5) or TNA (Tc1/Tc3), and the introns in *C. elegans* are AT rich (65% AT) compared to exons (57% AT).

Sleeping Beauty

Sleeping Beauty (SB) is a transposon in the Tc1/mariner family [52]. It was created from ancient non-active fish Tc1-like element through site-directed mutagenesis. SB can integrate into broad range of hosts. Vigdal et al. mapped 138 SB integration sites in human HeLa cells [53]. The target sites showed local sequence preference for a short, palindromic AT repeat: ATATATAT. But no preference for transcribed vs. non-transcribed DNA was observed. A recent survey of several hundred SB integrations in the mouse genome came to a similar conclusion [S. Yant, personal communication]. It appears (with the exception of Tc1's preference for chromosome V) that the transposable elements in *C. elegans* are more directly influenced by primary sequence than the retroviruses and retrotransposons are. However, P elements in *Drosophila* do seem to be an exception to this observation. Perhaps the more complex life cycles of the retroelements (having the RNA intermediate) required devising more sophisticated strategies for genome targeting than the transposable elements, which would improve their chance of survival.

Host co-factors

Several pieces of scientific evidence already discussed either suggest strongly or have already demonstrated that there are cellular factors that influence retroviral and retrotransposon site selection. It is less clear for the cut-and-paste transposable elements. Clear interactions with host proteins have been identified for the Ty elements, which have a strong influence on target selection. Yet it is still a wonder how this interaction is actually achieved. Imagining a huge Ty particle 'floating' around the nucleus trying to find a DNA-bound protein seems some-

what absurd, which would suggest that the interaction must be targeted in some way. Either the Ty particles are finding the DNA binding proteins out in the cytoplasm and are then co-imported to the nucleus where the host protein finds its DNA home, or there is some host machinery that is actively directing the Ty particles to their DNA target locations. There are many unanswered and exciting questions left on the cell biology of this phenomenon.

Another fascinating question is the difference between the HIV-1 and the MLV integration preferences. We showed in our work that HIV-1 integration exhibits a strong preference for active genes, but that preference drops off profoundly just 5′ or 3′ of the gene. Somehow the virus can tell the difference between the gene and the sequences just upstream and just downstream. One host factor that has been shown to interact with the HIV-1 integrase is Ini1, the human homolog of the yeast protein SNF5 [54]. Ini1 has been shown to be involved in chromatin remodeling [55], and deletion of the yeast SNF5 alleviates the toxicity caused by the expression of HIV-1 integrase in yeast cells [56], suggesting this is an important interaction for integrase targeting, or at least a subset of its functions. Other host factors such as HMG-1 [57] and hRad18 [58] have also been shown to interact with the integrase protein and may influence targeting to the active genes, but the roles for all of these proteins in site selection are not understood.

Even less is known about the MLV integrase and how it selects its target locations. The site preferences for MLV are less dramatic than for HIV-1 in that only roughly 20% of the integrations actually land in the 'preferred' regions, the remaining 80% of the integrations as of yet cannot be distinguished from a randomly generated set of integration sites. HMG-1 has been shown to also stimulate integration of MLV as it does with HIV-1 [57], but few other factors have been identified. By comparing integrations caused by infection to integrations generated by electroporation, Goetz et al. [59] were able to show that retroviral integrations favored chromosomal regions associated with the nuclear matrix. How that preference occurred is unknown. Bushman has proposed a tethering model for retroviral and retrotransposon integration where DNA binding proteins interact with the PICs of the retroelements and influence the actual site selection [60]. This tethering relationship has already been established for Ty elements, and the recent HIV-1, MLV, AAVr and P-element integration surveys are not inconsistent with such a model (see fig. 2). The interesting implication that some retroviruses used in gene therapy trials may be far enough

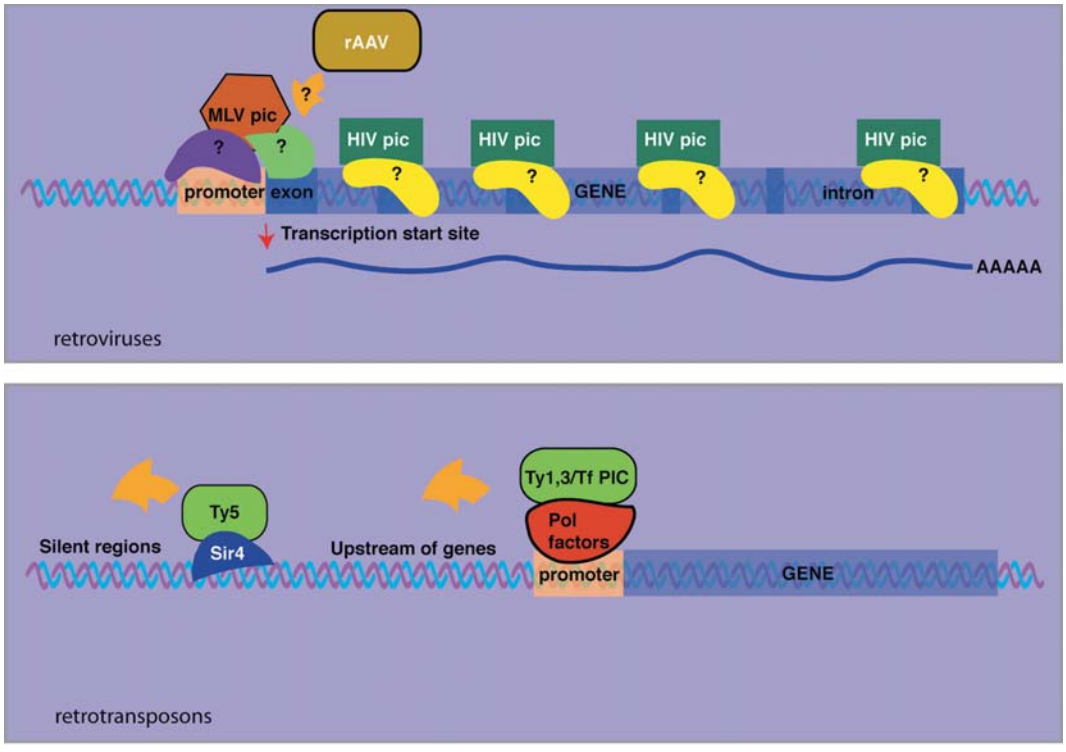

Figure 2. Model for host co-factors influencing integration. (*Top*) For both HIV-1 and MLV site selection, evidence suggests that cellular proteins are influencing site selection. Several cellular proteins have been identified, but no factors have been specifically shown to be involved in the site targeting. (*Bottom*) All Ty and Tf retrotransposon elements appear to use a similar mechanism, although the specific proteins they are binding to are different. All the retrotransposons appear to bind a polymerase factor and integrate immediately upstream of genes being transcribed by these factors.

removed from their original hosts that the tethering mechanisms may no longer be compatible and thus integration may not be targeted as efficiently. This could increase their relative safety from the standpoint of insertional mutagenesis. There are a great many roads that still need to be explored.

Targeting integration

It has become clear that retroviral vectors (MLV and HIV based) used in gene therapy can be deleterious, and the risk is higher than previously expected because of their preference for genes. One way to reduce the risk of insertional mutagenesis is to target retroviral integration into specific harmless sites in the human genome. Several studies have shown that it is possible to achieve certain target specificity by fusing integrase to sequence-specific DNA binding domains, including phage lambda repressor [61], LexA DNA binding domain [62, 63], transcription factor Zif268 [64] and synthetic zinc-finger binding protein E2C [65]. There are many limitations to this approach. For example, some of the DNA binding sequences such as lambda repressor and LexA binding sites do not exist in human genome. More recent studies use a mammalian zinc-finger binding domain [64] or a synthetic zinc-finger binding domain [65] (which have modifiable target specificity), and although the target sites may be localized to the desired position, there are too many variants of the target sites and the targeting effect is not absolute. However, even given perfect effectiveness, currently the major limitation for such chimeric integrase seems to be the adverse effects on the production of high titers of infectious virus.

Recently, Voytas's group took another approach for targeting integration of retroelements which explores protein-protein interactions, similar in principle to that of the yeast two-hybrid system [66]. Voytas's group found that the yeast retrotransposon Ty5 targets heterochomatin regions through interaction with the Sir4 protein. A targeting domain (TD) was mapped to a six-amino acid (aa) sequence near the end of the Ty5 integrase, which interacts with the C-terminal domain of Sir4p. In a series of elegant experiments, Voytas's group showed that integration could be targeted to a new LexA recognition site instead of the normal heterochomatic targets by expressing a LexA DNA binding domain/Sir4p C-terminal domain fusion protein. In this case, the DNA binding specificity was directed by the LexA DNA binding domain, and the targeting was achieved by the interaction between the IN TD and Sir4p C-terminal domain. The author went on to show that the 6-aa TD in the integrase could be replaced by other protein protein interaction domains such as the 13-aa motif of Rad9p that binds to the forkhead-associated domain (FHA1) of Rad53p, or the 12-aa motif of NpwBP, which binds to the WW domain of Npw38. Then the targeting specificity can accordingly be achieved by expressing fusion proteins of the LexA DNA binding domain with the corresponding binding partners. This approach opens new doors for targeting integration since there are many protein-protein interaction partners can be explored, and the DNA binding specificity can be modified by changing the DNA binding domain. However, it remains to be seen how well this approach can be tolerated by retroviral integrase and whether enough specificity can be generated by such an approach.

Conclusions

In summary, the genome-wide surveys for integration target sites of retroviruses and transposable elements have suggested that some retroviruses and transposable elements, including HIV-1, MLV, AAVr and the P element, can integrate into almost any region but prefer transcription active regions as integration target sites. Transcription start sites are especially hot regions for many viruses and transposable elements. This is either the result of open chromatin structure in the region or tethering by cellular factors to the region or both. The differences in the preference within the transcription active regions also suggest there are different cellular factors involved in the site selection for different viruses. Other elements, such as HTLV and Tc transposons, appear to integrate fairly randomly, perhaps influenced only by primary DNA sequence preferences such as restriction enzymes. Yet others, such as yeast Ty and Tf elements, integrate in a very site-specific fashion, clearly caused by interactions with host factors. With the increasing sensitivity of PCR cloning technology and the availability of more genomes, it will be helpful to survey many other retroviruses and transposable elements, both in their natural environment and in more distant host species. Comparison of the results from different viruses in different genomes will provide insights into the different mechanisms used for target selection and perhaps create a new understanding for the classification of the various types of integrating elements.

Acknowledgements. Research sponsored, at least in part, by the National Cancer Institute, the Department of Health and Human Services (DHHS), under contract N01-CO-12400 with SAIC-Frederick. The contents of this publication do not necessarily reflect the views or policies of the DHHS, nor does mention of trade names, commercial products or organizations imply endorsement by the US Government.

1 Hacein-Bey-Abina S., Von Kalle C., Schmidt M., McCormack M. P., Wulffraat N., Leboulch P. et al. (2003) LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. Science **302**(5644)**:** 415–419

2 Check E. (2003) Cancer fears cast doubts on future of gene therapy. Nature **421**(6924)**:** 678
3 Bor Y. C., Bushman F. D. and Orgel L. E. (1995) In vitro integration of human immunodeficiency virus type 1 cDNA into targets containing protein-induced bends. Proc. Natl. Acad. Sci. USA **92**(22)**:** 10334–10338
4 Bor Y. C., Miller M. D., Bushman F. D. and Orgel L. E. (1996) Target-sequence preferences of HIV-1 integration complexes in vitro. Virology **222**(1)**:** 283–288
5 Pryciak P. M., Sil A. and Varmus H. E. (1992) Retroviral integration into minichromosomes in vitro. EMBO J. **11**(1)**:** 291–303
6 Pryciak P. M. and Varmus H. E. (1992) Nucleosomes, DNA-binding proteins and DNA sequence modulate retroviral integration target site selection. Cell **69**(5)**:** 769–780
7 Pruss D., Reeves R., Bushman F. D. and Wolffe A. P. (1994) The influence of DNA and nucleosome structure on integration events directed by HIV integrase. J. Biol. Chem. **269**(40)**:** 25031–25041
8 Fitzgerald M. L. and Grandgenett D. P. (1994) Retroviral integration: in vitro host site selection by avian integrase. J. Virol **68**(7)**:** 4314–4321
9 Craigie R., Fujiwara T. and Bushman F. (1990) The IN protein of Moloney murine leukemia virus processes the viral DNA ends and accomplishes their integration in vitro. Cell **62**(4)**:** 829–837
10 Craigie R., Mizuuchi K., Bushman F. D. and Engelman A. (1991) A rapid in vitro assay for HIV DNA integration. Nucleic Acids Res. **19**(10)**:** 2729–2734
11 Bushman F. D., Fujiwara T. and Craigie R. (1990) Retroviral DNA integration directed by HIV integration protein in vitro. Science **249**(4976)**:** 1555–1558
12 Bushman F. D. and Craigie R. (1990) Sequence requirements for integration of Moloney murine leukemia virus DNA in vitro. J. Virol. **64**(11)**:** 5645–5648
13 Bushman F. D. and Craigie R. (1991) Activities of human immunodeficiency virus (HIV) integration protein in vitro: specific cleavage and integration of HIV DNA. Proc. Natl. Acad. Sci. USA **88**(4)**:** 1339–1343
14 Muller H. P. and Varmus H. E. (1994) DNA bending creates favored sites for retroviral integration: an explanation for preferred insertion sites in nucleosomes. EMBO J. **13**(19)**:** 4704–4714
15 Vijaya S., Steffen D. L. and Robinson H. L. (1986) Acceptor sites for retroviral integrations map near DNase I-hypersensitive sites in chromatin. J. Virol. **60**(2)**:** 683–692
16 Rohdewohld H., Weiher H., Reik W., Jaenisch R. and Breindl M. (1987) Retrovirus integration and chromatin structure: Moloney murine leukemia proviral integration sites map near DNase I-hypersensitive sites. J. Virol. **61**(2)**:** 336–343
17 Scherdin U., Rhodes K. and Breindl M. (1990) Transcriptionally active genome regions are preferred targets for retrovirus integration. J. Virol. **64**(2)**:** 907–912
18 Stevens S. W. and Griffith J. D. (1994) Human immunodeficiency virus type 1 may preferentially integrate into chromatin occupied by L1Hs repetitive elements. Proc. Natl. Acad. Sci. USA **91**(12)**:** 5557–5561
19 Stevens S. W. and Griffith J. D. (1996) Sequence analysis of the human DNA flanking sites of human immunodeficiency virus type 1 integration. J. Virol. **70**(9)**:** 6459–6462
20 Withers-Ward E.S., Kitamura Y., Barnes J. P. and Coffin J. M. (1994) Distribution of targets for avian retrovirus DNA integration in vivo. Genes Dev. **8**(12)**:** 1473–1487
21 Weidhaas J. B., Angelichio E. L., Fenner S. and Coffin J. M. (2000) Relationship between retroviral DNA integration and gene expression. J. Virol. **74**(18)**:** 8382–8389
22 Schroder A. R., Shinn P., Chen H., Berry C., Ecker J. R. and Bushman F. (2002) HIV-1 integration in the human genome favors active genes and local hotspots. Cell **110**(4)**:** 521–529
23 Wu X., Li Y., Crise B. and Burgess S. M. (2003) Transcription start regions in the human genome are favored targets for MLV integration. Science **300**(5626)**:** 1749–1751
24 Pruitt K. D., Katz K. S., Sicotte H. and Maglott D. R. (2000) Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. Trends Genet. **16**(1)**:** 44–47
25 Maglott D. R., Katz K. S., Sicotte H. and Pruitt K. D. (2000) NCBI's LocusLink and RefSeq. Nucleic Acids Res. **28**(1)**:** 126–128
26 Lund A. H., Turner G., Trubetskoy A., Verhoeven E., Wientjens E., Hulsman D. et al. (2002) Genome-wide retroviral insertional tagging of genes involved in cancer in Cdkn2a-deficient mice. Nat. Genet. **32**(1)**:** 160–165
27 Kim R., Trubetskoy A., Suzuki T., Jenkins N. A., Copeland N. G. and Lenz J. (2003) Genome-based identification of cancer genes by proviral tagging in mouse retrovirus-induced T-cell lymphomas. J. Virol. **77**(3)**:** 2056–2062
28 Suzuki T., Shen H., Akagi K., Morse H. C., Malley J. D., Naiman D. G. et al. (2002) New genes involved in cancer identified by retroviral tagging. Nat. Genet. **32**(1)**:** 166–174
29 Hwang H. C., Martins C. P., Bronkhorst Y., Randel E., Berns A., Fero M., et al. (2002) Identification of oncogenes collaborating with p27Kip1 loss by insertional mutagenesis and high-throughput insertion site analysis. Proc. Natl. Acad. Sci. USA **99**(17)**:** 11293–11298
30 Akagi K., Suzuki T., Stephens R. M., Jenkins N. A. and Copeland N. G. (2004) RTCGD: retroviral tagged cancer gene database. Nucleic Acids Res **32 Database issue:** D523–D527
31 Leclercq I., Mortreux F., Cavrois M., Leroy A., Gessain A., Wain-Hobson S. et al. (2000) Host sequences flanking the human T-cell leukemia virus type 1 provirus in vivo. J. Virol. **74**(5)**:** 2305–2312
32 Kotin R. M., Siniscalco M., Samulski R. J., Zhu X. D., Hunter L., Laughlin C. A. et al. (1990) Site-specific integration by adeno-associated virus. Proc. Natl. Acad. Sci. USA **87**(6)**:** 2211–2215
33 Samulski R. J., Zhu X., Xiao X., Brook J. D., Housman D. E., Epstein N. et al. (1991) Targeted integration of adeno-associated virus (AAV) into human chromosome 19. EMBO J. **10**(12)**:** 3941–3950
34 Weitzman M. D., Kyostio S. R., Kotin R. M. and Owens R. A. (1994) Adeno-associated virus (AAV) Rep proteins mediate complex formation between AAV DNA and its integration site in human DNA. Proc. Natl. Acad. Sci. USA **91**(13)**:** 5808– 5812
35 Linden R. M. and Berns K. I. (1996) Site-specific integration by adeno-associated virus. Proc. Natl. Acad. Sci. USA **93**(21)**:** 11288–11294
36 Nakai H., Moutini E., Fuess S., Storm A., Grompe M. and Kay M. A. (2003) AAV serotype 2 vectors preferentially integrate into active genes in mice. Nat. Genet. **34**(3)**:** 297–302
37 Boeke J. D. and Devine S. E. (1998) Yeast retrotransposons: finding a nice quiet neighborhood. Cell **93**(7)**:** 1087–1089
38 Devine S. E. and Boeke J. D. (1996) Integration of the yeast retrotransposon Ty1 is targeted to regions upstream of genes transcribed by RNA polymerase III. Genes Dev. **10**(5)**:** 620–633
39 Chalker D. L. and Sandmeyer S. B. (1993) Sites of RNA polymerase III transcription initiation and Ty3 integration at the U6 gene are positioned by the TATA box. Proc. Natl. Acad. Sci. USA **90**(11)**:** 4927–4931
40 Kirchner J., Connolly C. M. and Sandmeyer S. B. (1995) Requirement of RNA polymerase III transcription factors for in vitro position-specific integration of a retroviruslike element. Science **267**(5203)**:** 1488–1491
41 Zhu Y., Zou S., Wright D. A. and Voytas D. F. (1999) Tagging chromatin with retrotransposons: target specificity of the Saccharomyces Ty5 retrotransposon changes with the chromosomal localization of Sir3p and Sir4p. Genes Dev. **13**(20)**:** 2738–2749

42 Xie W., Gai X., Zhu Y., Zappulla D. C., Sternglanz R. and Voytas D. F. (2001) Targeting of the yeast Ty5 retrotransposon to silent chromatin is mediated by interactions between integrase and Sir4p. Mol. Cell. Biol. **21**(19)**:** 6606–6614
43 Singleton T. L. and Levin H. L. (2002) A long terminal repeat retrotransposon of fission yeast has strong preferences for specific sites of insertion. Eukaryot. Cell **1**(1)**:** 44–55
44 Behrens R., Hayles J. and Nurse P. (2000) Fission yeast retrotransposon Tf1 integration is targeted to 5′ ends of open reading frames. Nucleic Acids Res. **28**(23)**:** 4709–4716
45 O'Hare K. and Rubin G. M. (1983) Structures of P transposable elements and their sites of insertion and excision in the *Drosophila melanogaster* genome. Cell **34**(1)**:** 25–35
46 Berg C.A. and Spradling A. C. (1991) Studies on the rate and site-specificity of P element transposition. Genetics **127**(3)**:** 515–524
47 Spradling A.C., Stern D. M., Kiss I., Roote J., Laverty T. and Rubin G. M. (1995) Gene disruptions using P transposable elements: an integral component of the *Drosophila* genome project. Proc. Natl. Acad. Sci. USA **92**(24)**:** 10824–10830
48 Liao G. C., Rehm E. J. and Rubin G. M. (2000) Insertion site preferences of the P transposable element in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **97**(7)**:** 3347–3351
49 Plasterk R. H., Izsvak Z. and Ivics Z. (1999) Resident aliens: the Tc1/mariner superfamily of transposable elements. Trends Genet. **15**(8)**:** 326–332
50 Martin E., Laloux H., Couette G., Alvarez T., Bessou C., Hauser O. et al. (2002) Identification of 1088 new transposon insertions of *Caenorhabditis elegans*: a pilot study toward large-scale screens. Genetics **162**(1)**:** 521–524
51 The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 1998. **282**(5396)**:** 2012–2018
52 Ivics Z., Hackett P. B., Plasterk R. H. and Izsvak Z. (1997) Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. Cell **91**(4)**:** 501–510
53 Vigdal T. J., Kaufman C. D., Izsvak Z., Voytas D. F and Ivics C. (2002) Common physical properties of DNA affecting target site selection of sleeping beauty and other Tc1/mariner transposable elements. J. Mol. Biol. **323**(3)**:** 441–452
54 Kalpana G. V., Marmon S., Wang W., Crabtree G. R. and Goff S. P. (1994) Binding and stimulation of HIV-1 integrase by a human homolog of yeast transcription factor SNF5. Science **266**(5193)**:** 2002–2006
55 Wang W., Cote J., Xue Y., Zhou S., Khavari P. A., Biggar S. R. et al. (1996) Purification and biochemical heterogeneity of the mammalian SWI-SNF complex. EMBO J. **15**(19)**:** 5370–5382
56 Parissi V., Caumont A., Richard de, Soultrait V., Dupont C. H., Pichuantes S. and Litvak S. (2000) Inactivation of the SNF5 transcription factor gene abolishes the lethal phenotype induced by the expression of HIV-1 integrase in yeast. Gene **247**(1–2)**:** 129–136
57 Li L., Yoder K., Hansen M. S., Olvera J., Miller M. D. and Bushman F. D. (2000) Retroviral cDNA integration: stimulation by HMG I family proteins. J. Virol. **74**(23)**:** 10965–10974
58 Mulder L. C., Chakrabarti L. A. and Muesing M. A. (2002) Interaction of HIV-1 integrase with DNA repair protein hRad18. J. Biol. Chem. **277**(30)**:** 27489–27493
59 Goetze S., Huesemann Y., Baer A. and Bode J. (2003) Functional characterization of transgene integration patterns by halo fluorescence in situ hybridization: electroporation versus retroviral infection. Biochemistry **42**(23)**:** 7035–7043
60 Bushman F. D. (2003) Targeting survival: integration site selection by retroviruses and LTR-retrotransposons. Cell **115**(2)**:** 135–138
61 Bushman F. D. (1994) Tethering human immunodeficiency virus 1 integrase to a DNA site directs integration to nearby sequences. Proc. Natl. Acad. Sci. USA **91**(20)**:** 9233–9237
62 Goulaouic H. and Chow S. A. (1996) Directed integration of viral DNA mediated by fusion proteins consisting of human immunodeficiency virus type 1 integrase and *Escherichia coli* LexA protein. J. Virol. **70**(1)**:** 37–46
63 Katz R. A., Merkel G. and Skalka A. M. (1996) Targeting of retroviral integrase by fusion to a heterologous DNA binding domain: in vitro activities and incorporation of a fusion protein into viral particles. Virology **217**(1)**:** 178–190
64 Bushman F. D. and Miller M. D. (1997) Tethering human immunodeficiency virus type 1 preintegration complexes to target DNA promotes integration at nearby sites. J. Virol. **71**(1)**:** 458–464
65 Tan W., Zhu K., Segal D. J., Barbas C. F. 3rd and Chow S. A. (2004) Fusion proteins consisting of human immunodeficiency virus type 1 integrase and the designed polydactyl zinc finger protein E2C direct integration of viral DNA into specific sites. J. Virol. **78**(3)**:** 1301–1313
66 Zhu Y., Dai J., Fuerst P. G. and Voytas D. F. (2003) Controlling integration specificity of a yeast retrotransposon. Proc. Natl. Acad. Sci. USA **100**(10)**:** 5891–5895


To access this journal online:
http://www.birkhauser.ch