Weak Palindromic Consensus Sequences Are a Common Feature Found at the Integration Target Sites of Many Retroviruses

Xiaolin Wu,1* Yuan Li,2 Bruce Crise,2 Shawn M. Burgess,3 and David J. Munroe1

Laboratory of Molecular Technology1 and AIDS Vaccine Program,2 Scientific Application International Corporation, National Cancer Institute at Frederick, Frederick, and Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland

Received 2 September 2004/Accepted 14 December 2004

Integration into the host genome is one of the hallmarks of the retroviral life cycle and is catalyzed by virus-encoded integrases. While integrase has strict sequence requirements for the viral DNA ends, target site sequences have been shown to be very diverse. We carefully examined a large number of integration target site sequences from several retroviruses, including human immunodeficiency virus type 1, simian immunodeficiency virus, murine leukemia virus, and avian sarcoma-leukosis virus, and found that a statistical palindromic consensus, centered on the virus-specific duplicated target site sequence, was a common feature at integration target sites for these retroviruses.

Much is known about the sequence requirements at the end of the viral DNA for efficient integration of retroviruses (6). A dinucleotide CA is invariably positioned exactly 2 bp from both ends of the viral termini. The sequences internal to the CA dinucleotide extending for up to 15 bp also have significant roles. However, despite decades of effort, the mechanism of target site selection remains largely unknown. Evidence accumulated to date shows that most of the regions of the host genome are potential retroviral integration target sites, but integration is usually not random and the preferences appear to be specific to the individual viruses (12, 22, 26). Target site selection can be influenced by many factors, including DNA binding proteins (1, 13, 17, 20), the chromatin structure of DNA (18, 19, 24), and, perhaps most importantly, cellular targeting proteins (4, 25). It is still unclear how primary sequence at the target sites influences the target site selection, although weak consensus sequences have been reported for the target sites of several retroviruses (2, 5, 7, 18, 19, 23).

Recently, there have been several large-scale surveys of retroviral integration sites in the human genome (12, 15, 22, 26). We downloaded these sequences from GenBank and mapped these integration sites to the human genome, using the BLAT program on the University of California—Santa Cruz genome server (November 2003 freeze; UCSC Human Genome Project, http://genome.ucsc.edu). Included in our analysis were integration sites for human immunodeficiency virus type 1 (HIV-1; GenBank accession no. BH605938 to BH612086) (22), simian immunodeficiency virus (SIV) (GenBank accession no. AY679815 to AY680027), murine leukemia virus (MLV) (GenBank accession no. AY515855 to AY516880) (26), and avian sarcoma-leukosis virus (ASLV) (GenBank accession no. CLS28318 to CLS28772) (12). A total of 334 in vivo HIV-1 sites in SupT1 cells, 81 in vitro HIV-1 sites in naked DNA catalyzed by HIV preintegration complexes (PICs), 148 SIV sites in CEMx174 cells, 695 MLV sites in HeLa cells, and 357 ASLV sites in 293T-TVA cells were mapped.

The sequences upstream and downstream of proviral integration sites (same orientation as virus) were extracted for further analysis. All sequences were aligned at the integration sites (between base −1 and base 1), and the frequencies of A, C, G, and T at each position around the integration site were calculated. These values were compared to the expected value based on the total base frequency of the human genome or values from 500 computer-generated random integration site sequences in the human genome (Fig. 1). The human genome is relatively AT rich (60% AT and 40% GC). At any random site, the expected frequencies for A, C, G, and T are 30, 20, 20, and 30%, respectively. The base composition at each position around the 500 computer-generated random integration sites varies little from the expected value (Fig. 1A). In Fig. 1, we emphasized significant frequency changes at any base position adjacent to the precise integration site by highlighting changes of 10% or greater than expected (green for >10% increase and red for >10% decrease). Clear statistical preferences are observed for each virus, and they are different for each of the genera analyzed.

We compared the base compositions of in vivo HIV-1 integration sites in SupT1 cells to randomly generated sites. The frequencies of some bases at specific positions around HIV-1 target sites are significantly higher or lower than the expected value (Fig. 1B). For example, base position 1 shows a preference for G (40%) and avoidance for T (9%). Bases 2 and 4 show preferences for T (54%) and A (46%), respectively. Base 5 shows preference for C (41%) and avoidance for A (10%). These values are either 10% higher or 10% lower than the expected frequencies at random sites. To evaluate the statistical significance of these differences, we performed bootstrapping by randomly choosing 334 sites from the human genome (to match our in vivo HIV-1 sample size) and computed the base composition for each of the 20 positions surrounding the random sites. This process was repeated 10,000 times, and only 13 times were frequency changes >10% at any position among the 20-bp DNA: this corresponds to a P value of 0.0013. When
the changes are highlighted, it is easy to see that these preferences are symmetrically centered on base 3, forming a statistical palindrome. Interestingly, this palindrome is centered on the duplicated target site sequence, which comprises bases 1 to 5. The same statistical palindromed was also observed for HIV-1 integration sites in HeLa cells and mouse bone marrow cells (data not shown), suggesting that the preference is not cell line specific or species specific and may represent an intrinsic property of target site recognition by integrase.

We then calculated the base composition of HIV-1 integration sites in naked SupT1 genomic DNA catalyzed by PICs (22). The preference pattern is similar to that of in vivo HIV-1 integration sites (Fig. 1C). However, the preference outside the duplicated target site sequence differs slightly. Similar consensus sequences for HIV-1 target sites were reported previously, and synthetic oligonucleotides with the consensus sequence were shown indeed to be the favored target sites by PICs (5). These results with naked DNA targets suggest that preferences observed in vivo are due to recognition determinants of the integration machinery itself and not the influence of DNA binding proteins or chromatin structure.

SIV is closely related to HIV-1. The genome structure and proteins encoded by these two viruses share a great deal of homology. It was therefore interesting to see if SIV and HIV-1 shared the same target site preference. Our analysis showed that SIV integration sites had a similar statistical palindromic composition (Fig. 1D). However, there are differences between SIV and HIV-1, and they mainly lie outside of the 5-bp duplicated target site. For example, SIV has higher frequency of G at base –1 and T at base –2. Perhaps the most significant difference is at base –3, the third base outside the duplicated target site, where HIV-1 prefers T while SIV prefers G. The difference shows palindromic symmetry on the other side of the integration site, where SIV has a higher frequency of C, A, and C at bases 6, 7, and 8. The overall similarity of the target site sequence indicates that SIV and HIV-1 integration machinery may differ only slightly from each other.

MLV belongs to the genus Gammaretrovirus, which differs from lentiviruses in many aspects. Integration of MLV requires passage through mitosis, whereas lentiviruses do not (9, 21). MLV and HIV-1 also have distinct global target site preferences (22, 26). MLV highly prefers transcription start site regions, while HIV-1 prefers anywhere inside actively transcribed regions. Alignment of 695 MLV integration sites in HeLa cells revealed a different target site, but clearly significant consensus sequence (Fig. 1E). Like the HIV-1 sites, the consensus is also centered on the duplications that occur at the

---

**FIG. 1.** Palindromic consensus sequences at retroviral integration sites. Base compositions of the top DNA strand at each position around the integration sites were calculated. Integration occurs between positions –1 and 1 on the top strand. Colored positions have statistically different frequencies of bases from that of randomly generated positions (P < 0.01), which is 30%/20%/20%/30% for A/C/G/T. Bases with a greater than 10% increase of frequency at a position are green, and bases with a greater than 10% decrease of frequency at a position are red. Inferred duplicated target sites are in the blue box, and DNA strand transfer occurs at positions labeled by arrows. The base preference shows a palindromic pattern centered on the duplicated target sites, and symmetry is marked with the purple vertical line.
target site, which is 4 bp long for MLV instead of 5 bp long for HIV.

We also analyzed the target site sequence of another retrovirus, ASLV, which belongs to the genus Alpharetrovirus. From 148 integration sites, we deduced a weak palindromic target site consensus sequence for ASLV (Fig. 1F). This palindromic structure is centered on a 6-bp sequence fragment, which also coincides with the inferred duplicated target site for ASLV. As shown in Fig. 1, the palindromic target site consensus sequence for all four retroviruses extends beyond the target site duplications, suggesting that bases outside the very short target site duplications also contribute to target site selection.

It is interesting to note that a statistical palindromic consensus sequence has also been reported for the P transposable element in Drosophila (10), suggesting that a palindromic feature might be shared widely among many integrase or transposases. Like P element target sequences, we found very few individual retroviral target site sequences that correspond to the consensus, based on the favorite base at each position. For example, only 10 of 334 individual HIV in vivo sites have $G_T^2(A/T)_3A_4C_5$ at the duplicated target sites. There are several possible explanations for this discrepancy. First, the consensus is very weak and thus can only be found with large data sets like those used in this study. Second, there might be secondary DNA structures that can only be reflected partially by the primary sequences. To evaluate this possibility, we analyzed several known physical properties of the integration site DNA (Fig. 2).

We examined four different DNA structural properties, including A-philicity (8, 11), DNA bendability (3, 14), protein-induced deformability (16), and hydrogen bond (H-bond) potential patterns (10) for HIV, SIV, MLV, and ASLV integration site DNA (Fig. 2). A-philicity measures the propensity of DNA to form an A DNA-like double helix, which has a wide and shallow minor groove believed to give proteins easier access to form hydrogen bonds with bases within the DNA helix (8, 11). DNA bendability also changes the width and depths of the major groove and minor groove, affecting protein access (3, 14). Protein-induced deformability represents the impact of protein binding on DNA topology (16). H-bond potential patterns describe the potential hydrogen bond donors and acceptors of a base pair in the major groove of DNA that interacts with proteins (10). All of these properties are based on DNA primary sequence. However, H-bond potential is calculated based on single-nucleotide frequencies; A-philicity and protein-induced deformability are calculated based on dinucleotide frequencies; and DNA bendability is calculated based on trinucleotide frequencies. All four retroviruses showed significant signal change for the A-philicity score at integration sites when compared to computer-generated random integration sites (Fig. 2A). For the DNA bendability score, HIV and SIV showed more significant changes than MLV and ASLV at the integration sites (Fig. 2B). Significant changes were also observed for the measurement of protein-induced deformability at the integration sites of HIV, SIV, and MLV, while the change was less dramatic for ASLV. Also, H-bond potential exhibited palindromic patterns centered on the duplicated target sites for all four retroviruses (Fig. 2D). From these analyses, it is obvious that many structural properties are favored at the retroviral integration sites.

Our results suggest that the observed statistical palindromic primary sequence might reflect the influences of integrase on site selection at target sites. The symmetry of the target site sequence might reflect that the integrase complex works in symmetrical dimers, tetramers, or oligomers at the integration sites, such that each half-complex would have a similar preference for target DNA structure. Our results also imply that it may not be appropriate to think of the consensus sequences as...
the most favorite sequence at each base. It might be better to think of certain bases being excluded at certain positions to meet the spatial or energy requirements of the integration complexes. For example, all four retroviruses and even P element transposons do not prefer T at the first base of the duplicated target site sequence (or A at the last base). In fact, we only observed two individual target site sequences with T1N2N3N4A5 out of 334 (0.6%) HIV-1 integration sites. This is statistically lower than the frequency of random genomic site sequence, where T1N2N3N4A5 can be expected at 9% (30% A × 30% T) or 30 out of 334 (P < 0.01 using a chi-square test). Similarly, T1N2N3N4As, T1N2N3As, and T1N2N3N4As are observed at statistically lower frequencies for SIV (P < 0.05), MLV (P < 0.001), and ASLV (P < 0.001), respectively. This common avoidance may reflect the physical or chemical restraints for position 1 during the DNA cleavage and strand transfer reaction catalyzed by the integrase. All retroviruses showed very low A-philicity scores at base 1, and dinucleotides TA, TC, TG, and TT had high A-philicity scores. Thus, if low A-philicity is truly a requirement for base 1, T will be unlikely to appear at this position. Likewise, many other factors also may contribute to the selection of spatially “best-fit” target sites. The exact structural property of the integration sites will be better understood as our knowledge of the DNA physical structure advances.

This work has been funded in part with Federal Funds from the National Cancer Institute, National Institute of Health, DHHS, under contract N01-CO-12400.

The HbondView software was a kind gift from G. C. Liao.

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 U.S. Government.

REFERENCES


