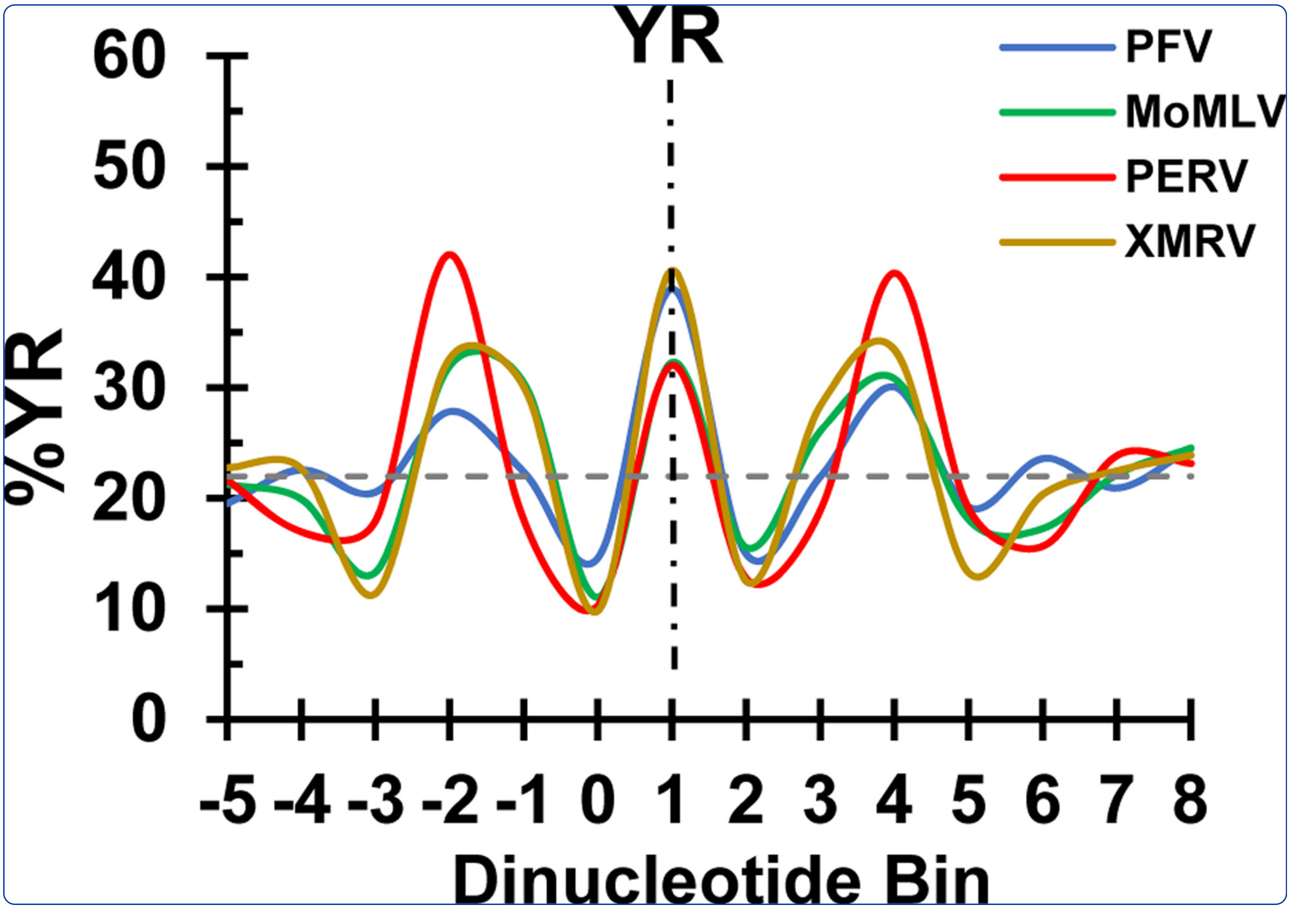
RETROVIROLOGY



Key determinants of target DNA recognition by retroviral intasomes

Serrao *et al.*

BioMed Central

Serrao *et al. Retrovirology* (2015) 12:39
DOI 10.1186/s12977-015-0167-3

Serrao *et al. Retrovirology* (2015) 12:39
DOI 10.1186/s12977-015-0167-3

RETROVIROLOGY

RESEARCH **Open Access**

Key determinants of target DNA recognition by retroviral intasomes

Erik Serrao[1], Allison Ballandras-Colas[1], Peter Cherepanov[2,3], Goedele N Maertens[2] and Alan N Engelman[1*]

Abstract

Background: Retroviral integration favors weakly conserved palindrome sequences at the sites of viral DNA joining and generates a short (4–6 bp) duplication of host DNA flanking the provirus. We previously determined two key parameters that underlie the target DNA preference for prototype foamy virus (PFV) and human immunodeficiency virus type 1 (HIV-1) integration: flexible pyrimidine (Y)/purine (R) dinucleotide steps at the centers of the integration sites, and base contacts with specific integrase residues, such as Ala188 in PFV integrase and Ser119 in HIV-1 integrase. Here we examined the dinucleotide preference profiles of a range of retroviruses and correlated these findings with respect to length of target site duplication (TSD).

Results: Integration datasets covering six viral genera and the three lengths of TSD were accessed from the literature or generated in this work. All viruses exhibited significant enrichments of flexible YR and/or selection against rigid RY dinucleotide steps at the centers of integration sites, and the magnitude of this enrichment inversely correlated with TSD length. The DNA sequence environments of *in vivo*-generated HIV-1 and PFV sites were consistent with integration into nucleosomes, however, the local sequence preferences were largely independent of target DNA chromatinization. Integration sites derived from cells infected with the gammaretrovirus reticuloendotheliosis virus strain A (Rev-A), which yields a 5 bp TSD, revealed the targeting of global chromatin features most similar to those of Moloney murine leukemia virus, which yields a 4 bp duplication. *In vitro* assays revealed that Rev-A integrase interacts with and is catalytically stimulated by cellular bromodomain containing 4 protein.

Conclusions: Retroviral integrases have likely evolved to bend target DNA to fit scissile phosphodiester bonds into two active sites for integration, and viruses that cut target DNA with a 6 bp stagger may not need to bend DNA as sharply as viruses that cleave with 4 bp or 5 bp staggers. For PFV and HIV-1, the selection of signature bases and central flexibility at sites of integration is largely independent of chromatin structure. Furthermore, global Rev-A integration is likely directed to chromatin features by bromodomain and extraterminal domain proteins.

Keywords: Retrovirus, Integrase, DNA flexibility, Dinucleotide steps, Integration sites, Nucleosomes, BET proteins

Background

The integration of a DNA copy of the viral RNA genome into a host cell chromosome is a critical step in the retroviral lifecycle. Retroviruses accordingly encode for an integrase (IN) enzyme, which is a specialized DNA recombinase. Integration begins with the formation of the intasome nucleoprotein complex, which consists of an IN tetramer assembled on the ends of the linear viral DNA (vDNA) [1-3]. The two inner subunits of the tetramer cleave the vDNA ends adjacent to invariant 5′-CA-3′ dinucleotides to yield reactive CA_{OH}-3′ hydroxyl groups [3-8]. The intasome is transported from the cytoplasm to the nucleus as part of a large assembly of viral and host proteins known as the preintegration complex [9-11]. In the nucleus the intasome engages host cell chromatin to form the target capture complex (TCC) [3,12]. The inner subunits of the IN tetramer utilize the vDNA CA_{OH}-3′ termini to cleave both strands of the target DNA (tDNA) in a staggered fashion, at the same time joining the vDNA ends to tDNA 5′-phosphates [13]. The resulting DNA recombination intermediate contains free vDNA 5′ ends abutting single stranded gaps in the tDNA, which vary in length from four to six nucleotides, depending on the retrovirus [14-17]. The single-stranded gaps are repaired

* Correspondence: alan_engelman@dfci.harvard.edu
[1]Department of Cancer Immunology and AIDS, Dana-Farber Cancer Institute, Boston, MA, USA
Full list of author information is available at the end of the article

by host cell enzymes to yield a target site duplication (TSD) of 4–6 bp flanking the provirus.

Several features of the animal cell genome, from the tDNA sequence at the site of integration to higher-order chromatin structure, can influence the selection of retroviral integration sites (see [18] for a recent review). Seven different genera, alpha through epsilon, lenti, and spuma, comprise the Retroviridae family, and preferential targeting of structural chromatin features is most evident for the lenti- and gammaretroviruses. Lentiviruses preferentially integrate along the bodies of actively transcribed genes [19], whereas the gammaretroviruses favor transcriptional start sites (TSSs) and active enhancer regions [20-22]. These preferences are in large part governed by interactions between IN proteins and cognate cellular factors [18]. The lentiviral IN-binding protein lens epithelium-derived growth factor (LEDGF)/p75 directs integration to active genes [23-26], whereas bromodomain and extraterminal domain (BET) proteins BRD2, 3, and 4 interact with Moloney murine leukemia virus (MoMLV) IN to affect TSS-proximal integration [27-29]. Viruses from the other profiled genera – integration site preferences of epsilonretroviruses have not been reported – show less propensity to target chromatin-specific features than do either the lentiviruses or gammaretroviruses, with the betaretrovirus mouse mammary tumor virus (MMTV) displaying the least selectivity of all [30].

Analyses of retroviral integration sites revealed weak palindromic tDNA sequence consensuses at the sites of vDNA joining [14-17,31]. A palindromic consensus implies dyadic symmetry within the IN nucleoprotein complex that engages tDNA, and crystallographic analysis of the prototype foamy virus (PFV) TCC revealed key features of the inner IN dimer within the tetramer that dictate the selection of the consensus PFV integration site (−3)KWK\VYRBMWM(+6) (written using International Union of Biochemistry base codes; the backslash indicates the position of vDNA plus-strand joining, and the italics mark the TSD, which is 4 bp for PFV) [12]. The tDNA is accommodated in a severely bent conformation, with the major groove widened such that the dinucleotide at the center of the integration site (YR) is unstacked. Given the relatively weak nature of nucleotide specificity at integration sites, it was not surprising that a number of IN main chain amide groups interacted with the tDNA backbone in the TCC structure. In addition, base specific contacts were observed for PFV IN residues Ala188, which resides in the catalytic core domain (CCD), and Arg329, which is part of the IN C-terminal domain. Ala188 in particular interacted with bases that lay 3 positions upstream from the points of vDNA joining, whereas Arg329 interacted with bases at either edge of the integration site, as well as those at symmetric nucleotide positions −2 and +5 [12].

The variety of dinucleotide steps differ in their propensity to support distortion of a DNA double helix, which reflects their inherent base stacking interactions [32]. Pyrimidine-purine (YR) and RY steps are the most and least distortable, respectively, whereas YY and RR display an intermediary level of flexibility. The strongly preferred YR at the center of PFV integration sites accommodates the sharp tDNA bend required for integration. The consensus target site sequence (−3)TDG\*(G/V)TWA(C/B)*CHA(+7) for human immunodeficiency virus type 1 (HIV-1) was subsequently shown to harbor the dinucleotide signature motif (0)RYXRY(+4), which selects against rigid RY dinucleotides at the center of integration sites (due to the odd number of bp between the sites of HIV-1 DNA joining, the location of the integration site center encompasses two overlapping positions: nucleotides +1 and +2 and nucleotides +2 and +3) [33].

Inherent curvature or bendability of DNA substrates positively correlates with frequency of integration targeting [34-36], and tDNA deformed by the binding of nucleosomes [37-39] or other DNA bending proteins [34,40] can be utilized preferentially by IN over naked DNA *in vitro*. Nucleosomes are favored sites for integration during MoMLV and HIV-1 infection [41-45]. Moreover, A/T-rich sequences that emanate outward from the central, local palindrome at the sites of vDNA insertion exhibit periodicity coincident with the outward-facing major grooves on the nucleosome surface [34,37,38,42-44]. However, because PFV [46] and HIV-1 [33] integration *in vivo* and using naked plasmid tDNA substrates *in vitro* generated similar palindrome signatures, the forces that govern the selection of particular bases at the sites of integration appear for the most part independent of tDNA chromatinization.

In this study we extended dinucleotide step analysis of retroviral integration sites to a total of 12 viruses. We find that central flexibility is a conserved feature and that it inversely correlates with the length of TSD. By comparing integration sites in naked plasmid or cellular tDNA to those generated during PFV and HIV-1 infection, we confirm that central flexibility and local nucleotide preferences are for the most part independent of nucleosome content. Furthermore, we report the integration site preferences of reticuloendotheliosis virus strain A (Rev-A) in infected cells, which paralleled those of previously reported gammaretroviruses despite the fact that Rev-A integration generates a 5 bp duplication of tDNA. Thus, Rev-A integration distribution mirrored that of MoMLV, and we accordingly show that Rev-A IN interacts with and is catalytically stimulated by a portion of the MoMLV integration host cofactor BRD4 that contains the IN-interacting extraterminal (ET) domain. Akin to MoMLV [27-29], IN binding to BET proteins likely directs Rev-A integration to chromatin features such as TSSs.

Results

Analytic strategy

In light of the similarity in the selection for central flexibility at sites of PFV and HIV-1 integration, we extended dinucleotide frequency analyses to a total of 12 retroviruses. Considering that retroviral TSDs vary from 4 to 6 bp, we analyzed four viruses that generate 4 bp duplications, four that generate 5 bp duplications, and four that generate 6 bp duplications. Viruses from two to three different genera comprise each of these subsets (Table 1). Four bp TSDs are yielded by the spumavirus PFV [47,48] as well as the gammaretroviruses MoMLV [49,50], porcine endogenous retrovirus (PERV) [51,52], and xenotropic murine leukemia virus-related virus (XMRV) [53,54]. The alpharetrovirus avian sarcoma-leukosis virus (ASLV) [55,56], deltaretrovirus human T-lymphotropic virus type 1 (HTLV-1) [57,58], and betaretroviruses human endogenous retrovirus K family (HERV-K) [59] and MMTV [60] yield 6 bp TSDs. In addition to HIV-1 [61,62], the lentiviruses simian immunodeficiency virus (SIV) [63] and equine infectious anemia virus (EIAV) [64], as well as the gammaretrovirus Rev-A [65,66], yield 5 bp TSDs.

Perusal of the literature revealed that the number of reported integration site datasets from virus-infected cells ranged from a small handful for Rev-A [65] to several million for MoMLV [21,22]. Rev-A is of particular interest, as it is a gammaretrovirus with a 5 bp TSD; all other gammaretroviruses yield 4 bp TSDs [14,15,17]. We accordingly initiated this study by determining the sequences of 834 unique integration sites from HEK293T cells infected with a Rev-A viral vector by ligation-mediated PCR. The targeting preferences of Rev-A for genomic annotations such as genes, TSSs, CpG islands, and gene density are described toward the end of the Results section. As the HEK293T DNA was fragmented by digestion with restriction endonucleases AvrII, NheI, and SpeI, a matched random control (MRC) of 282,824 unique sites was produced by selecting random positions in proximity of these restriction sites in human genome build 19 (hg19). Sequences extracted from GenBank (http://www.ncbi.nlm.nih.gov/genbank/) or obtained from the authors of prior studies yielded datasets for the remaining 11 viruses that encompassed from ~170 to 336,000 unique integration sites (Table 1).

Table 1 Retroviruses included in this study

Virus (genus)	TSD (bp)	Reference(s)	Number of sequences[a]
PFV (spuma)	4	[92,93]	2,924
MoMLV (gamma)	4	[44]	53,463
PERV (gamma)	4	[94]	1,668
XMRV (gamma)	4	[44,53]	5,487
EIAV (lenti)	5	[24,64]	1,172
HIV-1 (lenti)	5	[87]	335,968
Rev-A (gamma)	5	This study	834
SIV (lenti)	5	[95,96]	168
ASLV (alpha)	6	[97-99]	916
HERV-K (beta)	6	[100]	1,071
HTLV-1 (delta)	6	[17,101,102]	6,820
MMTV (beta)	6	[103]	178,574

[a] All sequences derived from virus-infected cells.

Sequence logos [67] were compiled to provide a visualization of base preferences at each nucleotide position at and surrounding the points of vDNA joining for the 12 different viruses (see Figures 1, 2 and 3, A-D). The position of vDNA insertion on the tDNA plus strand by convention was designated as 0, with upstream and downstream nucleotide positions extending in the negative and positive directions, respectively. As the most significant retroviral tDNA base preferences exist within and closely adjacent to the TSD [42] (the boundaries of which are marked by blue arrows in the sequence logos), we initially limited our analysis to base positions −5 through +9, or a total of 15 nucleotides. The statistical significance of nucleotide frequencies for each virus at each tDNA base position, which was compared to the MRC that was generated for the Rev-A integration site analysis, is presented in Additional file 1: Figure S1. These frequencies were calculated as observed-to-expected ratios and were thus normalized for human genomic DNA G/C content. The sequence logos by contrast should be primarily considered as visual aids because they depict raw base frequencies without normalization. The vast majority of positions within the 15 base windows displayed statistically significant variance from random across the viral integration sites (Additional file 1: Figure S1). For the dinucleotide step analysis, successive nucleotide positions were binned into groups of two; dinucleotide bin numbers are annotated below the sequence logo x-axes in Figures 1, 2 and 3, panels A-D. The frequencies of YR and RY dinucleotide usage at each bin position were compared to random frequencies using Fisher's exact test (Additional file 2: Figure S2).

Dinucleotide step analysis of viral integration sites with 4 bp TSDs

Symmetric base preferences project outward from the center of retroviral integration sites [14-17]. By convention, the center of an integration site with a 4 bp TSD is designated between nucleotide positions +1 and +2, which coincides with dinucleotide bin position +1 (Figure 1, vertical dotted line). As expected, PFV integration selected for flexible YR dinucleotides at this position [12] (Figure 1E and Additional file 2: Figure S2A). Perusal of the sequence logos indicated unique nucleotide signatures across the integration sites of viruses that generate 4 bp TSDs

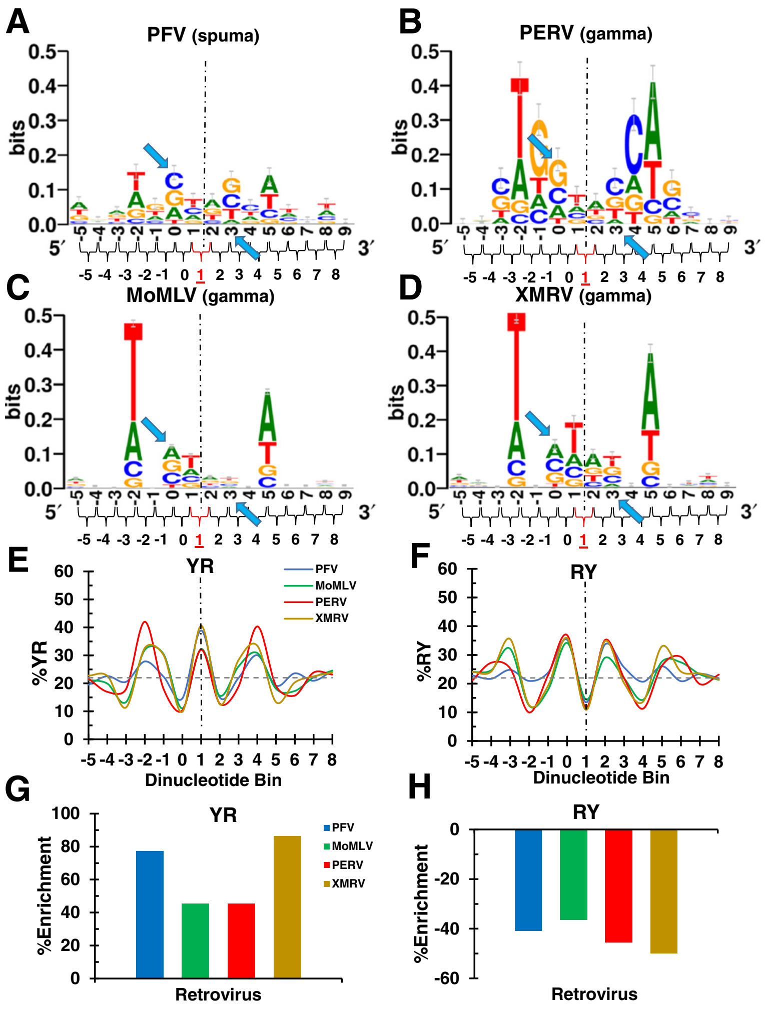


Figure 1 Sequence logos and dinucleotide step analysis of integration sites with 4 bp TSDs. **(A-D)** The height of each individual base at a given position is proportional to the frequency of the corresponding nucleotide within the sequences represented by the logos, and the height of each stack of base logos reflects the level of conservation at that position. **(E)** Percent YR utilization across the integration sites from panels **A-D** is shown relative to the calculated random value of 22% (dotted gray horizontal line). **(F)** Same as in panel **E**, except that the graph depicts RY utilization across the integration sites. Statistical analysis of panel **E** and **F** results are shown in Additional file 2: Figure S2 panels **A** and **B**, respectively. **(G and H)** The percent of YR (panel G) and RY (panel H) enrichment for each virus compared to random.

(Figure 1A-D). Nevertheless, commonalities among nucleotide and dinucleotide content were evident across these sites. Within the TSD, thymidine and adenosine were disfavored at symmetric positions 0 and +3, respectively. By contrast, outside of the TSD window T and A were preferred at symmetric positions −2 and +5, respectively (Additional file 1: Figure S1). Moreover, there was strong consensus for YR selectivity at the central dinucleotide step (Figure 1E and Additional file 2: Figure S2A). The calculated random frequency of YR dinucleotide occurrence in the human genome is 22% (Figure 1E, grey dashed horizontal line). YR utilization at bin position +1 by PFV, MoMLV, PERV, and XMRV were on average increased by 63% relative to this value (Figure 1G), equating to highly significant differences (*P* values ranging from 5 x 10^{-21} for PERV to >2.2 x 10^{-308} for MoMLV; Additional file 2: Figure S2A). On the contrary, all four viruses displayed strong selection against central RY steps, with an average value depressed by 43% relative to the expected value (Figure 1F, H, and Additional file 2: Figure S2B; *P* values of 1.9 x 10^{-27} for PERV to >2.2 x 10^{-308} for MoMLV).

Dinucleotide step analysis of viruses that yield 5 bp TSDs

Because viruses like HIV-1 join their vDNA ends across an odd number of tDNA bp, the center of their integration sites falls squarely on nucleotide position +2, which is a common element of dinucleotide bins +1 and +2 (Figure 2A-D). As recently elucidated the consensus sequence (0)RYXRY(+4), which resides at the center of HIV-1 integration sites [33], ensures for either YR or YY

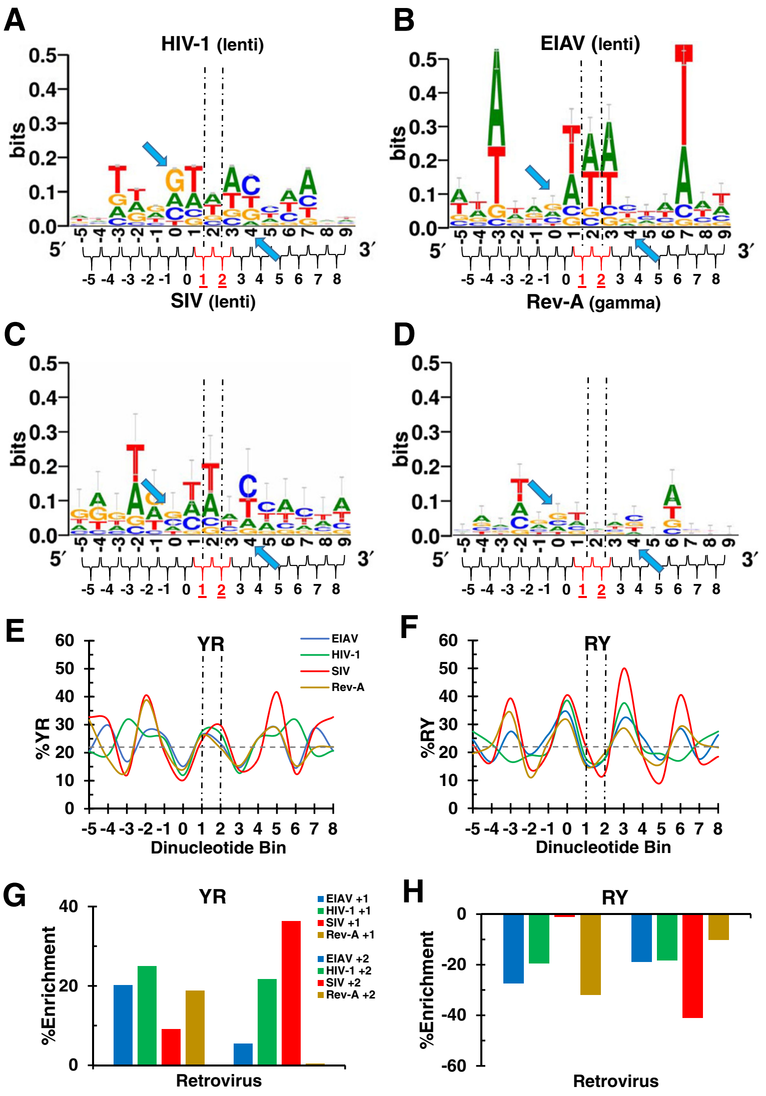


Figure 2 Sequence logos and dinucleotide step analysis of integration sites from retroviruses that yield 5 bp TSDs. Sequence logos for HIV-1 **(A)**, EIAV **(B)**, SIV **(C)**, and Rev-A **(D)**. **(E)** YR step analysis for the integration sites depicted in panels **A-D**. **(F)** Same as in panel **E**, except RY dinucleotide frequencies were calculated. Statistical analyses of panel **E** and **F** results are depicted in Additional file 2: Figure S2 panels **C** and **D**, respectively. Percent YR and RY enrichment for each virus compared to random is shown in **(G)** and **(H)**, respectively. Other labeling is as in Figure 1.

at nucleotide positions +1 and +2, and for YR or RR at nucleotide positions +2 and +3. Therefore, HIV-1 on average selects for a flexible YR step at one of the two central dinucleotide positions while strongly selecting against rigid RY steps. The tDNA sequences surrounding the integration sites of viruses that yield 5 bp TSDs were generally dissimilar from one another (Figure 2A-D). However, as was the case for the viruses that yield 4 bp TSDs, commonalities were evident among the sites that harbor 5 bp TSDs. HIV-1, EIAV, SIV, and Rev-A significantly disfavored T/A bases at the positions that delineate the external boundaries of the TSD, which in this case is positions 0 and +4, respectively (Additional file 1: Figure S1). All four viruses also displayed the preference for (0)RYXRY(+4) at the duplicated region and the symmetric preference for T/A adjacent to the TSD that was exhibited by 4-bp duplicating viruses. The positioning of this preference relative to the TSD window however varied among this set of viruses, falling two bases exterior to the TSD for SIV and Rev-A but three bases exterior for HIV-1 and EIAV (Additional file 1: Figure S1; Figure 2A-D).

Perhaps reflecting the fact that the center of these integration sites is spread over two dinucleotides, variable preference for YR/RY selectivity was apparent at the two central positions. At dinucleotide bin position +1, EIAV, HIV-1, and Rev-A each exhibited a similar enrichment for YR utilization, with an average relative

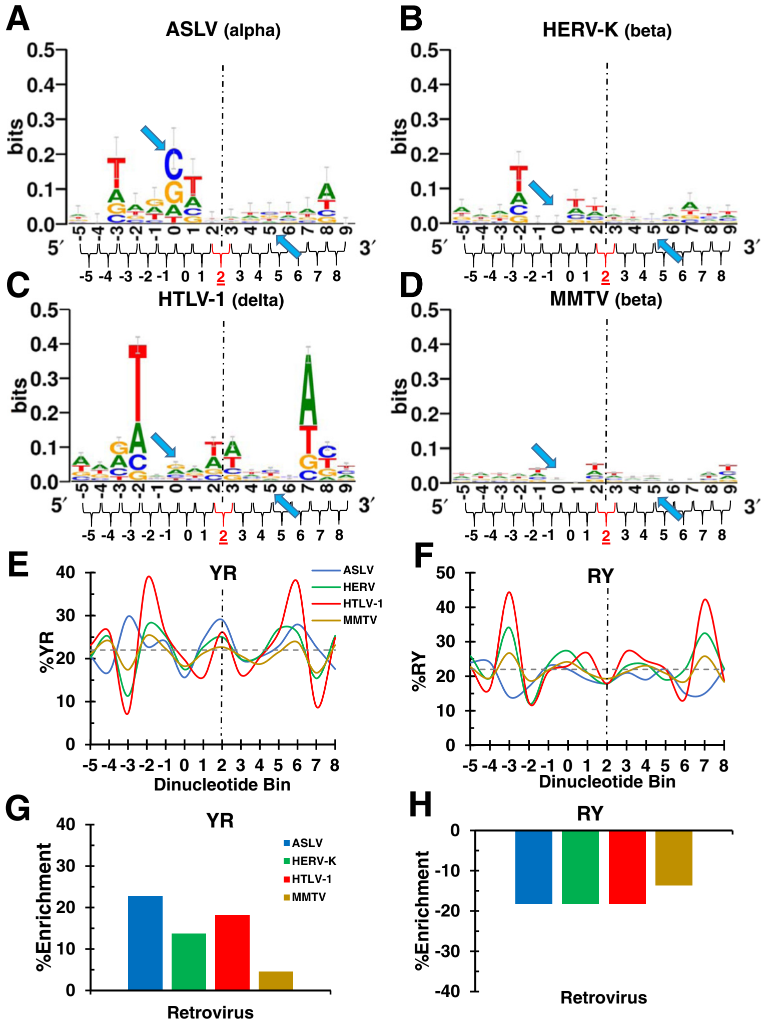


Figure 3 Sequence logos and YR/RY dinucleotide selectivities of viral integration sites with 6 bp TSDs. Sequence logos are shown for conglomerate integration sites of ASLV **(A)**, HERV-K **(B)**, HTLV-1 **(C)**, and MMTV **(D)**. **(E)** YR frequency utilization across the integration sites of viruses depicted in panels **A**-**D**. **(F)** Same as in panel **E**, except the plot is for RY dinucleotide utilization. Statistical analyses of panel **E** and **F** results are shown in Additional file 2: Figure S2 panels **E** and **F**, respectively. The percent YR and RY enrichment for each virus compared to random is in **(G)** and **(H)**, respectively. Other labeling is as in Figure 1.

increase of ~20% from the random value (Figure 2E, G, and Additional file 2: Figure S2C; *P* values from 5 x 10^{-3} for Rev-A to >2.2 x 10^{-308} for HIV-1). Although SIV also trended toward YR enrichment at bin position +1, this increase was not statistically different from random (Additional file 2: Figure S2C, *P* = 0.46). As reported [33], the enrichment for YR utilization at bin position +2 by HIV-1 was statistically significant (Figure 2E, G, and Additional file 2: Figure S2C; *P* >2.2 x 10^{-308}). While EIAV and SIV also trended toward YR enrichment at bin position +2, the difference only achieved statistical significance for SIV (*P* = 0.02). Rev-A by contrast did not exhibit YR enrichment at dinucleotide bin position +2. In terms of RY selectivity (Figure 2F, H, and Additional file 2: Figure S2D), EIAV and HIV-1 similarly avoided the rigid step at bin positions +1 and +2, averaging ~20% decreases from random (Additional file 2: Figure S2D, *P* values ranging from 5 x 10^{-4} for EIAV at bin position +2 to >2.2 x 10^{-308} for HIV-1 at both positions). While trending toward selection against RY at both dinucleotide positions, Rev-A registered as statistically different from random only at position +1 (*P* = 10^{-6}) while SIV registered as different only at position +2 (*P* = 0.002).

Dinucleotide step analysis of viral integration sites with 6 bp TSDs

The center of the integration site for viruses that yield 6 bp TSDs lies between nucleotide positions +2 and +3, which coincides with dinucleotide bin position +2 (Figure 3A-D). These integration sites on average yielded sequence logos with lower information content scores than the viruses that create 4 bp and 5 bp TSDs (compare Figure 3 to Figures 1 and 2). As with the previously discussed integration sites, T/A tended to be disfavored at the inner edges of the TSD window, though this was more evident for the sites generated by ASLV, HERV-K, and HTLV-1 than it was for MMTV (Additional file 1: Figure S1). ASLV, HERV-K, and HTLV-1 also revealed preference for T/A outside of the TSD window, two bases removed from the window for HERV-K and HTLV-1 but three bases removed for ASLV (Additional file 1: Figure S1),

All four viruses exhibited a significant enrichment for YR utilization at the center of their integration sites (Figure 3E, G, and Additional file 2: Figure S2E; *P* values ranged from 0.02 for HERV-K to 1.1 x 10^{-15} for HTLV-1). The selection against RY utilization at dinucleotide bin position +2 by these viruses was also significant (Figure 3F, H), yielding *P* values that ranged from 3.2 x 10^{-4} for HERV-K to 8.9 x 10^{-110} for MMTV (Additional file 2: Figure S2F).

Target DNA base preferences and central flexibility are determined by IN independent of nucleosome content

Integration sites on nucleosomal tDNA map to positions of DNA major groove distortion *in vitro* [35,37-39] and during virus infection [41-44]. Prior work with PFV [46] and HIV-1 [33] revealed that similar bases were selected in cells and when using the respective purified IN protein with naked tDNA *in vitro*, implying that nucleosome structure may not grossly influence the selection of tDNA bases at the sites of vDNA joining. However, the naked tDNA used in these studies was supercoiled plasmid with relatively low sequence diversity and little-to-no capacity to position native nucleosomes. Therefore, we accessed a panel of 22,117 unique integration sites from a reaction that utilized recombinant PFV intasomes and deproteinized human DNA [39], which served as an optimally diverse, nucleosome-free tDNA substrate.

The window of sequence logo analysis was extended from 15 bp (Figures 1, 2 and 3) to 50 bp (Figure 4) to assess signature tDNA sequences preferentially bound by nucleosomes, which show on average a 10.6 bp periodicity for A/T-rich sequences [68] (Figure 4A). Comparing integration sites derived from PFV infected cells (Figure 4B) to those generated *in vitro* with deproteinized cellular DNA (Figure 4C) confirmed that the preference for local tDNA sequences at the sites of virus insertion were in large part generated independent of nucleosome content. Re-analyzing 122 previously-reported *in vitro* concerted integration events [33] to sites derived from virus infected cells (Table 1) recapitulated the finding that the preference for local tDNA sequence at the sites of HIV-1 integration was independent of nucleosome content (Figure 4D, E). Both PFV and HIV-1 cell-based datasets exhibited cyclical A/T-rich sequences that extended symmetrically outward from the TSD with approximate 10 bp periodicity (Figure 4B, D), as described previously for HIV-1 [42]. These cyclical base preferences, which were absent from *in vitro* datasets (Figure 4C, E), and reminiscent of the A/T-rich periodicity exhibited by nucleosome-bound DNA (Figure 4A), indicated that PFV and HIV-1 IN select for their preferred local tDNA sequences in the context of nucleosomal DNA during virus infection [41,42] (Figure 4).

PFV and HIV-1 selected for marginally distinguishable flexibility profiles at integration sites in naked tDNA *in vitro* versus cellular DNA (Figure 5). As discussed above, raw frequencies of YR enrichment and RY avoidance for PFV at dinucleotide +1 equated to 39% and 13%, respectively (Figure 5A, B, blue curves). These values corresponded to a 77% increase in YR utilization and a 41% decrease in RY utilization relative to the MRC values (Figure 5C, D). The bias for YR utilization and against RY utilization at the center of integration sites was marginally greater when using recombinant PFV IN and naked cellular DNA than they were for virus-infected cells. Specifically, IN selected for YR and RY frequencies of 43% and 12% (Figure 5A, B), equating to a 95% increase and a 45% decrease from random, respectively (Figure 5C, D). These same trends also applied to HIV-1. Raw YR frequencies at central bins +1 and +2 were 27%/27% for virus and 32%/32% for recombinant IN protein (Figure 5E), and RY frequencies were 18%/18% for virus and 14%/14% for recombinant IN (Figure 5F). Comparing these raw frequencies to the MRC, YR was enriched by 23%/23% for virus and 45%/45% for recombinant IN, while RY was avoided by 18%/18% for virus and 36%/36% for recombinant IN (Figure 5G, H).

Genomic distribution of retroviral integration sites

Using various parameters linked to integration that include IN amino acid sequence, targeting of cellular chromatin features, and length of TSD, prior studies have phylogenetically linked subgroups of retroviral genera together [17,64]. We recently questioned the general applicability of this approach, as MoMLV and Rev-A, which are both gammaretroviruses, display similar tDNA base preferences but yield 4 and 5 bp TSDs, respectively [66]. It was therefore of interest to test if Rev-A integration distribution in cellular chromatin resembled that of MoMLV and/or other retroviruses. We accordingly mapped all of the integration sites used in this study, which included 834 unique sites from Rev-A-infected cells, with respect to several genomic annotations including RefSeq genes, CpG islands, TSSs, and gene density (Table 2). The statistical relevance of observed frequencies versus the MRC were determined by Fisher's exact test for RefSeq genes, CpG islands, and TSSs and

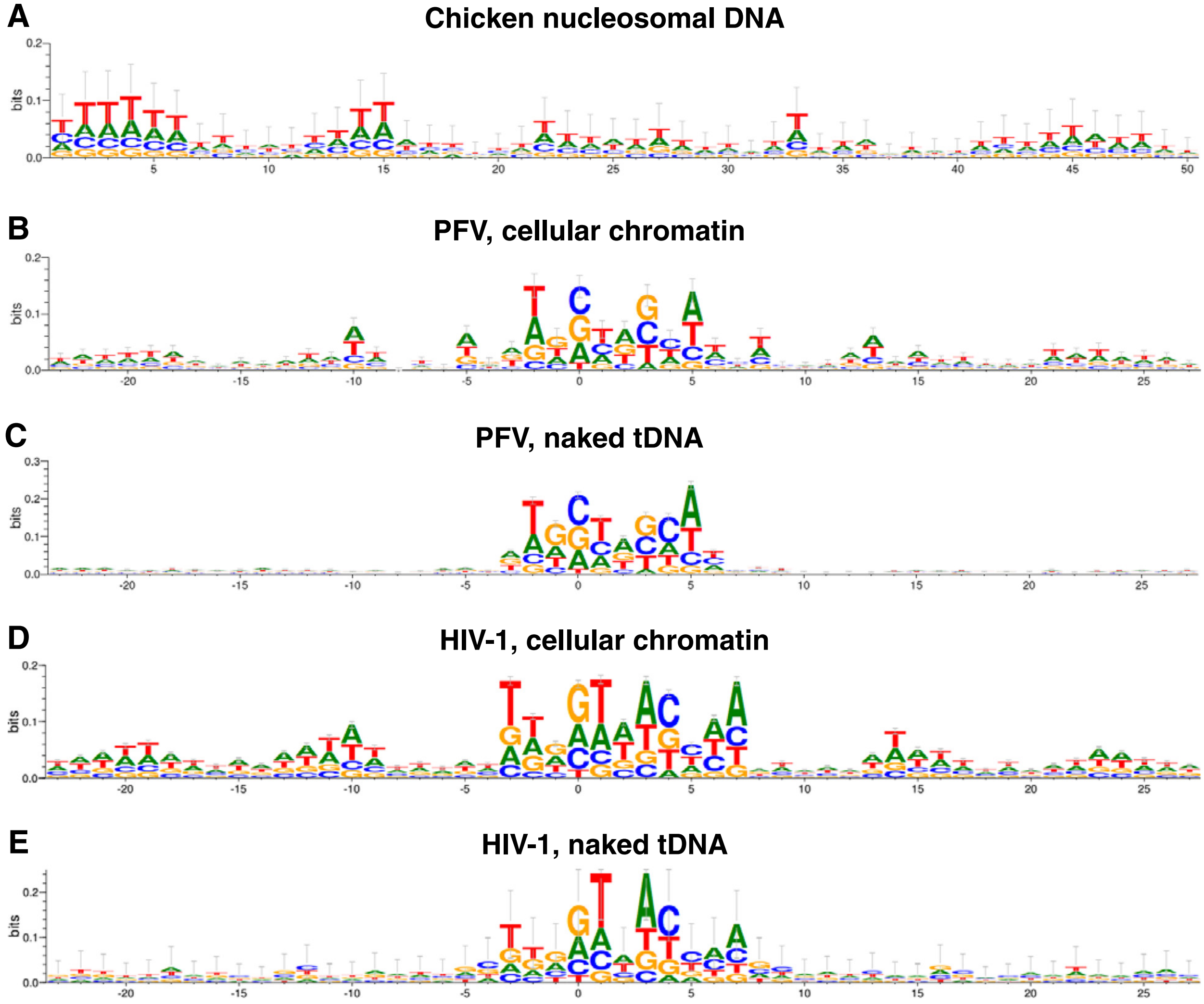


Figure 4 Sequence logos for PFV and HIV-1 integration sites in nucleosome-free versus chromatinized tDNA. **(A)** The logo illustrates the average nucleotide sequence of the first 50 nucleotides of center-aligned nucleosomal DNA sequences isolated from chicken erythrocytes [68]. **(B)** PFV integration sites derived from virus-infected cells [92,93]. **(C)** Integration sites from recombinant PFV intasomes and deproteinized cellular DNA. **(D)** HIV-1 integration sites from virus-infected cells [87]. **(E)** Concerted HIV-1 integration sites from recombinant HIV-1 IN and naked pGEM9Zf(–) plasmid DNA [33].

by Wilcoxon rank-sum test for gene density (*P* values listed in Additional file 3: Table S1).

The results of our analyses of 11 previously profiled retroviruses are in line with those of the prior reports (see Table 1 for the list of references). Distributions relative to CpG islands and TSSs were calculated by counting sites that fell within a 5 kb window (+/– 2.5 kb) of these features, while average gene density was calculated by counting the number of RefSeq genes falling within a 1 Mb window (+/– 500 kb) of each integration site, and then averaging this value for the entire dataset. Our MRC dataset revealed that 45.7% of human DNA comprised RefSeq genes (Table 2). Most of the viruses targeted RefSeq genes more frequently than this baseline value, with HIV-1 and SIV displaying the greatest levels of gene targeting (Table 2 and Additional file 3: Table S1). PFV and MMTV by contrast avoided this annotation, accomplishing only about 40% of their integrations within RefSeq genes ($P = 7.06 \times 10^{-09}$ for PFV and 3.51×10^{-282} for MMTV). MMTV also avoided CpG islands ($P = 9.94 \times 10^{-21}$) and TSSs ($P = 7.66 \times 10^{-62}$). All other viruses exhibited increased targeting of these genomic features over random, with the gammaretroviruses achieving the greatest levels. Over 40% of PERV integration sites mapped in the vicinity of a CpG island or a TSS ($P > 2.2 \times 10^{-308}$ for both comparisons to random). The average gene density surrounding MMTV integration sites (8.3 genes per Mb) was lower than the MRC value of 9.2 ($P > 2.2 \times 10^{-308}$) while all other viruses on average landed in regions that contained more genes per Mb than random. The average

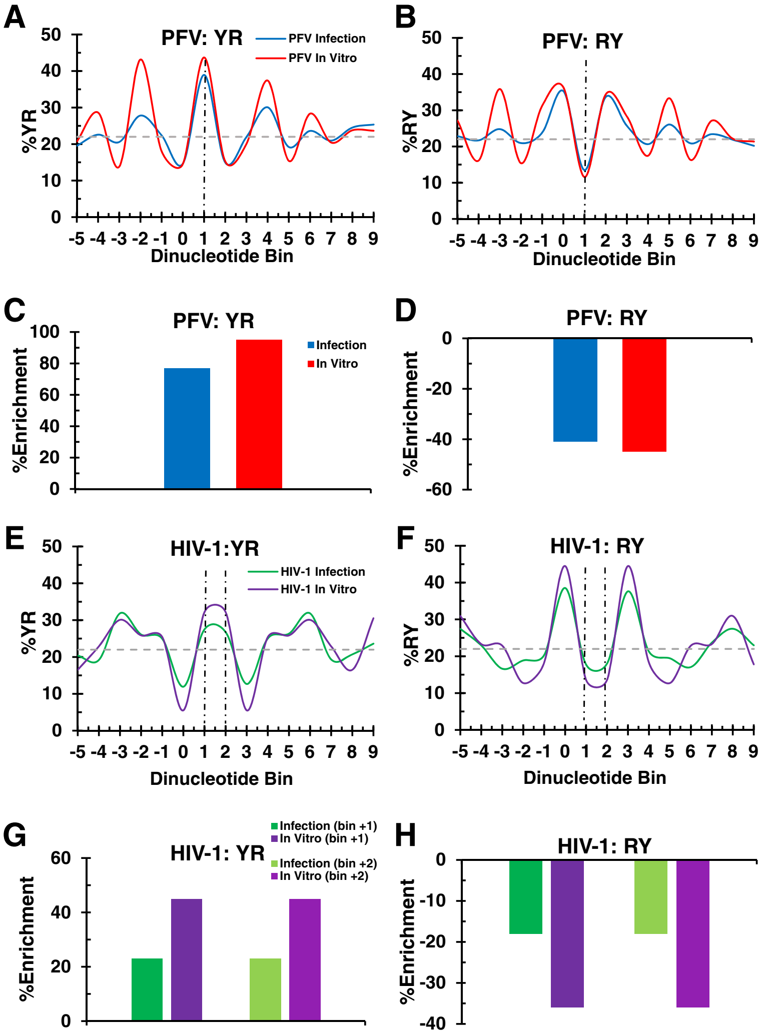


Figure 5 Flexibility profiles for PFV and HIV-1 integration sites in nucleosome-free versus chromatinized tDNA. **(A and B)** YR and RY frequency charts, respectively, for PFV integration sites into deproteinized genomic DNA (PFV *in vitro*) and from virus infection. Vertical dotted black line represents central dinucleotide step(s), and horizontal dotted grey line represents the MRC frequency of YR/RY utilization. **(C and D)**) Bar graphs illustrating the percent YR and RY enrichment, respectively, at the central dinucleotide step relative to MRC values. **(E and F)** YR and RY frequency charts, respectively, for HIV-1 integration sites into naked plasmid DNA (HIV-1 *in vitro*) and from virus infection. **(G and H)** Bar graphs illustrating the percent YR and RY enrichment, respectively, at the central dinucleotide steps compared to MRC.

gene densities selected by HTLV-1 and PFV, 10.7 and 10.8 genes/Mb, respectively, were only marginally greater than random (P = 3.33 × 10^{-40} and 2.95 × 10^{-24}, respectively), while the 23.4 genes/Mb value selected by PERV was the greatest among the viruses analyzed (P > 2.2 × 10^{-308}). As expected, the ability for PFV IN to target chromatin-specific features during integration was decreased significantly when the reaction was conducted with deproteinized cellular tDNA *in vitro* (Table 2; P values tabulated in Additional file 4: Figure S3).

Rev-A integrated within RefSeq genes at a frequency of 55.2%, which was significantly more frequent than the MRC value of 45.7% yet less frequent than the HIV-1 value of 74.6% (Table 2 and Additional file 5: Figure S4; P = 5.22 × 10^{-8} and 9.08 × 10^{-34}, respectively). Notably, this frequency was not significantly different from the other analyzed gammaretroviruses (P value range of 0.14 for MoMLV to 0.67 for PERV). The frequencies at which Rev-A and MoMLV targeted CpG islands were also statistically indistinguishable (25.9% vs. 28.4%, respectively; P = 0.12), yet were significantly different from the MRC value of 4.6% as well as the HIV-1 frequency of 5.6% (Table 2, Additional file 5: Figure S4). Approximately 25.9% of Rev-A integrations occurred within the 5 kb windows surrounding TSSs, a result that was again statistically indistinguishable from the MoMLV value of 27.8% (P = 0.23). The frequencies at which XMRV and PERV targeted CpG islands and TSSs were unique

Table 2 Genomic distribution of retroviral integration sites

Library	Unique sites	Within Refseq genes (%)	Within 5 kb (+/− 2.5 kb) of CpG (%)	Within 5 kb (+/− 2.5 kb) of TSS (%)	Average gene density per Mb (+/− 0.5 Mb) of integration sites
PFV	2,924	1,180 (40.4)	450 (15.4)	425 (14.5)	10.8
PFV (*in vitro*)	22,117	10,506 (47.5)	1,349 (6.1)	1,150 (5.2)	10.3
MoMLV	53,463	30,865 (57.7)	15,157 (28.4)	14,870 (27.8)	15.4
PERV	1,668	936 (56.1)	811 (48.6)	676 (40.5)	23.4
XMRV	5,487	3,086 (56.2)	1,022 (18.6)	1,100 (20.1)	14.5
EIAV	1,172	689 (58.8)	70 (6.0)	69 (5.9)	12.1
HIV-1	335,968	250,552 (74.6)	18,871 (5.6)	13,882 (4.1)	19.9
Rev-A	834	460 (55.2)	216 (25.9)	216 (25.9)	15.7
SIV	168	142 (84.5)	4 (2.4)	3 (1.8)	17.8
ASLV	916	320 (54.7)	55 (9.4)	44 (7.5)	11.3
HERV-K	1,071	541 (50.5)	131 (12.2)	108 (10.1)	17.8
HTLV-1	6,820	3148 (49.8)	523 (8.3)	467 (7.4)	10.7
MMTV	178,574	72,035 (40.3)	7,131 (4.0)	6,957 (3.9)	8.3
MRC	282,824	129,287 (45.7)	12,914 (4.6)	13,942 (4.9)	9.2

among the viruses studied, including both MoMLV and Rev-A (Table 2, Additional file 5: Figure S4). Rev-A, MoMLV, and XMRV integrated into similarly gene-dense regions of chromatin, whereas the 23.4 gene/Mb value displayed by PERV was more similar to the 19.9 gene/Mb value exhibited by HIV-1 (Table 2, Additional file 5: Figure S4).

Rev-A IN interacts with and is catalytically stimulated by the BRD4 ET domain *in vitro*

The interaction between MoMLV IN and BET proteins (BRD2-4) in large part determines the promoter-proximal integration profile of this virus [27-29]. The amino acid sequence of the IN C-terminal region, WxϕxxpxxPLbϕbϕR (x, non-conserved position; ϕ, small hydrophobic; p, small polar; b, basic), which dictates binding to BET proteins [29,69,70], is a conserved feature of gammaretroviral IN proteins including Rev-A IN [18]. The BET proteins comprise well-characterized bromodomain (BD) I and II and the ET domain, the latter of which accounts for IN binding (Figure 6A) [27-29,68]. To test if BET proteins interact with Rev-A IN, we expressed and purified hexahistidine-tagged IN and the C-terminal fragment of BRD4$_{462-720}$ that contains the ET domain (Figure 6A). As a control, we substituted glutamic acid for conserved residue Leu-630; the analogous L662E amino acid substitution in BRD2 negated binding to both MoMLV and feline leukemia virus IN proteins *in vitro* [28].

Utilizing a nickel-nitrilotriacetic acid (Ni-NTA) pull down format, Rev-A IN recovered input BRD4$_{462-720}$ protein from the solution but failed to bind detectable levels of BRD4$_{462-720/L630E}$ (Figure 6B). The BET proteins as well as their isolated ET domains can stimulate the concerted integration activity of MoMLV IN *in vitro* [27,28,70]. We accordingly assessed the ability of Rev-A IN to insert oligonucleotide vDNA substrates into supercoiled plasmid tDNA in the presence of BRD4$_{462-720}$ or BRD4$_{462-720/L630E}$. Two major types of integration products were expected under these reaction conditions [1,2,46,66]: the integration of one vDNA end molecule into one strand of plasmid DNA yields a tagged circle that co-migrates with nicked plasmid molecules in an agarose gel whereas the concerted integration of two vDNA ends yields a population of products that migrate as linearized plasmid molecules. As expected [66], Rev-A IN catalyzed a low level of single vDNA end and concerted integration activity in the absence of added BET protein (Figure 7A, compare lane 2 to lane 1). BRD4$_{462-720}$ stimulated IN concerted integration activity in a dose dependent manner, with an approximate fivefold boost in catalysis in the presence of 0.5 μM protein (Figure 7B; $P < 0.01$). On the contrary BRD4$_{462-720/L630E}$ did not appreciably stimulate Rev-A IN activity, even at the highest concentration tested (Figure 7).

Discussion

Different aspects of the nuclear environment, from global chromatin structure to local tDNA sequence, can influence where retroviruses integrate [18]. Prior work indicated that the forces that dictate these phenotypes may very well operate in an independent fashion. As example, the consensus palindromic sequence at sites of HIV-1 integration was unchanged in cells knocked out for expression of the dominant chromatin targeting factor LEDGF/p75 [24,25]. In this report we have expanded the analysis of local tDNA site determinants across a representative sampling of Retroviridae.

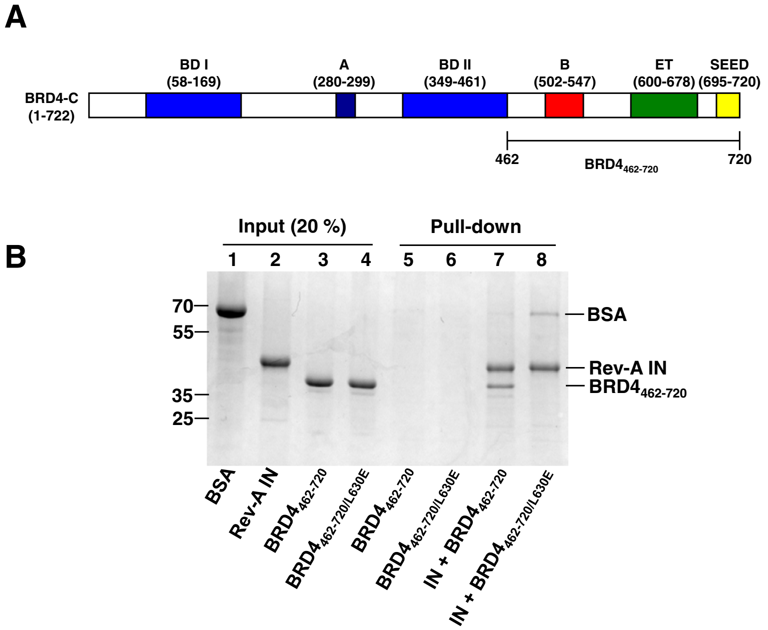


Figure 6 BRD4 protein and interaction with Rev-A IN. **(A)** Schematic of human BRD4 isoform C (NCBI reference sequence NP_055114.1) highlighting various protein domains and the 426–720 fragment used in this study. **(B)** Ni-NTA pull-down of purified BRD4$_{462–720}$ by His$_6$-tagged Rev-A IN. Migration positions of standards (in kDa) are labeled to the left. **Lanes 1–4**: 20% reaction input of the indicated proteins. **Lanes 5 and 6**: the indicated BRD4$_{462–720}$ protein was incubated with Ni-NTA beads in the absence of Rev-A IN. **Lanes 7 and 8**: the indicated BRD4$_{462–720}$ protein was incubated with Rev-A IN-bound beads. The gel is representative of results obtained from three independent experiments.

Central flexibility is a conserved feature of retroviral integration sites

Dinucleotide flexibility is arguably the most crucial sequence-dependent determinant of DNA conformation, with results of numerous studies confirming the specific role of YR steps in DNA bending [71-74]. Base stacking interactions play an even greater role in enforcing the conformation of the DNA double helix than Watson-Crick base pairing or phosphodiester backbone integrity [75-78]. Due to their relative lack of base overlap, YR dinucleotide steps possess the greatest level of inherent flexibility of the four purine/pyrimidine dinucleotides [32]. On the contrary, RY steps share the greatest surface area (RR and YY steps exhibit intermediate base overlap) and consequently display significantly lower propensities for roll, twist, slide, and other DNA bendability characteristics [71-73]. YR steps have accordingly been observed with near 50% frequency at sections of severe kinking in histone-wrapped DNA, while RY steps are by far the least represented at about 14% [79]. PFV selects for YR dinucleotides at the center of its integration sites [12] (Figure 1E), which facilitates IN-mediated minor groove compression of tDNA within the TCC to the point of central base pair unstacking. Our subsequent observation that central YR dinucleotides are enriched within HIV-1 integration sites [33] led us to extend the dinucleotide step analysis of integration sites to 10 additional retroviruses.

HIV-1, SIV, MoMLV, and ASLV integration sites were previously examined for physical properties related to DNA flexibility including A-philicity, protein-induced deformability, and bendability [15]. Both A-philicity and protein-induced deformability are based on dinucleotide frequencies. While all four sets of integration sites displayed similar A-philicity profiles, the protein-induced deformability profile was less dramatic for ASLV than for HIV-1, SIV, and MoMLV, which is consistent with the results reported here. Bendability scores – which are based on trinucleotide frequencies [80] – were higher for HIV-1 and SIV than for MoMLV or ASLV. By contrast, our findings indicate that tDNA is likely to undergo severe bending during MoMLV integration. Although this prior work importantly indicated the general bendable nature of retroviral integration sites, the YR/RY step analyses performed here pinpoint salient bendable tDNA phosphodiester bonds that contribute to vDNA integration. Quantitatively comparing levels of enriched flexibility at specific tDNA bonds using trinucleotide-based metrics may prove suboptimal, as only the integration sites of viruses that yield 5 bp TSDs contain a central trinucleotide. Furthermore, trinucleotide parameters as relating to bendability were originally deduced by probabilistic modeling of DNase I digestion data [80], which is a fairly indirect approach for classifying global levels of bendability for runs of nucleotides. To this point we tallied the frequency of all possible trinucleotide combinations falling at the

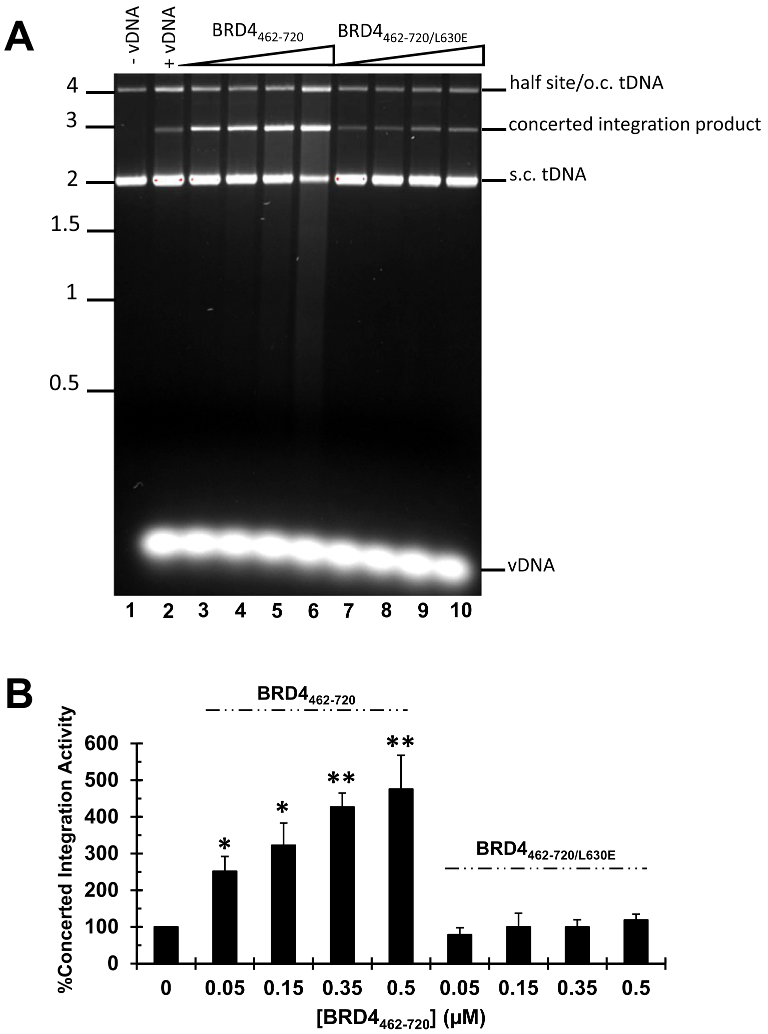


Figure 7 Concerted integration activity of Rev-A IN. **(A)** Ethidium bromide stained image of Rev-A IN integration reactions in the presence of increasing concentrations of indicated BRD4$_{462–720}$ protein. Migration positions of standards (in kb) are shown to the left, and the positions of single vDNA end, or half-site, and concerted vDNA integration products, as well as supercoiled (s.c.) and open circular (o.c) forms of the plasmid tDNA, are to the right. **Lanes 1 and 2**: Rev-A IN and tDNA incubated without vDNA or with vDNA, respectively. **Lanes 3–6:** increasing concentrations (0.05, 0.15, 0.35, 0.5 μM) of BRD4$_{462–720}$ incubated with Rev-A IN plus vDNA and tDNA. **Lanes 7–10**: same as lanes 3–6 but with BRD4$_{462–720/L630E}$. **(B)** Strand transfer activities for three independent experiments ± standard error of the mean, measured by quantification of DNA band intensity. The results were normalized to the level of Rev-A IN concerted integration activity in the absence of BRD4$_{462–720}$ proteins, which was set to 100%. *P* values of <0.05 and <0.01 are indicated by * and **, respectively, as determined by one-tailed t-test.

center of the integration sites for the viruses studied here that yield 5 bp TSDs (EIAV, HIV-1, Rev-A, and SIV), and found that A/T-rich trinucleotides (TAA/TTA, ATA/TAT, and AAA/TTT) were on average preferred by all four viruses compared to the MRC (Additional file 6: Figure S5). Because these sequences do not correlate well with DNase I-based trinucleotide flexibility classification [80], we propose that DNA bendability as applied to retroviral integration sites is better judged by YR/RY dinucleotide step analysis.

Based on the overall similar nature of the YR/RY histograms for the integration sites with 4 bp TSDs (Figure 1), we conclude that each of these viral intasomes is likely to accommodate tDNA with a severe central bend, akin to that observed in the PFV TCC crystal structure [12]. The two PFV IN active sites in the TCC are separated by approximately 26.3 Å, and the central dinucleotide accrues ~55° in negative roll to unstack the tDNA to position scissile phosphodiester bonds that are separated by 4 bp at these two positions [12]. The tDNA bend is considerably more extreme than at any position on nucleosomal DNA. A recent structural study revealed ancillary IN-histone and IN-DNA contacts on the sides of the tDNA-binding groove of the PFV intasome that help to lift nucleosomal DNA from the surface of the histone octamer for productive engagement between the IN active sites [39]. This explains why the nature of tDNA base preferences in the immediate vicinity of the integration site is independent of chromatinization (Figures 4 and 5), and also why nucleosomes may dampen the extent of such preferences (Figure 5).

Our prior work with HIV-1 indicated that the central positions of these integration sites are also likely to kink severely [33]. Due to the odd number of intervening bp, two adjacent tDNA dinucleotide steps seemingly collaborate to elicit the bend required for HIV-1 integration. While the other analyzed viruses that yield 5 bp TSDs in general displayed similar YR/RY histograms as HIV-1, our findings indicate that in these cases one of two central steps contributes more significantly to tDNA deformation during integration. As examples, flexibility at dinucleotide bin position +2 was significantly more important for SIV integration, whereas Rev-A selected for flexibility signatures at dinucleotide bin position +1 (Figure 2E-H). Given the lack of tactile TCC structures, the reasons for this apparent asymmetry are unclear and, in the case of SIV, may be a consequence of the relatively limited number of sites analyzed (Table 1).

Our results indicate that viruses that yield 6 bp TSDs likely require less flexibility at the center of their integration sites than do the viruses that cut tDNA with a 4 bp or 5 bp stagger. Why would some retroviruses require less central flexibility than others? Because phosphodiester bonds apposed by 6 bp are separated by 20.4 Å in canonical B-form DNA, it seems possible that significantly less deformation is required to fit chromosomal DNA into the two active sites of a viral intasome that generates a 6 bp TSD. The positions of the IN active sites in viral intasomes that yield 6 bp TSDs could also be further apart than those observed in PFV TCC structure. Furthermore, the extent of chromatin compaction can influence IN activity *in vitro* [81,82]. Viruses that yield 6 bp TSDs might accordingly display decreased affinity for nucleosomes, which are relative hotspots of YR dinucleotide content [68,79]. To assess if viruses such as ASLV that generate 6 bp TSDs display evidence for relatively nucleosome-depleted regions of chromatin during integration, we extended the boundaries of the tDNA sequence logos to encompass 50 nucleotides (Additional file 7: Figure S6A-C). Most of the analyzed integration sites, including those generated by ASLV, HERV-K, and HTLV-1 (Additional file 7: Figure S6C), showed evidence for nucleosome content adjacent to the local regions of TSD. MMTV and PERV by contrast failed to reveal evidence for periodic A/T-enriched peaks emanating outward from the integration sites. Based on this analysis, the ability to integrate into nucleosomal arrays does not seem to track with the size of TSD. We note that a report published after the submission of this paper indicated that MLV and PFV integration can occur in regions of compact chromatin that tend to disfavor HIV-1 and ASLV integration [45].

IN-tDNA interactions during retroviral integration

The modeling of thymidine for the preferred tDNA cytidine at the point of vDNA joining suggested that the methyl group at the C5 position of thymine may clash with the scissile phosphodiester bond, providing a structural account for why PFV avoids integration at T residues [12]. As each of the analyzed viruses – with the exception of MMTV – prominently disfavored insertion at a T residue (Additional file 1: Figure S1), this aspect of intasome nucleoprotein structure appears nearly universal. We additionally observed a symmetric preference for thymidine/adenosine at either two or three bases exterior to the TSD boundary across the majority of integration sites (Figures 1, 2 and 3, A-D). Through mutagenesis experiments we previously revealed that HIV-1 IN residue Ser119, which is structurally analogous to PFV IN residue Ala188, helps to determine the identity of tDNA bases three positions upstream from the points of vDNA insertion [33]. Neutral, compact amino acids occupy the analogous position of the CCD α2 helix across retroviruses (Additional file 8: Figure S7), and correlating these residues with tDNA sequence preferences suggests that the polarity of the amino acid side chain dictates IN-tDNA interactions. Specifically, non-polar residues alanine and proline dictate preference for thymidine/adenosine at two bases upstream/downstream of the vDNA insertion sites regardless of TSD length, while polar amino acids serine and threonine shift this preference one base further outward from the tDNA cut, to three bases upstream/downstream. Scrutiny of the corresponding YR/RY peaks reinforces that although serine and threonine dictate the T/A preference at the same relative tDNA position, the specifics of the nucleoprotein contacts are not identical. Serine yields the T/A preference at positions −3/+7 for HIV-1, while the S119T IN mutation [33] as well as the threonine that is naturally present in EIAV IN, switches this preference to A/T (Figure 2). Recent studies have determined integration sites of HIV-1 IN mutants that harbor all possible small amino acid substitutions for Ser119 in the CCD (S119A, S119T, S119P, S119G) [33,83]. Consistent with our observations, Ser119 as well as the Thr substituent were modeled to interact with bases at tDNA positions −3/+7 while Ala and Gly were modeled to interact with positions −2/+6 [83].

Rev-A integration distribution and BET proteins

Our results reveal that Rev-A shares similar integration site distribution patterns with other gammaretroviruses (Table 2). Of the four gammaretroviruses analyzed in this study, the targeting of chromatin-specific features was most similar between Rev-A and MoMLV (Table 2 and Additional file 5: Figure S4). Purified BRD4$_{462-730}$ interacted with Rev-A IN and stimulated its concerted integration activity, and the L630E amino acid substitution in

BRD4 counteracted the protein-protein interaction (Figures 6 and 7). Rev-A and MoMLV are thus likely directed to integrate into signature chromatin features using similar BET protein-IN mediated interactions. The unique chromatin targeting preferences of PERV and XMRV suggest that these viruses may interact with additional host factors to guide promoter-proximal integration events.

Conclusions

The work presented here clarifies that conserved palindromic sequences at sites of retroviral DNA integration reflect the requirement for a central tDNA bend and that the sharpness of the required deformation is reduced for viruses that generate 6 bp TSDs. It seems plausible that retroviruses have convergently evolved to select not necessarily a specific sequence of nucleotides at integration sites, but rather combinations of bases that yield a flexibility pattern that is favorable for tDNA incorporation into the TCC. Retroviral base preferences and associated flexibility profiles appear largely independent of tDNA chromatinization, suggesting that IN interactions with tDNA nucleotides dictate integration site selection on the local scale.

Methods

Plasmids, cell culture, and virus infection

A human BRD4 expression construct was purchased from Addgene (plasmid #14441). The region of BRD4 corresponding to amino acids 462–720 was PCR-amplified using primers AE5271 (5′-GATATACCCGGGGAGGAGCCAG TGGTGGCCGTG) and AE5270 (5′-GCAGCACTCGAG TTACTCTGTTTCGGAGTCTTC). The PCR product was cleaved with XhoI and XmaI, and ligated to XhoI/XmaI-digested pCPH6P [84]. The L630E amino acid substitution was introduced by site-directed mutagenesis (Stratagene QuickChange II kit) using primers AE5362 (5′-AAGCTCC CCGGCGAGAAGGAGGGCCGCGTGGTGCACATC) and AE5363 (5′-GATGTGCACCACGCGGCCCTCCTTC TCGCCGGGGAGCTT). Plasmids pCPH6P-RevA-IN [66], pJD215 [85], pSW253 [86], and pCG-VSV-G [25] were previously described.

HEK293T and HeLa cells were cultured in Dulbecco's Modified Eagle Media supplemented to contain 10% fetal bovine serum, 100 IU/mL penicillin, and 100 μg/mL streptomycin. HEK293T cells were transfected with pJD215, pSW253, and pCG-VSV-G at the ratio of 4.5:4.5:1 using PolyJet (SignaGen). After 48 h, the cell-free supernatant was filtered through 0.45 μm filters, concentrated by ultracentrifugation at 200,000 g for 1 h, and treated with 40 U/mL DNase Turbo (Ambion). Viral titer was determined by infecting 3 x 10^5 HeLa cells with two-fold dilutions of virus in the presence of 500 μg/mL G418 (Life Technologies) and counting neomycin-resistant colony forming units after 7 d. For integration site sequencing, HEK293T cells (5 x 10^6) were infected and treated similarly for 7 d to select for transductants and to allow for dissolution of unintegrated vDNA.

Expression and purification of recombinant proteins

His$_6$-tagged Rev-A IN was expressed in *Escherichia coli* and purified essentially as previously described [66], but without hexahistidine tag cleavage. BRD4$_{462-720}$ and BRD4$_{462-720/L630E}$ were expressed in *E. coli* strain PC2 [84] by overnight induction with 1 mM isopropyl-β-D-thiogalactopyranoside at 18°C. The bacterial pellets were dissolved in 50 mM Tris HCl, pH 7.5, 500 mM NaCl, and 1 mM phenylmethanesulfonylfluoride, lysed by sonication, and clarified by centrifugation at 40,000 g for 1 h. BRD4$_{462-720}$ and BRD4$_{462-720/L630E}$ were bound to a HisTrap HP column and eluted using a 20–500 mM imidazole gradient on an AKTA purifier liquid chromatography system (GE Healthcare). Fractions containing the protein of interest were pooled together and quickly diluted 5-fold with 50 mM Tris HCl, pH 7.5 to reduce the concentration of NaCl to 100 mM and then loaded on a HiTrap Heparin column and eluted with a 100–1000 mM NaCl gradient. Finally, the proteins were purified by gel filtration using a Superdex 200 column in 50 mM Tris HCl, pH 7.5, 500 mM NaCl, 2 mM dithiothreitol (DTT). Chromatography columns were purchased from GE Healthcare.

Pull-down assay

For *in vitro* Ni-NTA pull-down assays, 10 μg of His$_6$-tagged Rev-A IN in 100 μL pull-down buffer (25 mM Tris HCl pH 7.5, 150 mM NaCl, 25 μM ZnCl$_2$, 0.1% (v/v) Nonidet P40, 20 mM imidazole) was mixed with 10 μL settled volume of Ni-NTA beads (Thermo Scientific) previously washed with pull-down buffer. Following incubation at 4°C for 2 h with gentle agitation, 10 μg BSA and 10 μg of BRD4$_{462-720}$ or BRD4$_{462-720/L630E}$ were added, and mixtures were incubated overnight at 4°C. The beads were washed five times with pull-down buffer and briefly centrifuged for 1 min at 1,300 g, and were then resuspended in 20 μL 2X sodium dodecyl sulfate (SDS) gel loading buffer and boiled for 10 min. The resulting supernatant was analyzed by denaturing gel electrophoresis on a 10% acrylamide gel. Proteins were detected by staining with Coomassie blue.

DNA strand transfer activity assay

The *in vitro* concerted integration assays were carried out as previously for described Rev-A IN following removal of the His$_6$-tag by site-specific proteolysis [66]. Briefly, 1 μM Rev-A IN mixed with 0.5 μM vDNA and 4 nM pGEM-3 tDNA, in 40 μL of 20 mM HEPES pH 7.4, 50 mM NaCl, 5 mM MgCl$_2$, 4 μM ZnCl$_2$, and 10 mM DTT, was incubated for 1 h at 37°C with varying concentrations of BRD4$_{462-720}$

or BRD4$_{462–720/L630E}$ (0.05/0.15/0.35/0.5 μM). Reactions were stopped by adding 25 mM EDTA and 0.5% SDS, and deproteinized by digestion with proteinase K. Products were precipitated with ethanol and analyzed by electrophoresis through 1.5% agarose gels. DNA was detected by staining with ethidium bromide. Concerted integration products were measured by band intensity quantification using the Molecular Imager® Gel Doc TM XR+ System with Image Lab software (Bio-Rad).

Sequencing of Rev-A integration sites

Genomic DNA (20 μg) isolated using the DNeasy Blood and Tissue Kit (Qiagen) was digested overnight with AvrII, NheI, and SpeI and purified using the QIAquick PCR Purification Kit (Qiagen). A double-stranded asymmetric linker was prepared by annealing 10 μM of oligonucleotides AE5237 (5′-[Phosp]CTAGGCAGCCCG[Am C7-Q]-3′) and AE5238 (GTAATACGACTCACTATAG GGCACGCGTGGTCGACGGCCCGGGCTGC) by heating to 90°C in 10 mM Tris–HCl, pH 8.0 and 0.1 mM EDTA and slowly cooling to room temperature. Linker DNA (1.5 μM) was ligated with digested cellular DNA (1 μg) overnight at 16°C in four parallel reactions, and the DNAs were pooled and re-purified using the QIAquick PCR Purification Kit. Nested PCR was used to selectively amplify integration sites, with reactions multiplexed into eight separate samples per PCR stage. First- and second-round linker primers were AE5240 (5′-GACTCACTATA GGGCACGCGT) and AE5242 (5′-GTCGACGGCCCGG GCTGCCTA), and first- and second round Rev-A U5 primers were AE6121 (5′-GCAGGGATCCGGACTG) and AE6122 (5′-CCGTAGTACTTCGGTACAAC), respectively. PCRs were incubated at 94°C for 2 min, followed by 30 cycles at 94°C for 15 sec, 55°C for 30 sec, and 68°C for 45 sec, which was followed by a final extension for 10 min at 68°C. Pooled PCRs were purified using the QIAquick PCR Purification Kit, and standard Illumina adapters were ligated onto the amplicons prior to sequencing on the Illumina MiSeq platform at the Data-Farber Cancer Institute Molecular Biology Core Facilities. Sequences were mapped to hg19 version of human genome using BLAT, ensuring that the genomic match starts immediately after TACTTCGG-TACAACA sequence, which corresponds to the processed Rev-A U5 end. Bioinformatics analysis of Rev-A integration sites was performed as described previously [87].

Statistical analysis

An MRC dataset of 282,824 sites was created by selecting random genomic positions in proximity (<500 bp) of a AvrII, NheI or SpeI recognition site. The sequence immediately abutting each random site, truncated to 97 bp or less (to simulate Illumina read length), was subjected to the genomic alignment procedure described above. Differences in nucleotide sequence from random among retroviral integration sites were determined by chi-square analysis relative to the above random control. Statistical differences with respect to YR/RY frequency plots and genomic distribution of integration sites was calculated by Fisher's Exact Test using *R* [88].

Additional files

Additional file 1: Figure S1. Nucleotide preferences at tDNA base positions surrounding retroviral integration sites. The number in parentheses indicates the number of unique integration sites analyzed for each virus. Chi-square analysis was used to calculate *P* values by comparing observed nucleotide frequencies to the expected frequency based on 282,824 random points in the human genome. Significant differences are marked by blue shade and bold character ($P < 0.05$). Nucleotide frequencies that differed from the neutral value of 100% by more than 40 are highlighted in red (<60%) and green (>140%).

Additional file 2: Figure S2. Statistical analysis of YR/RY dinucleotide frequencies across retroviral integration sites. (A and B) *P* values calculated by Fisher's Exact Test for statistical comparison of YR and RY dinucleotide frequency profiles as compared to the MRC of 282,824 randomly-generated sites for viruses that yield 4 bp TSDs. (C and D) Statistical analysis of dinucleotide frequencies for viruses that yield 5 bp TSDs. (E and F) Analysis of dinucleotide frequency statistics for viruses with 6 bp TSDs. The dotted gray horizontal line in all panels demarcates the statistical cutoff value of 1.3 ($-\log_{10}(0.05)$). Curve flattening results from the statistical cutoff of 2.2×10^{-308} of the utilized statistical package [88]. Asterisks denote curves graphed on the secondary y-axis to the right of the charts.

Additional file 3: Table S1. *P* values for comparison of retroviral integration site distributions versus MRC. Values > 0.05 are highlighted in bold italics.

Additional file 4: Figure S3. *P* values for comparison of PFV integration site distribution in deproteinized tDNA and from virus-infected cells to the MRC dataset. Counts of integration sites within RefSeq genes and relative to CpG islands and TSSs, as well as regional gene densities, are listed in Table 2.

Additional file 5: Figure S4. *P* values for comparison of Rev-A integration site distribution to other gammaretroviruses, HIV-1, PFV, and to the MRC dataset. Numbers of integration sites within RefSeq genes and nearby CpG islands and TSSs, as well as regional gene density profiles, are listed in Table 2.

Additional file 6: Figure S5. Trinucleotide frequencies at centers of 5 bp TSDs. The frequency of all possible trinucleotides residing at the central three bases of EIAV, HIV-1, Rev-A, and SIV integration sites was computed and compared to the random distribution given by the MRC dataset. Y-axis values represent the percent increase in usage over random. The x-axis displays the most frequently-utilized trinucleotides arranged from left to right.

Additional file 7: Figure S6. Extended sequence logos depicting base preferences at a total of 50 bases surrounding retroviral integration sites. The Y-axis scales were set to 0.2 bits for all logos in order to highlight T/A periodicity, and thus some of the overly prominent base preferences at the centers of the integration sites are slightly obscured. (A) Sequence logos for viruses that yield 4 bp TSDs are compared to the average sequence of chicken erythrocyte nucleosomal DNA and to the *in vitro* dataset of recombinant PFV integration sites. (B) Same as in panel A, except the analyzed viruses generate 5 bp TSDs. The *in vitro* HIV-1 integration dataset is from ref. [33]. (C) Extended sequence logos for viruses that yield 6 bp TSDs, compared to the average nucleosome content of chicken DNA.

Additional file 8: Figure S7. Amino acid sequence alignment of CCD α2 helix residues for the viruses analyzed in this study. The location of secondary structural elements has been documented crystallographically for HIV-1 [89], ASLV [90], SIV [91], and PFV [2] INs. Residues analogous to Ala188 in PFV and Ser119 in HIV-1 INs are highlighted in yellow. The active site residue that is analogous to Asp185 in PFV IN and Asp116 in HIV-1 IN is in red type.

Abbreviations

IN: Integrase; vDNA: Viral DNA; TCC: Target capture complex; tDNA: Target DNA; TSD: Target site duplication; TSS: Transcriptional start site; LEDGF: Lens epithelium-derived growth factor; BET: Bromodomain and extraterminal domain; MoMLV: Moloney murine leukemia virus; MMTV: Mouse mammary tumor virus; PFV: Prototype foamy virus; CCD: Catalytic core domain; Y: Pyrimidine; R: Purine; HIV-1: Human immunodeficiency virus type 1; Rev-A: Reticuloendotheliosis virus strain A; ET: Extraterminal; PERV: Porcine endogenous retrovirus; XMRV: Xenotropic murine leukemia virus-related virus; ASLV: Avian sarcoma-leukosis virus; HTLV-1: Human T-lymphotropic virus type 1; HERV-K: Human endogenous retrovirus K family; SIV: Simian immunodeficiency virus; EIAV: Equine infectious anemia virus; MRC: Matched random control; hg19: Human genome build 19; BD: Bromodomain; Ni-NTA: Nickel-nitrilotriacetic acid; DTT: Dithiothreitol; SDS: Sodium dodecyl sulfate.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

ES, AB-C, GNM, PC, and ANE conceived the study, analyzed the data, and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Joseph Dougherty for the generous gift of pJD215 plasmid DNA. We additionally thank the following individuals for graciously sharing with us the coordinates of the following retroviral integration site datasets: Frederic Bushman, MoMLV and EIAV; Charles Bangham, HTLV-1, Jeroen de Ridder, MMTV. This work was supported by US National Institutes of Health grants AI070042 (to A.E. and P.C.) and AI060354 (Harvard University Center for AIDS Research).

Author details

[1]Department of Cancer Immunology and AIDS, Dana-Farber Cancer Institute, Boston, MA, USA. [2]Division of Infectious Diseases, Imperial College London, London, UK. [3]Clare Hall Laboratories, The Francis Crick Institute, London, UK.

Received: 1 January 2015 Accepted: 20 April 2015
Published online: 30 April 2015

References

1. Li M, Mizuuchi M, Burke Jr TR, Craigie R. Retroviral DNA integration: reaction pathway and critical intermediates. EMBO J. 2006;25:1295–304.
2. Hare S, Gupta SS, Valkov E, Engelman A, Cherepanov P. Retroviral intasome assembly and inhibition of DNA strand transfer. Nature. 2010;464:232–6.
3. Hare S, Maertens GN, Cherepanov P. 3′-processing and strand transfer catalysed by retroviral integrase in crystallo. EMBO J. 2012;31:3020–8.
4. Fujiwara T, Mizuuchi K. Retroviral DNA integration: structure of an integration intermediate. Cell. 1988;54:497–504.
5. Roth MJ, Schwartzberg PL, Goff SP. Structure of the termini of DNA intermediates in the integration of retroviral DNA: dependence on IN function and terminal DNA sequence. Cell. 1989;58:47–54.
6. Brown PO, Bowerman B, Varmus HE, Bishop JM. Retroviral integration: structure of the initial covalent product and its precursor, and a role for the viral IN protein. Proc Natl Acad Sci U S A. 1989;86:2525–9.
7. Pauza CD. Two bases are deleted from the termini of HIV-1 linear DNA during integrative recombination. Virology. 1990;179:886–9.
8. Lee YM, Coffin JM. Relationship of avian retrovirus DNA synthesis to integration in vitro. Mol Cell Biol. 1991;11:1419–30.
9. Bowerman B, Brown PO, Bishop JM, Varmus HE. A nucleoprotein complex mediates the integration of retroviral DNA. Genes Dev. 1989;3:469–78.
10. Bukrinsky MI, Sharova N, Dempsey MP, Stanwick TL, Bukrinskaya AG, Haggerty S, et al. Active nuclear import of human immunodeficiency virus type 1 preintegration complexes. Proc Natl Acad Sci U S A. 1992;89:6580–4.
11. Miller MD, Farnet CM, Bushman FD. Human immunodeficiency virus type 1 preintegration complexes: studies of organization and composition. J Virol. 1997;71:5382–90.
12. Maertens GN, Hare S, Cherepanov P. The mechanism of retroviral integration from X-ray structures of its key intermediates. Nature. 2010;468:326–9.
13. Engelman A, Mizuuchi K, Craigie R. HIV-1 DNA integration: mechanism of viral DNA cleavage and DNA strand transfer. Cell. 1991;67:1211–21.
14. Holman AG, Coffin JM. Symmetrical base preferences surrounding HIV-1, avian sarcoma/leukosis virus, and murine leukemia virus integration sites. Proc Natl Acad Sci U S A. 2005;102:6103–7.
15. Wu X, Li Y, Crise B, Burgess SM, Munroe DJ. Weak palindromic consensus sequences are a common feature found at the integration target sites of many retroviruses. J Virol. 2005;79:5211–4.
16. Berry C, Hannenhalli S, Leipzig J, Bushman FD. Selection of target sites for mobile DNA integration in the human genome. PLoS Comput Biol. 2006;2, e157.
17. Derse D, Crise B, Li Y, Princler G, Lum N, Stewart C, et al. Human T-cell leukemia virus type 1 integration target sites in the human genome: comparison with those of other retroviruses. J Virol. 2007;81:6731–41.
18. Kvaratskhelia M, Sharma A, Larue RC, Serrao E, Engelman A. Molecular mechanisms of retroviral integration site selection. Nucleic Acids Res. 2014;42:10209–25.
19. Schroder AR, Shinn P, Chen H, Berry C, Ecker JR, Bushman F. HIV-1 integration in the human genome favors active genes and local hotspots. Cell. 2002;110:521–9.
20. Wu X, Li Y, Crise B, Burgess SM. Transcription start regions in the human genome are favored targets for MLV integration. Science. 2003;300:1749–51.
21. De Ravin SS, Su L, Theobald N, Choi U, Macpherson JL, Poidinger M, et al. Enhancers are major targets for murine leukemia virus vector integration. J Virol. 2014;88:4504–13.
22. LaFave MC, Varshney GK, Gildea DE, Wolfsberg TG, Baxevanis AD, Burgess SM. MLV integration site selection is driven by strong enhancers and active promoters. Nucleic Acids Res. 2014;42:4257–69.
23. Ciuffi A, Llano M, Poeschla E, Hoffmann C, Leipzig J, Shinn P, et al. A role for LEDGF/p75 in targeting HIV DNA integration. Nat Med. 2005;11:1287–9.
24. Marshall HM, Ronen K, Berry C, Llano M, Sutherland H, Saenz D, et al. Role of PSIP1/LEDGF/p75 in lentiviral infectivity and integration targeting. PLoS One. 2007;2, e1340.
25. Shun MC, Raghavendra NK, Vandegraaff N, Daigle JE, Hughes S, Kellam P, et al. LEDGF/p75 functions downstream from preintegration complex formation to effect gene-specific HIV-1 integration. Genes Dev. 2007;21:1767–78.
26. Schrijvers R, De Rijck J, Demeulemeester J, Adachi N, Vets S, Ronen K, et al. LEDGF/p75-independent HIV-1 replication demonstrates a role for HRP-2 and remains sensitive to inhibition by LEDGINs. PLoS Pathog. 2012;8, e1002558.
27. Sharma A, Larue RC, Plumb MR, Malani N, Male F, Slaughter A, et al. BET proteins promote efficient murine leukemia virus integration at transcription start sites. Proc Natl Acad Sci U S A. 2013;110:12036–41.
28. Gupta SS, Maetzig T, Maertens GN, Sharif A, Rothe M, Weidner-Glunde M, et al. Bromo- and extraterminal domain chromatin regulators serve as cofactors for murine leukemia virus integration. J Virol. 2013;87:12721–36.
29. De Rijck J, de Kogel C, Demeulemeester J, Vets S, El Ashkar S, Malani N, et al. The BET family of proteins targets moloney murine leukemia virus integration near transcription start sites. Cell Rep. 2013;5:886–94.
30. Faschinger A, Rouault F, Sollner J, Lukas A, Salmons B, Gunzburg WH, et al. Mouse mammary tumor virus integration site selection in human and mouse genomes. J Virol. 2008;82:1360–7.
31. Stevens SW, Griffith JD. Sequence analysis of the human DNA flanking sites of human immunodeficiency virus type 1 integration. J Virol. 1996;70:6459–62.
32. Johnson RC, Stella S, Heiss JK. Bending and compaction of DNA by Proteins. In: Rice PA, Correll CC, editors. Protein-nucleic acid interactions. London: RCS Publishing; 2008. p. 176–220.
33. Serrao E, Krishnan L, Shun MC, Li X, Cherepanov P, Engelman A, et al. Integrase residues that determine nucleotide preferences at sites of HIV-1 integration: implications for the mechanism of target DNA binding. Nucleic Acids Res. 2014;42:5164–76.
34. Muller HP, Varmus HE. DNA bending creates favored sites for retroviral integration: an explanation for preferred insertion sites in nucleosomes. EMBO J. 1994;13:4704–14.
35. Pruss D, Reeves R, Bushman FD, Wolffe AP. The influence of DNA and nucleosome structure on integration events directed by HIV integrase. J Biol Chem. 1994;269:25031–41.
36. Katz RA, Gravuer K, Skalka AM. A preferred target DNA structure for retroviral integrase in vitro. J Biol Chem. 1998;273:24190–5.

37. Pryciak PM, Varmus HE. Nucleosomes, DNA-binding proteins, and DNA sequence modulate retroviral integration target site selection. Cell. 1992;69:769–80.
38. Pruss D, Bushman FD, Wolffe AP. Human immunodeficiency virus integrase directs integration to sites of severe DNA distortion within the nucleosome core. Proc Natl Acad Sci U S A. 1994;91:5913–7.
39. Maskell DP, Renault L, Serrao E, Lesbats P, Matadeen R, Hare S, et al. Structural basis for retroviral integration into nucleosomes. Nature. 2015, in press.
40. Bor YC, Bushman FD, Orgel LE. In vitro integration of human immunodeficiency virus type 1 cDNA into targets containing protein-induced bends. Proc Natl Acad Sci U S A. 1995;92:10334–8.
41. Pryciak PM, Müller HP, Varmus HE. Simian virus 40 minichromosomes as targets for retroviral integration in vivo. Proc Natl Acad Sci U S A. 1992;89:9237–41.
42. Wang GP, Ciuffi A, Leipzig J, Berry CC, Bushman FD. HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. Genome Res. 2007;17:1186–94.
43. Wang GP, Levine BL, Binder GK, Berry CC, Malani N, McGarrity G, et al. Analysis of lentiviral vector integration in HIV+ study subjects receiving autologous infusions of gene modified CD4+ T cells. Mol Ther. 2009;17:844–50.
44. Roth SL, Malani N, Bushman FD. Gammaretroviral integration into nucleosomal target DNA in vivo. J Virol. 2011;85:7393–401.
45. Benleulmi MS, Matysiak J, Henriquez DR, Vaillant C, Lesbats P, Calmels C, et al. Intasome architecture and chromatin density modulate retroviral integration into nucleosome. Retrovirology. 2015;12:13.
46. Valkov E, Gupta SS, Hare S, Helander A, Roversi P, McClure M, et al. Functional and structural characterization of the integrase from the prototype foamy virus. Nucleic Acids Res. 2009;37:243–55.
47. Schweizer M, Fleps U, Jäckle A, Renne R, Turek R, Neumann-Haefelin D. Simian foamy virus type 3 (SFV-3) in latently infected Vero cells: reactivation by demethylation of proviral DNA. Virology. 1993;192:663–6.
48. Neves M, Périès J, Saïb A. Study of human foamy virus proviral integration in chronically infected murine cells. Res Virol. 1998;149:393–401.
49. Dhar R, McClements WL, Enquist LW, Vande Woude GF. Nucleotide sequences of integrated Moloney sarcoma provirus long terminal repeats and their host and viral junctions. Proc Natl Acad Sci U S A. 1980;77:3937–41.
50. Shoemaker C, Goff S, Gilboa E, Paskind M, Mitra SW, Baltimore D. Structure of a cloned circular Moloney murine leukemia virus DNA molecule containing an inverted segment: implications for retrovirus integration. Proc Natl Acad Sci U S A. 1980;77:3932–6.
51. Niebert M, Rogel-Gaillard C, Chardon P, Tönjes RR. Characterization of chromosomally assigned replication-competent gamma porcine endogenous retroviruses derived from a large white pig and expression in human cells. J Virol. 2002;76:2714–20.
52. Gorbovitskaia M, Liu Z, Bourgeaux N, Li N, Lian Z, Chardon P, et al. Characterization of two porcine endogenous retrovirus integration loci and variability in pigs. Immunogenetics. 2003;55:262–70.
53. Kim S, Kim N, Dong B, Boren D, Lee SA, Das Gupta J, et al. Integration site preference of xenotropic murine leukemia virus-related virus, a new human retrovirus associated with prostate cancer. J Virol. 2008;82:9964–77.
54. Kim S, Rusmevichientong A, Dong B, Remenyi R, Silverman RH, Chow SA. Fidelity of target site duplication and sequence preference during integration of xenotropic murine leukemia virus-related virus. PLoS One. 2010;5, e10255.
55. Hishinuma F, DeBona PJ, Astrin S, Skalka AM. Nucleotide sequence of acceptor site and termini of integrated avian endogenous provirus ev1: integration creates a 6 bp repeat of host DNA. Cell. 1981;23:155–64.
56. Hughes SH, Mutschler A, Bishop JM, Varmus HE. A Rous sarcoma virus provirus is flanked by short direct repeats of a cellular DNA sequence present in only one copy prior to integration. Proc Natl Acad Sci U S A. 1981;78:4299–5303.
57. Chou KS, Okayama A, Tachibana N, Lee TH, Essex M. Nucleotide sequence analysis of a full-length human T-cell leukemia virus type I from adult T-cell leukemia cells: a prematurely terminated PX open reading frame II. Int J Cancer. 1995;60:701–6.
58. Chou KS, Okayama A, Su IJ, Lee TH, Essex M. Preferred nucleotide sequence at the integration target site of human T-cell leukemia virus type I from patients with adult T-cell leukemia. Int J Cancer. 1996;65:20–4.
59. Ono M. Molecular cloning and long terminal repeat sequences of human endogenous retrovirus genes related to types A and B retrovirus genes. J Virol. 1986;58:937–44.
60. Majors JE, Varmus HE. Nucleotide sequences at host-proviral junctions for mouse mammary tumour virus. Nature. 1981;289:253–8.
61. Vincent KA, York-Higgins D, Quiroga M, Brown PO. Host sequences flanking the HIV provirus. Nucleic Acids Res. 1990;18:6045–7.
62. Vink C, Groenink M, Elgersma Y, Fouchier RA, Tersmette M, Plasterk RH. Analysis of the junctions between human immunodeficiency virus type 1 proviral DNA and human DNA. J Virol. 1990;64:5626–7.
63. Regier DA, Desrosiers RC. The complete nucleotide sequence of a pathogenic molecular clone of simian immunodeficiency virus. AIDS Res Hum Retroviruses. 1990;6:1221–31.
64. Hacker CV, Vink CA, Wardell TW, Lee S, Treasure P, Kingsman SM, et al. The integration profile of EIAV-based vectors. Mol Ther. 2006;14:536–45.
65. Shimotohno K, Temin HM. No apparent nucleotide sequence specificity in cellular DNA juxtaposed to retrovirus proviruses. Proc Natl Acad Sci U S A. 1980;77:7357–61.
66. Ballandras-Colas A, Naraharisetty H, Li X, Serrao E, Engelman A. Biochemical characterization of novel retroviral integrase proteins. PLoS One. 2013;8, e76638.
67. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14:1188–90.
68. Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, Moore IK, et al. A genomic code for nucleosome positioning. Nature. 2006;442:772–8.
69. Aiyer S, Swapna GV, Malani N, Aramini JM, Schneider WM, Plumb MR, et al. Altering murine leukemia virus integration through disruption of the integrase and BET protein family interaction. Nucleic Acids Res. 2014;42:5917–28.
70. Larue RC, Plumb MR, Crowe BL, Shkriabai N, Sharma A, DiFiore J, et al. Bimodal high-affinity association of Brd4 with murine leukemia virus integrase and mononucleosomes. Nucleic Acids Res. 2014;42:4868–81.
71. Suzuki M, Yagi N. Stereochemical basis of DNA bending by transcription factors. Nucleic Acids Res. 1995;23:2083–91.
72. el Hassan MA, Calladine CR. Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. J Mol Biol. 1996;259:95–103.
73. Packer MJ, Dauncey MP, Hunter CA. Sequence-dependent DNA structure: dinucleotide conformational maps. J Mol Biol. 2000;295:71–83.
74. Olson WK, Gorin AA, Lu XJ, Hock LM, Zhurkin VB. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. Proc Natl Acad Sci U S A. 1998;95:11163–8.
75. Aymami J, Coll M, van der Marel GA, van Boom JH, Wang AH, Rich A. Molecular structure of nicked DNA: a substrate for DNA repair enzymes. Proc Natl Acad Sci U S A. 1990;87:2526–30.
76. Zhang Y, Crothers DM. High-throughput approach for detection of DNA bending and flexibility based on cyclization. Proc Natl Acad Sci U S A. 2003;100:3161–6.
77. Protozanova E, Yakovchuk P, Frank-Kamenetskii MD. Stacked-unstacked equilibrium at the nick site of DNA. J Mol Biol. 2004;342:775–85.
78. Mills JB, Hagerman PJ. Origin of the intrinsic rigidity of DNA. Nucleic Acids Res. 2004;32:4055–9.
79. Richmond TJ, Davey CA. The structure of DNA in the nucleosome core. Nature. 2003;423:145–50.
80. Brukner I, Sánchez R, Suck D, Pongor S. Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. EMBO J. 1995;14:1812–8.
81. Taganov KD, Cuesta I, Daniel R, Cirillo LA, Katz RA, Zaret KS, et al. Integrase-specific enhancement and suppression of retroviral DNA integration by compacted chromatin structure in vitro. J Virol. 2004;78:5848–55.
82. Lesbats P, Botbol Y, Chevereau G, Vaillant C, Calmels C, Arneodo A, et al. Functional coupling between HIV-1 integrase and the SWI/SNF chromatin remodeling complex for efficient in vitro integration into stable nucleosomes. PLoS Pathog. 2011;7, e1001280.
83. Demeulemeester J, Vets S, Schrijvers R, Madlala P, De Maeyer M, De Rijck J, et al. HIV-1 integrase variants retarget viral integration and are associated with disease progression in a chronic infection cohort. Cell Host Microbe. 2014;16:651–62.
84. Cherepanov P. LEDGF/p75 interacts with divergent lentiviral integrases and modulates their enzymatic activity in vitro. Nucleic Acids Res. 2007;35:113–24.

85. Dougherty JP, Temin HM. High mutation rate of a spleen necrosis virus-based retrovirus vector. Mol Cell Biol. 1986;6:4387–95.
86. Watanabe S, Temin HM. Construction of a helper cell line for avian reticuloendotheliosis virus cloning vectors. Mol Cell Biol. 1983;3:2241–9.
87. Matreyek KA, Wang W, Serrao E, Singh P, Levin HL, Engelman A. Host and viral determinants for MxB restriction of HIV-1 infection. Retrovirology. 2014;11:90.
88. R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2011.
89. Dyda F, Hickman AB, Jenkins TM, Engelman A, Craigie R, Davies DR. Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases. Science. 1994;266:1981–6.
90. Bujacz G, Jaskolski M, Alexandratos J, Wlodawer A, Merkel G, Katz RA, et al. High-resolution structure of the catalytic domain of avian sarcoma virus integrase. J Mol Biol. 1995;253:333–46.
91. Chen Z, Yan Y, Munshi S, Li Y, Zugay-Murphy J, Xu B, et al. X-ray structure of simian immunodeficiency virus integrase containing the core and C-terminal domain (residues 50–293) - an initial glance of the viral DNA binding platform. J Mol Biol. 2000;296:521–33.
92. Trobridge GD, Miller DG, Jacobs MA, Allen JM, Kiem HP, Kaul R, et al. Foamy virus vector integration sites in normal human cells. Proc Natl Acad Sci U S A. 2006;103:1498–503.
93. Nowrouzi A, Dittrich M, Klanke C, Heinkelein M, Rammling M, Dandekar T, et al. Genome-wide mapping of foamy virus vector integrations into a human cell line. J Gen Virol. 2006;87:1339–47.
94. Moalic Y, Felix H, Takeuchi Y, Jestin A, Blanchard Y. Genome areas with high gene density and CpG island neighborhood strongly attract porcine endogenous retrovirus for integration and favor the formation of hot spots. J Virol. 2009;83:1920–9.
95. Hematti P, Hong BK, Ferguson C, Adler R, Hanawa H, Sellers S, et al. Distinct genomic integration of MLV and SIV vectors in primate hematopoietic stem and progenitor cells. PLoS Biol. 2004;2, e423.
96. Crise B, Li Y, Yuan C, Morcock DR, Whitby D, Munroe DJ, et al. Simian immunodeficiency virus integration preference is similar to that of human immunodeficiency virus type 1. J Virol. 2005;79:12199–204.
97. Mitchell RS, Beitzel BF, Schroder AR, Shinn P, Chen H, Berry CC, et al. Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. PLoS Biol. 2004;2, E234.
98. Narezkina A, Taganov KD, Litwin S, Stoyanova R, Hayashi J, Seeger C, et al. Genome-wide analyses of avian sarcoma virus integration sites. J Virol. 2004;78:11656–63.
99. Barr SD, Leipzig J, Shinn P, Ecker JR, Bushman FD. Integration targeting by avian sarcoma-leukosis virus and human immunodeficiency virus in the chicken genome. J Virol. 2005;79:12035–44.
100. Brady T, Lee YN, Ronen K, Malani N, Berry CC, Bieniasz PD, et al. Integration target site selection by a resurrected human endogenous retrovirus. Genes Dev. 2009;23:633–42.
101. Meekings KN, Leipzig J, Bushman FD, Taylor GP, Bangham CR. HTLV-1 integration into transcriptionally active genomic regions is associated with proviral expression and with HAM/TSP. PLoS Pathog. 2008;4, e1000027.
102. Gillet NA, Malani N, Melamed A, Gormley N, Carter R, Bentley D, et al. The host genomic environment of the provirus determines the abundance of HTLV-1-infected T-cell clones. Blood. 2011;117:3113–22.
103. de Jong J, Akhtar W, Badhai J, Rust AG, Rad R, Hilkens J, et al. Chromatin landscapes of retroviral and transposon integration profiles. PLoS Genet. 2014;10, e1004250.