Unraveling the palindromic and nonpalindromic motifs of retroviral integration site sequences by statistical mixture models

Dalibor Miklík, Jirí Grim, Daniel Elleder, et al.

Genome Res. 2023 33: 1395-1408 originally published online July 18, 2023
Access the most recent version at doi:10.1101/gr.277694.123

Method

Unraveling the palindromic and nonpalindromic motifs of retroviral integration site sequences by statistical mixture models

Dalibor Miklík,[1] Jiří Grim,[2] Daniel Elleder,[1] and Jiří Hejnar[1]

[1] *Laboratory of Viral and Cellular Genetics, Institute of Molecular Genetics of the Czech Academy of Sciences, Prague 4, 142 20, Czech Republic;* [2] *Pattern Recognition Department, Institute of Information Theory and Automation of the Czech Academy of Sciences, Prague 8, 182 08, Czech Republic*

A weak palindromic nucleotide motif is the hallmark of retroviral integration site alignments. Given that the majority of target sequences are not palindromic, the current model explains the symmetry by an overlap of the nonpalindromic motif present on one of the half-sites of the sequences. Here, we show that the implementation of multicomponent mixture models allows for different interpretations consistent with the existence of both palindromic and nonpalindromic submotifs in the sets of integration site sequences. We further show that the weak palindromic motifs result from freely combined site-specific submotifs restricted to only a few positions proximal to the site of integration. The submotifs are formed by either palindrome-forming nucleotide preference or nucleotide exclusion. Using the mixture models, we also identify HIV-1-favored palindromic sequences in *Alu* repeats serving as local hotspots for integration. The application of the novel statistical approach provides deeper insight into the selection of retroviral integration sites and may prove to be a valuable tool in the analysis of any type of DNA motifs.

[Supplemental material is available for this article.]

Retroviruses integrate the DNA copy of their genome into the genome of the host. The distribution of proviruses is genome-wide with global preferences toward specific chromatin states and genomic features (Elleder et al. 2002; Schröder et al. 2002; Mitchell et al. 2004; Trobridge et al. 2006; Derse et al. 2007; De Ravin et al. 2014; LaFave et al. 2014). Globally, tethering to chromatin by cellular proteins drives the preferences (Ciuffi et al. 2005; De Rijck et al. 2013; Gupta et al. 2013; Sharma et al. 2013; Sowd et al. 2016; Winans et al. 2017). Locally, the occupancy of nucleosomes and the physical attributes of DNA affect the target site selection (Pryciak and Varmus 1992; Benleulmi et al. 2015; Naughtin et al. 2015; Michieletto et al. 2019; Jóźwik et al. 2022).

Retroviral enzyme integrase (IN) is the key protein that processes viral DNA ends and catalyzes strand transfer reaction. During the strand transfer reaction, 3′ hydroxyl groups of viral DNA attack phosphodiester bonds of target DNA (tDNA), leading to a covalent connection of the host and viral DNA. Typically, the cleavage sites on both strands of tDNA are 4 to 6 nucleotides (nt) apart depending on the retrovirus. Structural studies showed that IN assembles into symmetric multimeric complexes that bend the tDNA, enabling the strand transfer reaction inside a widened major groove of tDNA (Maertens et al. 2010; Hare et al. 2012; Ballandras-Colas et al. 2016, 2017; Passos et al. 2017; Wilson et al. 2019; Bhatt et al. 2020; Pandey et al. 2021; Ballandras-Colas et al. 2022; Jóźwik et al. 2022).

After the alignment of preintegration tDNA sequences, a weak palindromic motif arises as a hallmark of sites recognized by IN (Fig. 1A; Fitzgerald et al. 1992; Pryciak et al. 1992; Stevens and Griffith 1996; Carteau et al. 1998; Holman and Coffin 2005; Wu et al. 2005). The appearance of the palindromic motifs is dependent on IN amino acids that are in contact with tDNA (Harper et al. 2001, 2003; Maertens et al. 2010; Demeulemeester et al. 2014; Serrao et al. 2014; Aiyer et al. 2015). Hence, the selection of the precise position for integration shows a pattern consistent with the IN-dictated preference for tDNA composition. Given the symmetry of IN multimers, the preference for palindromic motifs may be expected. In addition, the palindromic motif is observed at target sites of other transposable elements (Vigdal et al. 2002; Miyao et al. 2003; Linheiro and Bergman 2008; Gangadharan et al. 2010; Mularoni et al. 2012; Riggs et al. 2021) and is a marker of binding sites of some transcription factors (Datta and Rister 2022).

Palindromic motifs are expected to be formed by the alignment of palindromic sequences. Palindromes are, however, rarely present in individual sequences targeted by retroviral integration. It was thus suggested that the weak palindromic motif results from the nonpalindromic motif being present on either tDNA strand, namely, on one half-site of the tDNA (Kirk et al. 2016). According to this hypothesis, the statistical distribution mixture model of two reverse-complementary components (Fig. 1B) was proposed to decompose the set of sequences. In this way, however, the constrained model estimates only the first submotif, and the second submotif is strictly defined by its reverse-complement. Although this approach can generate two identical palindromic submotifs if the first submotif is palindromic, the model produces a pair of nonpalindromic submotifs. The results thus led to the conclusion that nonpalindromic submotifs overlap and mix to produce a weak palindromic consensus.

Here, we have used unconstrained mixture models of multiple (more than two) components to unravel heterogeneous

Corresponding author: jiri.hejnar@img.cas.cz
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.277694.123. Freely available online through the *Genome Research* Open Access option.

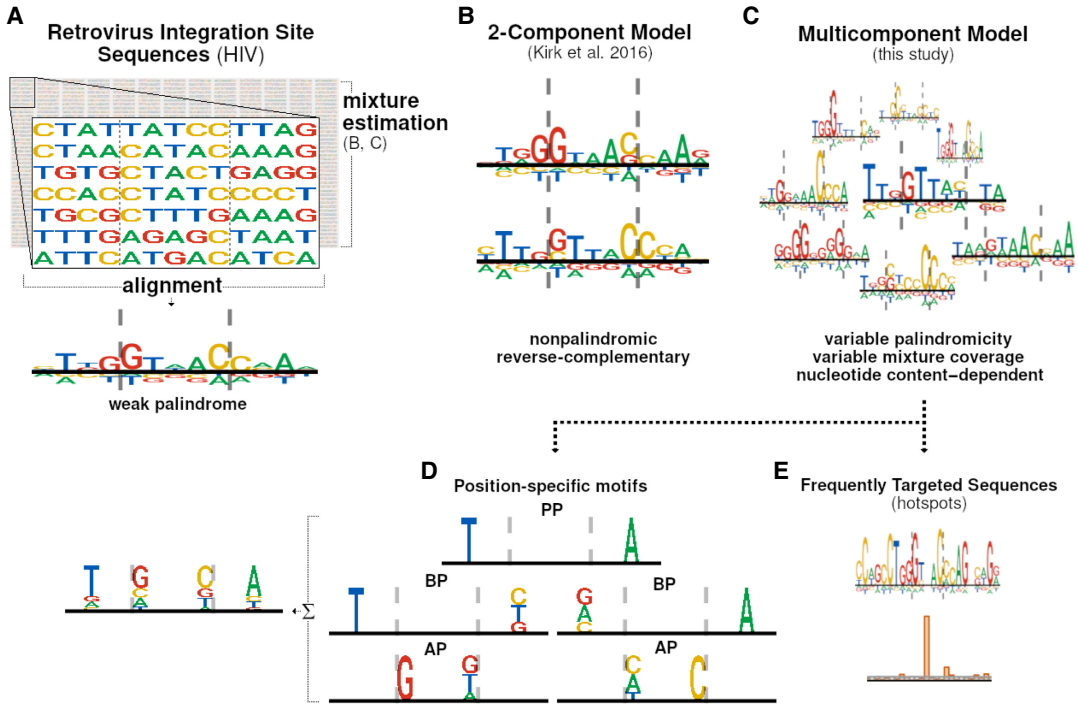

Figure 1. Origin of palindrome consensus motifs at sites of retroviral integration. Generally, a palindromic DNA motif composed of two reverse-complementary half-sites appears when palindromic sequences are aligned. (*A*) A weak palindromic motif appears when the sequences retrieved from the integration sites of retroviruses are aligned. The nature of the motif suggests the low frequency of the palindrome in aligned DNA sequences. Instead of whole-set alignments, mixture model estimates can be used to describe the possible submotifs in the sequence population. (*B*) A constrained two-component mixture model was used to analyze the consensus-forming submotifs, where an asymmetric motif appears on either of the half-sites of the target DNA (tDNA). (*C*) In this work, we used unconstrained multicomponent mixture models formed by at least two components (an eight-component mixture is depicted here). (*D*) Based on the submotifs appearing in multicomponent mixtures, we further performed quantitative analysis of the position-specific motifs. The major position-specific motifs observed include positional palindromes (PPs), broken palindromes (BPs), and asymmetric pairs (APs). (*E*) We described subpopulations of the frequently targeted sequences represented by low-abundant components in highly decomposed mixtures. We subsequently showed that one such component represents an abundant local hotspot of HIV-1 integration. Dashed lines in the sequence logos mark the cleavage sites of retroviral IN.

subpopulations of sequences targeted by retroviral integration (Fig. 1C). Using the information from the multicomponent mixture models, we describe and quantify isolated position-specific motifs and subpopulations of the frequently targeted sequences forming local hotspots. We hypothesize that the multicomponent mixture-driven description of tDNA motif/sequence subpopulations will provide novel insight into the preferences in retroviral integration site (IS) selection.

Results

EM algorithm uncovers palindromic and asymmetric submotifs in retroviral IS

To uncover sequence motifs at retroviral ISs, we estimated the mixture models of IS data sets using the expectation-maximization (EM) algorithm. We collected data representing ISs of human immunodeficiency virus type 1 (HIV) (Zhyvoloup et al. 2017), human T cell leukemia virus type 1 (HTLV) (Kirk et al. 2016), maedi-visna virus (MVV) (Ballandras-Colas et al. 2022), Moloney murine leukemia virus (MLV) (De Ravin et al. 2014), prototype foamy virus (PFV) (Lesbats et al. 2017), and avian sarcoma and leukosis virus (ASLV) (Malhotra et al. 2017).

Each IS set is characterized by a distribution mixture with a fixed number of components. Because a reliable choice of a "true" number of components is not justified in our case, we estimated and characterized mixture models consisting of a variable number of components (up to eight components owing to the reasons discussed throughout the work). A component is defined by (1) a nucleotide position probability matrix (PPM) and (2) component weight representing the proportion of mixture covered by the component. To visualize the PPMs, we developed a sequence logo based on the Kullback–Leibler information divergence (KLID) that emphasizes the divergence of the observed frequencies to the genomic probabilities of nucleotides. Additionally, we calculated the palindromic defect, which is defined as the dissimilarity between the PPM and its reverse-complement, indicates the level of the PPM asymmetry, and is equal to zero for palindromic PPMs. The PPMs representing entire retroviral IS sets showed weak sequence logos (Fig. 2A) and close-to-zero palindromic defects (Fig. 2B) in correspondence to the weak palindromic motifs at IS alignments.

The palindromic defect was next calculated for each PPM of estimated mixtures (Fig. 2B; Supplemental Fig. 1). The two-component mixtures were represented by components with increased palindromic defect compared with the whole-set PPM. The increased palindromic defect in two-component and other multicomponent mixtures indicates the presence of nonpalindromic motifs. Starting at the number of three components, the multicomponent mixtures often constitute also one component displaying a low palindromic defect—a marker of a possibly palindromic motif. This

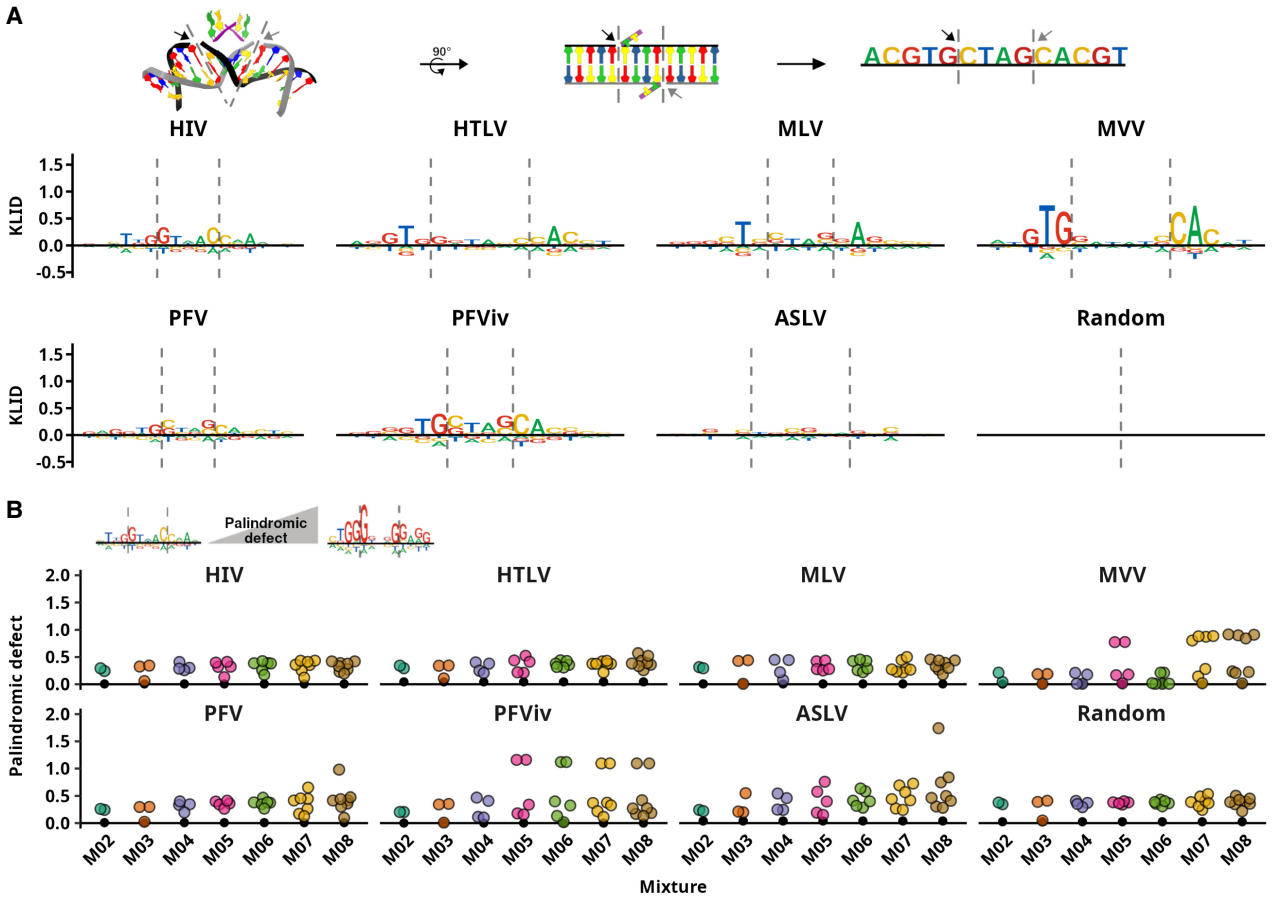

Figure 2. Identification of palindromic and asymmetric mixture components. (*A*) Scheme of retroviral integration and Kullback–Leibler information divergence (KLID)-based sequence logos representing complete sets of retroviral IS and a set of random genomic sequences. The arrows and the dashed lines mark the cleavage sites where the strand transfer reaction takes place. Logos represent IS sequences spanning 8 nt to each side from the center of the sequences. KLID values express the informativity of the *n*th position in sequence alignment relative to the global nucleotide frequencies. The character heights are proportional to the respective contributions of the nucleotides to the value of KLID. (*B*) Representation of palindromic defect that is defined as the dissimilarity between the PPM and its reverse-complement and is equal to zero for palindromic PPMs. All PPMs of retroviral IS component mixtures are represented. On the *x*-axis, mixtures are ordered by the number of components from two-component (M02) to eight-component (M08) mixture. The palindromic defect of the complete IS set (sequence logos in Fig. 1A) is included as a dark circle in each column.

variability in palindromic defect values among the components of identical mixtures suggests the coexistence of palindromic and nonpalindromic motifs in mixtures with three or more components.

Next, we examined the sequence logos of the mixture components displaying the highest and the lowest palindromic defect observed with a particular retroviral IS set (Supplemental Fig. 2B). The most palindromic components identified by the lowest palindromic defect are often derived from three-component mixtures and (except for HIV) are enriched in A and T nucleotides with dominant T/A palindromic motif. In contrast, the PPMs with the highest palindromic defect (presumably representing nonpalindromic motifs) are often derived from seven- or eight-component mixtures and are asymmetrically enriched in G/C nucleotides on either of the tDNA half-sites (Supplemental Fig. 2C). This observation validates the existence of diverse palindromic and nonpalindromic/asymmetric motifs among the components of retroviral IS mixture models.

Initial analysis identified both asymmetric (i.e., nonpalindromic with asymmetric nucleotide enrichment) and palindromic components in estimated IS mixtures. Although two-component mixtures preferably create pairs of asymmetric components, three-component mixtures reveal at least one palindromic component. Because multicomponent mixtures seem to produce diverse components, the components of two- and three-component mixtures could still artificially reflect the overlap of distinct motifs and the introduction of multicomponent mixtures might thus be necessary to minimize the creation of artificial motifs.

Multiple motifs are present in mixtures of IS sequences

To get a complete picture of the mixture decomposition, we examined the sequence logos of all components forming the mixtures (Supplemental Sequence Logos). We considered several features as indirect "interpretational" evidence of high decomposition: (1) frequent observation of the motifs that resemble another motif in the mixture—the motif is observed repetitively, and (2) emergence of components with low weight—the small populations of highly similar sequences start to be classified separately. We observed a high degree of motif decomposition already in mixtures formed by eight or fewer components. We have also found that in most cases the eight-component mixture is sufficient to reveal

all relevant properties of the IS data. Therefore, we present the two- to eight-component mixture models throughout the study.

Visual examination reveals the high heterogeneity of the sequence motifs both inside the mixtures and between the retroviral IS sets. First, there is a clear separation of A-, T-, G-, or C-enriched components. The simple nucleotide content thus may be a common feature of some IS sequences. For instance, A/T-rich components are present in eight-component mixtures of HIV (components C04, C02), HTLV (C02, C05), and MVV (C03) IS sets (Fig. 3). In particular, the A/T-rich palindromic component resembling the C03 component of MVV eight-component mixture appears in every MVV IS mixture that includes at least three components. This component covers 7%–11% of the respective mixtures and lacks otherwise ubiquitous enrichment for palindrome-forming G/C

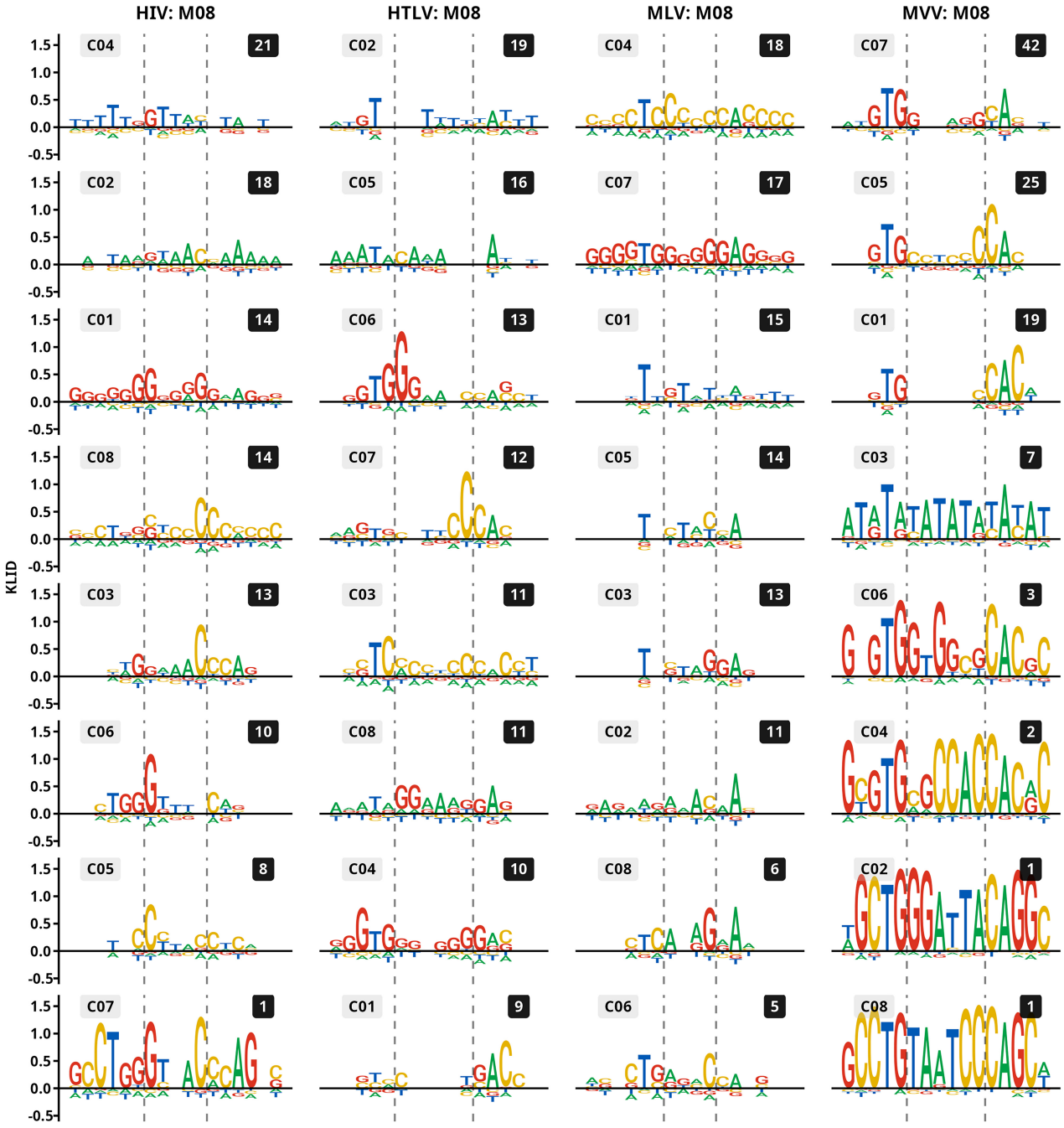

Figure 3. Sequence logo representations of the mixture model components of IS data sets. Eight-component (M08) mixtures of HIV, HTLV, MLV, and MVV IS sets are shown as examples of highly decomposed mixtures. KLID is used to construct the logos. Sequence logos of the two- to eight-component mixture models can be found in the Supplemental Sequence Logos file. Symbols in the gray field mark arbitrary component names as they appear in the output of the EM algorithm. Numbers in the black field mark the percentage of the mixture covered by the component. Logos represent IS sequences spanning 8 nt to each side from the center of the sequences. Vertical dashed lines mark sites where a strand transfer reaction takes place.

nucleotides upstream to the cleavage sites. This observation of separated A/T-rich components thus suggests that A/T-rich sequences may form a special category of target substrates.

The enrichment for certain nucleotides at specific positions was also repeatedly observed. The palindrome-forming enrichment of the T/A nucleotides at positions outside the area between the cleavage sites is one frequently present motif. This palindromic motif is most prominent in MVV and MLV mixtures, where it is present even in G/C-rich components (MLV components C04 and C07 and MVV components C07, C05, and C01) (Fig. 3) and rarely segregates into an asymmetric motif. On the other hand, HIV and HTLV mixtures contain components with asymmetric enrichment of G/C nucleotides inside the area between the cleavage sites (HIV components C03/C06 and HTLV components C06/C07) (Fig. 3). The aforementioned components form the bona fide (i.e., unforced) asymmetric reverse-complementary pairs that together cover 20%–40% of the respective mixtures. Other IS set-specific positional motifs like asymmetric G/C enrichment outside the area between the cleavage sites (HTLV components C04/C01) (Fig. 3) or TA enrichment at central positions in the area for in vitro PFV IS sequences (PFViv eight-component mixture components C02 and C06) (Supplemental Sequence Logos) were observed. The components enriched for the central TA dinucleotide were specifically observed in the PFV IS data set derived from in vitro integration into deproteinized DNA but not in the IS set obtained after cell infection.

Visual examination of the mixture models revealed the complexity of the retroviral IS sets and validated the existence of both palindromic and asymmetric motifs in the multicomponent mixtures. The observation of motifs segregating into distinct components indicates the existence of independent separable submotifs. Importantly, the described motifs appear repeatedly in independently estimated mixtures. Similar submotifs were observed also in independent IS data sets derived from HIV- and HTLV-infected cell cultures or patient blood cells (Supplemental Sequence Logo; Vansant et al. 2020; Melamed et al. 2022). Because the outcome of the mixture models is highly dependent on the component interpretability, the repeated observation of matching submotifs across the mixtures and independent IS sets adds to the reliability of the described submotifs.

Preferred nucleotide combinations at specific positions of IS sequences

The nature of the mixture components obtained in our analysis admits the existence of both palindromic and asymmetric motifs at the IS sequences. Because the results of the mixture models rely on their interpretability, we sought to quantify motifs directly and in a position-specific manner. We therefore determined the frequencies of nucleotide combinations at complementary positions proximal to the cleavage sites (in text, the positions are labeled by vertical bars; for detailed description, see Methods) (see legends of Fig. 4 and Supplemental Fig. 3 for schematic explanations).

First, we calculated KLID as an indicator of the cleavage site–relative, position-specific enrichment for nucleotide combinations (Fig. 4A; Supplemental Fig. 3). Generally, we observed that high KLID values are caused by enrichment for either single dominant combination or multiple nucleotide combinations. Single dominant contribution is predominantly observed at cleavage site–upstream positions |−2| or |−3| and is formed by the palindrome-forming T–T combination. The palindrome-forming G–G combination is also enriched at the position |−1| of the MVV IS sequences. On the other hand, the KLID value at position |1| is often higher than the value of a single contribution, pointing to the fact that KLID contributions of multiple nucleotide combinations add to its total value. In contrast to other sets of IS sequences, we observed the HIV IS sequences consistently displayed the highest KLID value at the position |1| (Fig. 4A; Supplemental Fig. 3). Although the palindrome-forming G–G combination is dominant here, other combinations certainly contribute to the total KLID value at this position.

Next, we examined the frequencies of the nucleotide combinations at the positions with elevated KLID values. The cleavage site upstream T–T combination is present in 22% of HIV IS at position |−3| and in 27%–60% of PFVin, HTLV, MLV, and MVV IS sequences at position |−2| (Fig. 4B,C; Supplemental Fig. 4). At the T–T enriched sites, frequencies of other nucleotides were increased when in combination with T. In total, the T nucleotide is observed at one of the sites forming |−2|/|−3| positions in up to 95% of IS sequences (Fig. 4D, top). On the other hand, G is the most disfavored nucleotide at the position (Fig. 4D, bottom). Besides the T-rich |−2| position, MVV IS sequences are enriched in G at position |−1| where G–G and G–A combinations are observed in 35% and 33% of IS sequences (Supplemental Fig. 4).

Position |1| is the site where asymmetric components of HIV and HTLV mixtures were enriched in G nucleotides. We observed G–G palindrome-forming combination at position |1| in 17% of HIV but only in 8% of HTLV IS sequences. More frequent are, however, the asymmetric G–A and G–C combinations that are each present in 20% of HIV and 18% and 16% of HTLV IS sequences. In total, 67% of HIV and 53% of HTLV IS sequences contain G at position |1|. The most striking, however, is the disfavor of combinations containing T nucleotide at position |1| observed also in MLV and PFV IS sequences that otherwise display only mild preferences for nucleotide combinations at the position (Fig. 4D, bottom; Supplemental Fig. 5B).

In summary, we quantified the nucleotide combinations in a position-specific way and showed that the preferences for nucleotide composition concentrate on a few specific positions surrounding the cleavage sites. We reported two types of motif-forming mechanisms that we refer to as broken palindromes and T-exclusion. The broken palindromes appear upstream of the cleavage sites and are created by less preferred nucleotides disrupting the preferred palindrome-forming combinations. On the other hand, the exclusion of T at positions downstream to the cleavage sites causes the increase of other nucleotide frequencies at the position.

Component mixture models identify hotspot of HIV-1 integration

Inspection of the multicomponent mixtures revealed low-weight components with strong motifs in the sequence logo. We hypothesize the components may describe small groups of highly similar sequences possibly serving as local hotspots for integration. To challenge the hypothesis, we characterized IS associated with the palindromic component C07 of the eight-component HIV IS mixture (Fig. 3).

We identified 1587 IS sequences associated with the C07 component. First, we examined the genomic localization of the IS. The majority of the C07-associated IS (96%) overlap with *Alu* repeats (Fig. 5A). When mapped to the *Alu* consensus sequence (Price et al. 2004), the ISs concentrated to two loci (Fig. 5B; Supplemental Table 2). The dominant locus with 62% of intra-*Alu* ISs mapped at consensus position 262 and is characterized by the sequence CTGGGCGACAGAG. Next, we reduced the IS sequences to the 13-bp motifs formed by palindrome-forming positions |1|, |−2|,

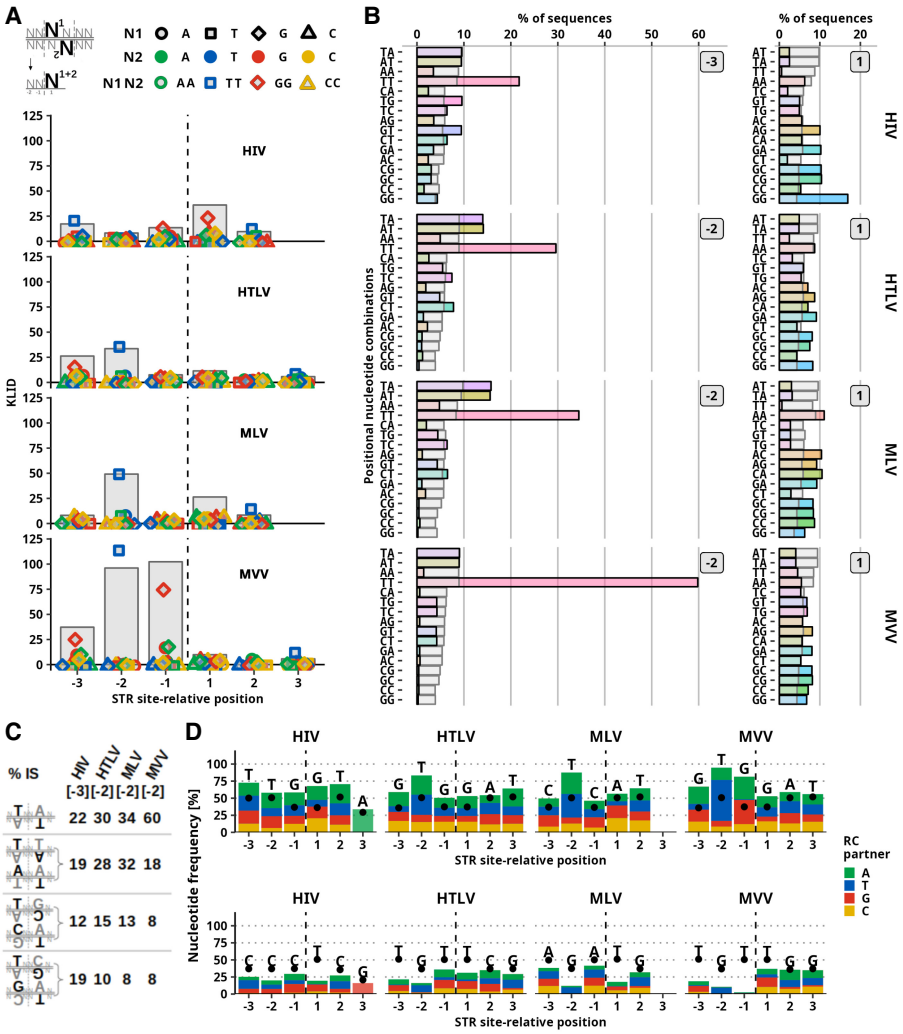

Figure 4. Preferred nucleotide combination at positions proximal to the cleavage site. Frequency of nucleotide combinations at complementary positions of both DNA strands. (*A*) KLID score of dinucleotide combinations at complementary sites marked by position relative to the cleavage site. The position of the cleavage site is marked by the vertical dashed line. Positions upstream to the cleavage site are marked with negative values. Gray bars represent the total KLID value at the position. Colored points represent individual contributions of each of the nucleotide combinations. Colors and shapes of points marking palindromic combinations are depicted in the legend. (*B*) Frequency of sequences with marked positional nucleotide combination. Gray bars represent the frequency observed in a control (random) set of sequences. The numbers to the *right* of the bars show the position relative to the cleavage site. The *left* column represents the positions with the maximum KLID score upstream of the cleavage site, and the *right* column represents the cleavage site–proximal position |1|. (*C*) Representation of frequency of T-containing dinucleotide combinations at position |−3| of HIV and position |−2| of HTLV, MLV, and MVV IS sequences. (*D*) Representation of frequencies of the most (*top*) and the least (*bottom*) frequent nucleotide at a given cleavage site–relative position. The height of the bar represents the marked nucleotide frequency, and the colored area of the bar represents the frequency of nucleotide present at complementary cleavage site–relative position (i.e., position with the same number but on opposite DNA strand). Black points represent the nucleotide frequency in the random set of sequences.

and |−3| (Figs. 3, 4A). The palindromic CT..G…C..AG motif was the most frequent motif in C07-associated sequences (Fig. 5C) as well as in intra-*Alu* IS sequences from the whole HIV IS set (Fig. 5E). Following our previous observations, most of the sequence motifs of the top 10 intra-*Alu*-targeted motifs contained one half-site of the palindromic motif (5′-CT…G…-3′). Although ISs are enriched at the positions of the palindromic motif, enrichment is not observed in the proximity to the sites of the motif (Fig. 5F). These results suggest that the exact positions, not the surrounding loci, are locally preferred HIV integration hotspots.

To test the effect of preintegration complex tethering to chromatin, we extracted HIV IS data sets in which capsid-CPSF6 (N74D) (Zhyvoloup et al. 2017) or IN-LEDGF/p75 (LEDGIN treatment or LEDGF/p75 KD) (Vansant et al. 2020) interactions were disrupted. Uniform preferential integration into the CT..G…C…AG motif was documented in each of the analyzed IS sets (Fig. 5G,H; Supplemental Figs. 6, 7). The substitutions of IN amino acids (Demeulemeester et al. 2014), however, impaired the targeting of the motif. It is noteworthy that every S119 mutant analyzed displayed the shift away from the palindromic hotspot caused by either more random selection for nucleotide combination at ISs (S119G/I/K/R/V) or enhanced preference for nucleotide combinations at certain positions relative to ISs (S119A/T) (Supplemental Fig. 8).

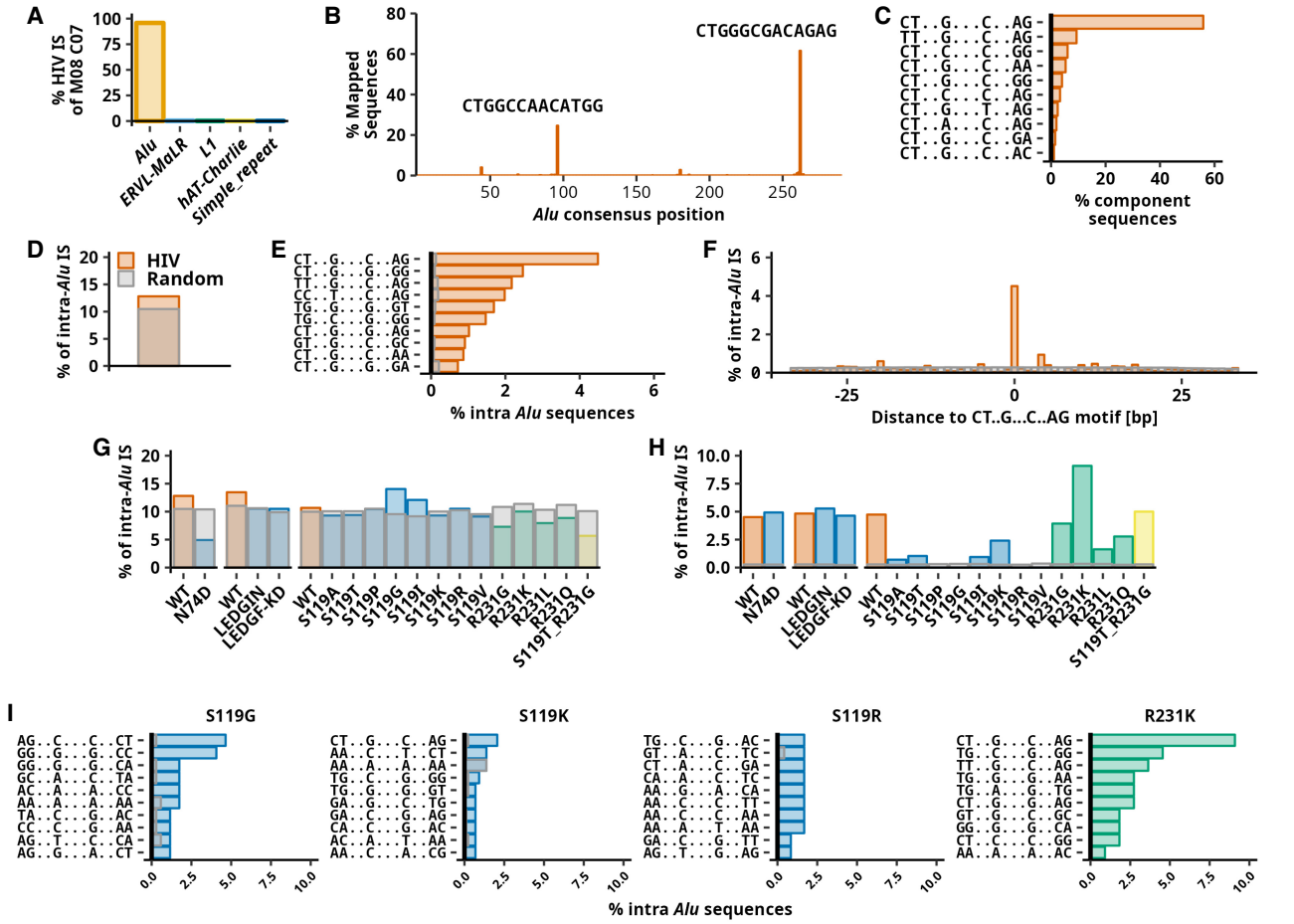

Figure 5. Characterization of HIV integration hotspot in *Alu* repeats. (*A*–*C*) The characterization of HIV-1 IS sequences associated with component C07 of an eight-component mixture. (*A*) Barplot showing the frequency of 1596 component-associated ISs in repetitive elements. Five repeat families with the most ISs are depicted. (*B*) Intra-*Alu* component-associated sequences mapped to *Alu* consensus sequence. The plot shows single-base binned consensus positions (length 290 bp) and a percentage of 1063 sequences mapped to individual positions of *Alu* consensus. Sequences of the two most frequent positions in the *Alu* consensus sequence are displayed. (*C*) Frequency of sequence motifs in sequences associated with component C07. The 10 most frequent sequence motifs are shown. In sequence motifs, dots (".") substitute any nucleotide. (*D*–*F*) Data for HIV-1 obtained from the complete HIV IS set. (*D*) Barplot showing a frequency of IS found in *Alu* repeats. The gray bar represents the expected random targeting of *Alu* repeats. (*E*) Frequency of the sequence motifs among intra-*Alu* IS. Gray bars represent the mean expected frequency. The 10 most frequent sequence motifs are shown. (*F*) Frequency of integration into and in proximity to the most frequently targeted motif CT..G…C..AG. Plots show the frequency of intra-*Alu* IS 33 bp downstream from and upstream of the motif relative to *Alu* repeat orientation. (*G*–*I*) Results of the same analysis as shown for *D*–*F* performed for different studies. Each experimental set contains a control with normal (WT) integration preference. N74D is HIV-1 capsid mutation, LEDGIN marks the usage of LEGFF/p75-IN interaction inhibitor, and LEDGF-KD transduction of cells in which LEDGF/p75 protein is knocked down. The *right* part of the bar plots represents IN mutants. Mutants of S119 are shown in blue, mutants of R231 in green, and mutant of both S119 and R231 in yellow. (*G*) Frequency of HIV-1 IS in *Alu* repeats. (*H*) Frequency of intra-*Alu* repeats in palindromic motif CT..G…C..AG. (*I*) Frequency of 10 most frequent sequence motifs.

Using the information from the mixture components, we identified a local integration hotspot of HIV. The hotspot has a form of CT..G…C..AG palindromic motif, which is preferentially located near the 3′ end of *Alu* repeats. The sequence hotspot is observed irrespective of the tethering of the HIV preintegration complex to chromatin and is an intrinsic feature of HIV IN. Under conditions not disrupting integration preference, about one in 175 HIV ISs (Supplemental Table 3) is, on average, located at the hotspot.

Discussion

The selection of the tDNA sequence is the last step in the complex process of retroviral IS determination. Using multicomponent mixture models, we unraveled the submotifs present in the sequences targeted by retroviral integration. Results retrieved from the mixture models and subsequent analysis showed that the weak palindromic motif of retroviral ISs is formed by separate motifs formed by broken palindromes and nucleotide exclusion. In addition, we showed that the models can be used to identify locally preferred sequences of retroviral integration. In this way, we characterized HIV-1 palindromic hotspot in *Alu* repeats.

A previous study suggested that the weak palindromic motif appears as a consequence of a nonpalindromic motif occurring on one of the tDNA half-sites (Kirk et al. 2016). The mixture model used was imposed to create two reverse-complementary motifs, and although the unconstrained clustering (rather than the unconstrained mixture model) was considered in the work, the investigators interpreted the two-component model as a correct explanation of the weak palindromic motif at retroviral IS. Here,

we showed that even the unconstrained two-component mixture models are selective toward the creation of asymmetric pairs of motifs, confirming that a significant fraction of target sequences are nonpalindromic. The introduction of multicomponent models, however, reveals that the low number of components possibly creates the averaged motifs, as is true for the weak palindromic consensus motif. With the higher number of components, the components describe smaller portions of the whole mixture, and we can observe a decomposition of the motifs observed with the lower number of components. The asymmetric pairs then describe the consensus motif only partially and various motifs, including the palindromic ones, contribute to the formation of the consensus motif.

We cannot define the definitive number of submotifs from the estimated mixture models, as a reliable choice of a "true" number of components is not justified in the case of categorical data. Instead, we rely on the interpretability of components that was shown to be essential for inferring biological implications from mixture models of categorical data (Gyllenberg et al. 1994). With interpretability as a major qualitative result of the mixture model analysis, we used various multicomponent models to observe the motifs in independently estimated mixtures. Both the more precise components and the recurrence of motifs in independently estimated mixtures contribute to interpreting the components with increased confidence. Critically, we challenged the interpretations with a direct mixture-independent quantitative analysis of the positional combinations that confirmed the presence of position-specific motifs.

One explanation for the nucleotide preferences might reside in specific interactions at the IN-tDNA interface. About 3 bp upstream of the cleavage site, structurally conserved intrusion into the minor groove of small amino acids like serine (HIV, RSV), proline (HTLV, MVV, MLV [Aiyer et al. 2015], and MMTV [Jóźwik et al. 2022]), and alanine (PFV) was documented. Albeit no nucleotide-specific interactions seem to be present at the position, mutations of HIV-1 IN S119 (Harper et al. 2001; Demeulemeester et al. 2014; Serrao et al. 2014) affect the preference of IN target site selection. The substitutions of S119 can either enhance or cancel the preference for nucleotides at positions proximal to the cleavage site and can disrupt the T-exclusion at the position downstream from the cleavage site. The identity of the minor groove-intruding amino acid is thus the major determinant of IN preference for tDNA composition.

The exclusion of T at the position downstream from the cleavage site was previously explained by hindering the target phosphodiester by the C5-methyl group of T (Maertens et al. 2010). Additional preference for G at the position of HIV-1 IS could potentially be explained by the interaction of G N5 with the major groove-intruding arginine (R231) (Passos et al. 2017). Although R231G mutation was previously associated with mild global retargeting of HIV-1 integration (Demeulemeester et al. 2014), we did not observe altered targeting of *Alu* hotspot nor altered nucleotide preferences of IN R231 mutants. In addition, no significant preference for G downstream from the cleavage site is present in IS sets of PFV and MVV, where the major groove-intruding arginine is present. Moreover, obstruction of the strand transfer reaction by the T C5-methyl group may be overcome by substitutions of other amino acids. Hence, the contribution of major groove-invading arginine to G preference at the position is disputable.

Because there are no clear correlates between IN–tDNA interface composition and the strength of tDNA motifs, the motif may reflect a need for tDNA to adopt specific conformation. Yet, there is no major distortion of the B-form tDNA structure at the position occupied by preferred T (Jóźwik et al. 2022). Also, the number of documented H-bond-forming interactions between IN and tDNA does not correlate with the preference for T-palindrome upstream to the cleavage site.

Given the symmetry of retroviral intasome, the existence of nonpalindromic submotifs is an interesting feature of the tDNA sequences. Imperfect palindromes are known to regulate the binding of transcription factors, where changes in one half-site affect the binding of the transcription factor to the other half-site of the motif (Datta and Rister 2022). Unlike the motifs of transcription factors, retroviral IS motifs are limited to a few positions and are probably not the consequence of the sequence-specific IN–tDNA interactions. However, we showed that the presence of the preferred nucleotide at one half-site of tDNA can increase the chances of the less preferred nucleotide appearing at the other half-site of tDNA. We thus consider the model in which triggering the strand transfer reaction after tDNA recognition by IN can act asymmetrically.

In vitro experiments suggested that HIV-1 IN preference toward tDNA patterns shapes the local distribution of HIV-1 integration (Leavitt et al. 1992; Carteau et al. 1998). Using the mixture models, we identified a group of highly similar sequences within *in cellulo* IS set. This hotspot was located in *Alu* repeats and represents the sequence with a palindromic pattern that might be an ideal template for HIV-1 integration (Carteau et al. 1998). We observed the hotspot with constant frequency (about one in 175 ISs) in independent IS sets, and we showed that the targeting of the hotspot is an attribute of HIV-1 IN. The existence of such a hotspot highlights the fact that, at least for HIV-1 integration, the sequence composition shapes the local distribution of proviruses and that scanning of tDNA by IN may take place before integration.

In summary, we showed a novel statistical approach for the analysis of retroviral IS sequences, which can form a valuable tool in the analysis of DNA motifs. So far, studies exploring tDNA composition approached the tDNA as the perfect palindrome or worked on the averaged global alignment of IS sequences. We propose that the shown variability of target site sequences should be considered to understand the natural process of retroviral integration.

Methods

IS sequences

Sources of IS data are described in Supplemental Table 1. Only ISs of HTLV were obtained as genomic preintegration sequences from Supplemental Data from Kirk et al. (2016). For the rest of IS sets, genomic coordinates of LTR-proximal nucleotides were obtained from published coordinates, sets of mapped reads, or raw sequencing reads. In case the data were not previously publicly available (requested data), we made coordinates of mapped IS available in the form of BED files along with published IS sequences. Details on the data processing can be found in Supplemental Methods. Briefly, genomic coordinates of nucleotides proximal to proviral LTRs were first obtained from available data. In the case of HIV data, IS coordinates were obtained from the Supplemental Table from Zhyvoloup et al. (2017). Replicates of DMSO-treated samples of wt and capsid N74D mutant were joined to create wt and N74D IS sets. Genomic coordinates of HIV IS data were obtained from the Supplemental Table of the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) entry (Vansant et al. 2020). wt, LEDGIN, and LEDGF-KD data sets were created joining bulk, GFP^+, and GFP^- samples from the identical treatment group. HIV IS coordinates from Demeulemeester et al. (2014) were

extracted from tables provided by the study's investigators. For each variant, data from the transduction of different cell lines were mixed. Coordinates of MVV ISs from HEK293T cells (Ballandras-Colas et al. 2022) and PFV IS sets (Lesbats et al. 2017) were obtained from the GEO database. MLV (De Ravin et al. 2014) set was obtained from raw sequencing reads that were mapped to hg38 human genome assembly using Bowtie 2 (Langmead and Salzberg 2012). Only reads mapping from the first nucleotide and mapping uniquely to the genome were accepted. Data from CEF cells transduced by RCASC (Malhotra et al. 2017) were used as data representing ASLV integration. Mapped reads in SAM format were obtained from the investigators of the study. Alignments were filtered and converted to BED format using SAMtools view (Danecek et al. 2021) and BEDTools (Quinlan and Hall 2010) bamtobed tools. Another ASLV IS set (Moiani et al. 2014) was retrieved as raw reads and processed as IS raw reads of MLV and PFV. However, this IS set was not used in the main study as some uncertainties were observed during the analysis of mixture models (Supplemental Fig. 9).

Next, single–base pair coordinates of LTR-proximal nucleotides in BED format were transformed to ranges spanning 13 bp to each side. As a result, ranges of length 26 nt (HTLV, MLV, MVV, PFV, and ASLV) and 27 nt (HIV) were created. Sequences from respective genomes were obtained using BEDTools getfasta tools. Whenever IS coordinates were available in the original publication, we obtained sequences from the reference genome used in the publication and reproduced the IS set reported in a given publication. The genome references used are listed in Supplemental Table 1 along the data set information and include hg18, hg19, hg38, and galGal4. FASTA files were then transformed into tables of sequences. For the purpose of mixture modeling, nucleotides were encoded to numeric strings where A = 1, C = 2, G = 3, T = 4, and N = 5.

Random genomic sequences and shuffled controls

Sets of randomly selected genomic sequences were created to estimate the expected frequencies of nucleotides at loci targeted by retroviral integration. First, 9000 random genomic ranges of lengths 26 and 27 from hg19 and 10,000 ranges of length 26 bp from galGal4 were created using BEDTools random tool. Ranges were then transformed to FASTA sequences using the BEDTools getfasta tool. Sequences containing N's were removed, and sequences were then encoded to numeric strings as described for IS sequences. Sequences derived from the hg19 assembly were used as controls for all IS sets derived from the human genome, including those mapped to the hg38 assembly. Background nucleotide frequencies used for the creation of the sequence logo were obtained as a mean nucleotide frequency across all positions of 26-bp random sequence–derived PPMs obtained from hg19 genome assembly (Supplemental Table 4).

To control for the expected frequency of nucleotide combinations at complementary positions of the IS, the frequency of the combination was calculated in sets of random sequences derived from hg19 or galGal4 assemblies for human or chicken genomes, respectively.

To control for random targeting of genomic features, a BEDTools shuffle tool was used with a sample BED file and a file containing chromosome lengths of the targeted genome obtained from UCSC goldenpath (https://hgdownload.soe.ucsc.edu/downloads.html). In the case of shuffled controls created to ISs inside genomic features (i.e. *Alu* repeats), -incl and -f 0.6 options were included in the BEDTools shuffle command. The BEDTools getfasta tool was used to retrieve genomic sequences of generated ranges. When stated, shuffling was repeated several times, and target frequency was calculated as a mean of the frequency obtained in a particular iteration.

The mixture of multiple product components

Given a collection of aligned nucleotide sequences of length N, the basic statistical description of the data follows from the relative frequencies of the four bases A, C, G, and T at each position. We denote the set of bases as $\mathcal{B} = \{A,\ C,\ G,\ T\}$. The probabilities of bases are usually presented in the form of a PPM of size $4 \times N$. Denoting $\mathcal{N} = \{1,\ 2,\ \ldots,\ N\}$, we can write

$$PPM = (p_n(\xi)),\quad \xi \in \mathcal{B},\quad n \in \mathcal{N},\quad \sum_{\xi \in \mathcal{B}} p_n(\xi) = 1. \tag{1}$$

Here the rows of *PPM* correspond to the four bases A, C, G, and T, and the columns relate to the respective nucleotide positions $n = 1, 2, \ldots, N$. In each column, the probabilities sum to one. The *PPM* of the whole collection—in the following, we use the notation PPM_0—can be viewed as a basic statistical model of IS sequences of the considered retrovirus. To describe the statistical properties of IS sequence populations in a more specific way, we have applied a general unconstrained mixture model of multiple components, which can be estimated using the EM algorithm. We have found that the reverse-complementary components tend to occur in mixtures spontaneously, without any enforced condition, and the additional components may uncover other interesting properties of the IS sequences. For this purpose, we assume the multivariate probability distribution $P(\boldsymbol{x})$ in the form of a weighted sum of multiple product components $P(\boldsymbol{x}|m),\ \ m \in \mathcal{M} = \{1,\ 2,\ \ldots,\ M\}$:

$$P(\boldsymbol{x}) = \sum_{m \in \mathcal{M}} w_m P(\boldsymbol{x}|m),\ \ \boldsymbol{x} \in \mathcal{X}, \tag{2}$$

$$P(\boldsymbol{x}|m) = \prod_{n \in \mathcal{N}} p_n(x_n|m),\quad \sum_{\xi \in \mathcal{B}} p_n(\xi|m) = 1. \tag{3}$$

Here, w_m is the probabilistic weight of the mth component, and $p_n(x_n|m),\ \ n \in \mathcal{N}$ are the related component-specific position probabilities of the bases A, C, G, T. In other words, the distributions $p_n(\cdot|m),\ \ n \in \mathcal{N}$ define the component PPM_m of the underlying subpopulation.

EM algorithm

The distribution mixture (Equation 2) is a widely used general statistical model to approximate unknown discrete probability distributions. The standard way to estimate the mixture parameters $w_m, p_m(\cdot|m),\ \ m \in \mathcal{M}$ is to maximize the log-likelihood criterion

$$L = \frac{1}{|S|} \sum_{x \in S} \log\left[\sum_{m \in \mathcal{M}} w_m P(\boldsymbol{x}|m)\right] \tag{4}$$

using the iterative EM algorithm (Grim 2017). Let S be a collection of IS sequences of a retrovirus:

$$S = \{\boldsymbol{x}^{(1)},\ \boldsymbol{x}^{(2)},\ \ldots,\ \boldsymbol{x}^{(K)}\},\ \boldsymbol{x} = (x_1,\ \ \ldots,\ x_N),\ \ x_n \in \mathcal{B}. \tag{5}$$

To compute the $m.-l.$ estimates of the unknown parameters $w_m, p_n(\cdot|m)$, we repeat the basic EM iteration equations:

$$q(m|\boldsymbol{x}) = \frac{w_m P(\boldsymbol{x}|m)}{\sum_{j \in \mathcal{M}} w_j P(\boldsymbol{x}|j)},\quad m \in \mathcal{M}, \tag{6}$$

$$w'_m = \frac{1}{|S|} \sum_{x \in S} q(m|\boldsymbol{x}),\quad m \in \mathcal{M},\ \ \boldsymbol{x} \in S, \tag{7}$$

$$p'_n(\xi|m) = \frac{1}{\sum_{x \in S} q(m|\boldsymbol{x})} \sum_{x \in S} \delta(\xi, x_n) q(m|\boldsymbol{x}), \tag{8}$$

where $|S|$ is the number of sequences, w'_m, $p'_n(\cdot|m)$ are the new parameter values, and $\delta(\xi, x_n)$ denotes the usual delta-function, namely, $\delta(\xi, x_n) = 1$ for $\xi = x_n$ and otherwise is $\delta(\xi, x_n) = 0$.

The EM algorithm generates a nondecreasing sequence $\{L^{(t)}\}_0^\infty$, and the iterations are stopped when the relative increment

of the criterion is less than a chosen threshold. As Criterion 4 is bounded above ($L<0$), the monotonic property implies convergence of the sequence $\{L^{(t)}\}_0^\infty$ to a possible local maximum of the criterion, whereby the local maximum may be starting point dependent. To decrease the risk of locally optimal solutions, the initial parameters $p_n(\xi|m)$ are usually generated randomly, and the EM optimization is repeated several times with different initial values.

A difficult problem is a proper choice of the number of components M. Unlike Gaussian mixtures, the mixtures of discrete product components are not identifiable. As shown by Grim (2006), any discrete mixture can be equivalently described in infinitely many ways, and therefore, a reliable choice of a "true" model or "true" number of components is not justified. By increasing the number of components, we can increase the model accuracy in the statistical sense in terms of higher likelihood criterion. However, the number of low-weight marginal components would also increase, and some reasonably interpretable components could decompose. In this sense, the estimated components provide an opportunity to uncover the properties of underlying sequences in a transparent way, because the statistical properties of each component are defined by its *PPM*, and the related sequences are easily identified. It is known that in practical experiments the well-defined components are reasonably identifiable from data (Carreira-Perpiñán and Renals 2000), but there is no exact formulation of this property.

The situation is very similar to the application of multivariate Bernoulli mixtures to bacterial taxonomy (Gyllenberg et al. 1994). We recall that Bernoulli mixtures, as a special case of discrete mixtures, are not identifiable, and consequently, different classes of bacteria could become questionable. However, like in bacterial taxonomy, our problem is not a statistical but a genetic one, and therefore, we accept the interpretability of the resulting mixture model in terms of data as the ultimate criterion.

We make an essential modification to the EM algorithm, where we use the structural mixture model to suppress the influence of less informative noisy variables. The structural model essentially reduces the variability of resulting mixtures. In the computational experiments, we have found that in most cases the mixture of M = 8 components is sufficient to reveal all relevant properties of the IS data. Indirect evidence in this sense is the occurrence of several components of low or very low weight describing marginal properties.

Extraction of component-associated sequences

Given the estimated distribution mixture (Equation 2), we can characterize any sequence $\boldsymbol{x} \in S$ in terms of the conditional probabilities $q(m|\boldsymbol{x})$. The conditional posterior weight $q(m|\boldsymbol{x})$ can be interpreted as a measure of the affinity of the sequence $\boldsymbol{x}$ with the mth component or, in other words, as a membership value of $\boldsymbol{x}$ for the mth subpopulation. In this sense, we can decompose the original data S into subcollections S_m according to the maximum values $q(m|\boldsymbol{x})$, namely, using the Bayes formula. The membership value was calculated by custom R script, and sequences with $q(m|\boldsymbol{x}) \geq 0.9$ were identified and separated from IS data set as a component-associated IS. Given Equation 7, the component weight w_m can be interpreted as an estimate of the relative size of S_m.

Reverse-complement distance and palindromic defect

For any two PPMs P_1, P_2, the reverse-complement distance is defined as the sum of absolute differences of reverse-complementary position probabilities:

$$d(P_1, P_2) = \sum_{\xi \in \mathcal{B}} \frac{1}{N} \sum_{n \in \mathcal{N}} |\, p_n(\xi|1) - p_{N+1-n}(\tilde{\xi}|2)|. \tag{9}$$

If the distance (Equation 9) is zero then, by definition, the two matrices P_1 and P_2 are reverse-complementary.

If we apply the Equation 9 to the same matrix $P_1 = P_2 = P_0$, then the corresponding value $d(P_0, P_0)$ can be viewed as a measure of violation of the palindromic condition, namely, as a palindromic defect of the matrix P_0. We denote

$$D(P_0) = \sum_{\xi \in \mathcal{B}} \frac{1}{N} \sum_{n \in \mathcal{N}} |\, p_n(\xi|0) - p_{N+1-n}(\tilde{\xi}|0)|. \tag{10}$$

Here we ignore the central position if the number of positions N is not even. The palindromic defect $D(P_0)$, $(0 \leq D \leq 2)$ is zero for any palindromic P_0 and is positive otherwise.

Kullback–Leibler information divergence

If $p_n(\xi|m)$, $n \in \mathcal{N}$, $\xi \in \mathcal{B}$ are the position probabilities of the mth component and $p_n(\xi|0)$ the position probabilities of the background, then the KLID of the two distributions $p_n(\cdot\,|m)$, $p_n(\cdot\,|0)$ at the position $n \in \mathcal{N}$ is given by the formula

$$I(p_n(\cdot|m), p_n(\cdot|0)) = \sum_{\xi \in \mathcal{B}} p_n(\xi|m) \log \frac{p_n(\xi|m)}{p_n(\xi|0)} \geq 0. \tag{11}$$

The KLID is nonnegative, equals zero only if the two distributions are identical, and can be interpreted as a measure of information that is lost if the IS probabilities $p_n(\xi|m)$ are replaced by the global background probabilities $p_n(\xi|0)$.

Sequence logo based on KLID

A popular way to illustrate the related viral preferences is the well-known sequence logo derived from PPM_0, which highlights the overall IS motif. It is a histogram of N columns, each column displays proportionally the four probabilities of bases in descending order and the total height of the column is proportional to the Shannon information contained. The Shannon information is zero in the case of uniform probabilities, and it is higher if some bases are preferred.

However, in the present context, the KLID formula is a more informative tool to illustrate the properties of mixture components graphically by a histogram and can be used as a sequence logo. Unlike the standard sequence logo, the value of the expression $I(p_n(\cdot\,|m), p_n(\cdot\,|0))$ can be viewed as a measure of information importance of the nth position in the mth component, concerning the statistical properties of the background. The KLID formula (Equation 11) includes four terms that reflect the role of the four possible nucleotides. The term is positive if the IS probability $p_n(\xi|m)$ is greater than the global (background) probability $p_n(\xi|0)$, and it is negative if it is smaller. In the case of nonspecific background distribution, characterized by $\phi_{mn} = 0$, we have $p_n(\xi|m) = p_n(\xi|0)$, and the related KLID value is zero. At each position of the histogram, the column height corresponds to the informativity $I(p_n(\cdot\,|m), p_n(\cdot\,|0))$. The four color-parts are proportional to the respective contributions of the 4 nt to the value of KLID. In contrast to the standard sequence logo, the contributions of the less-probable nucleotides are displayed as negative, to emphasize their relation to the background.

The graphical representation of the sequence logo based on KLID was created with the ggseqlogo R package (Wagih 2017; R Core Team 2021). The functions of the package were modified to display the contribution of each nucleotide to KLID at the position. Given the set background probabilities (see section "Structural mixture model"), the maximum values for each nucleotide are 1.2226 for A, 1.5865 for C, 1.5714 for G, and 1.2271 for T.

Structural mixture model

The statistical analysis of a collection of aligned IS sequences based on a mixture model is a suitable approach to identifying the local IS preferences of a retrovirus. However, in this way, the statistical properties of the near IS neighborhood may be influenced by random properties of the more distant parts of genomic sequences. To suppress the possible noisy influence of less informative positions of IS sequences we have used a structural modification of the product mixture (Equation 2). In particular, using binary structural parameters $\phi_{mn} \in \{0, 1\}$, we can confine the estimation of component parameters only to some informative variables (Grim 2017). If we define

$$\begin{aligned} P(\boldsymbol{x}, \Phi) &= \sum_{m\in\mathcal{M}} w_m P(\boldsymbol{x}|m, \varphi_m), \quad \boldsymbol{x} \in \mathcal{X}, \\ P(\boldsymbol{x}|m, \varphi_m) &= \prod_{n\in\mathcal{N}} p_n(x_n|m)^{\varphi_{mn}} p_n(x_n|0)^{1-\varphi_{mn}}, \end{aligned} \tag{12}$$

then by setting the structural parameter $\phi_{mn} = 0$, we can replace any component-specific distribution $p_n(\cdot|m)$ by a common fixed "background" distribution $p_n(\cdot|0)$. In our case, we use the four global genomic probabilities $p(A|0) = 0.2968$, $p(C|0) = 0.2049$, $p(G|0) = 0.2028$, $p(T|0) = 0.2955$ as a background distribution:

$$\sum_{\xi\in\mathcal{B}} p_n(\xi|0) = 1, \quad n \in \mathcal{N}.$$

Structural EM algorithm

The structural mixture (Equation 12) can be optimized by the EM algorithm in full generality; namely, we can estimate both the component-specific distributions $p_n(x_n|m)$, $m \in \mathcal{M}$, and the optimal binary structural parameters ϕ_{mn}. For this purpose, we can use the standard EM iteration equations with the only difference that, in addition, we choose in each iteration the most informative parameters using KLID. In particular, making the substitution

$$P(\boldsymbol{x}|m, \varphi_m) = P(\boldsymbol{x}|0) G(\boldsymbol{x}|m, \varphi_m), \quad m \in \mathcal{M}, \tag{13}$$

$$G(\boldsymbol{x}|m, \varphi_m) = \prod_{n\in\mathcal{N}} \left[\frac{p_n(x_n|m)}{p_{n(x_n|0)}}\right]^{\varphi_{mn}}, \tag{14}$$

in Equation 2, we can write the structural mixture in the form

$$\begin{aligned} P(\boldsymbol{x}, \Phi) &= P(\boldsymbol{x}|0) \sum_{m\in\mathcal{M}} w_m G(\boldsymbol{x}|m, \varphi_m), \\ P(\boldsymbol{x}|0) &= \prod_{n\in\mathcal{N}} p_n(x_n|0), \quad \boldsymbol{x} \in \mathcal{X}, \end{aligned} \tag{15}$$

where $P(\boldsymbol{x}|0)$ is a fixed nonzero "background" probability distribution, and the component functions $G(\boldsymbol{x}|m, \phi_m)$ include the binary structural parameters $\phi_{mn} \in \{0, 1\}$. Obviously, by considering Equation 2, the structural mixture model (Equation 15) can be viewed formally as a product mixture again.

In this sense, we can estimate both the component-specific distributions $p_n(x_n|m)$ and the binary structural parameters ϕ_{mn} (Grim 2017) by means of the standard EM iteration Equations 6 through 8. Using substitution (Equation 15), we can write the structural log-likelihood criterion in the form

$$\begin{aligned} L &= \frac{1}{|S|} \sum_{x\in S} \log P(\boldsymbol{x}, \Phi) \\ &= \frac{1}{|S|} \sum_{x\in S} \log \left[\sum_{m\in M} w_m P(\boldsymbol{x}|0) G(\boldsymbol{x}|m, \varphi_m)\right], \end{aligned} \tag{16}$$

and, by using the formula of Equation 13, we can reduce the iterative Equation 6 to informative variables only:

$$q(m|\boldsymbol{x}) = \frac{w_m G(\boldsymbol{x}|m, \varphi_m)}{\sum_{j\in\mathcal{M}} w_j G(\boldsymbol{x}|j, \varphi_j)}, \quad m \in \mathcal{M}, \quad \boldsymbol{x} \in \mathcal{X}. \tag{17}$$

The next two equations, Equations 7 and 8, are unchanged. The only difference is that, in addition, we choose in each iteration the most informative parameters using the well-known KLID.

In each iteration, the KLID formula (Equation 11) is evaluated for all components and variables, and the following weighted mean is used to derive a suitable threshold value:

$$\tau = \frac{1}{NM} \sum_{m\in\mathcal{M}} \sum_{n\in\mathcal{N}} w_m I(p_n(\cdot|m), p_n(\cdot|0)). \tag{18}$$

In the next iteration step, only the sufficiently informative component distributions $p_n(\cdot|m)$ are used to satisfy the inequality:

$$w_m I(p_n(\cdot|m), p_n(\cdot|0)) > \alpha\tau, \quad n \in \mathcal{N}, \quad m \in \mathcal{M}. \tag{19}$$

Here, α is an optional coefficient that can be used to control the level of suppressed noise. By setting $\alpha = 0.0$, we would estimate all component-specific parameters, as in the general mixture. The value $\alpha = 1.0$ would eliminate all position distributions $p_n(\cdot|m)$ of below-average informativity. In our case, we have used $\alpha = 0.2$ to replace only the low informative position distributions by background.

KLID is widely used in statistics and information theory as a measure of discrimination information. However, it should be emphasized that the application of KLID in the present context is not intuitively motivated but directly follows from the monotonic condition of the EM algorithm (Grim 2017).

The EM iterations have been stopped when the relative increment of the criterion was less than 10^{-5}. Again, to decrease the risk of locally optimal solutions, the EM algorithm has been initialized randomly and repeated 100 times with different initial values. The maximum log-likelihood has been used as a criterion to select the representative mixture.

Positional nucleotide combinations

To provide information about the frequency of nucleotide combinations at cleavage site–relative positions, cleavage site–relative PPMs were created, denoted as PPM_{CS}. First, the alignments of IS sequences were split into equally long left A_L and right A_R halves. The middle position (marked by 0 in PPM) is ignored if the original number of positions N is odd. The length of each half-alignment A_L, A_R is $H = \lfloor N/2 \rfloor$. Positions of both half-alignments were given cleavage site–relative values: Positions downstream from the cleavage site were given positive values that increase with the distance from the cleavage site; positions upstream of the cleavage site were given negative values that decrease with the distance to the cleavage site. To analyze nucleotides forming the DNA strand on which strand transfer reaction takes place, nucleotides contained in A_R were transformed into the complementary nucleotides, creating $\tilde{A}_R$. Given that an identical cleavage site–relative position exists in A_L and $\tilde{A}_R$, nucleotide combination C is formed by a pair of nucleotides at identical positions h of the same sequence $\boldsymbol{x}$ (represented by a row in an alignment). PPM_{CS} is then defined as a matrix of size $16 \times H$ and is formed by nucleotide combination probabilities at positions: $PPM_{CS} = (p_h(c))$, $c \in \mathcal{C}$, $h \in \mathcal{H}$. The positional matrices PPM_{CS} were calculated also for the alignments of randomly selected genomic sequences. Here, we refer to the probabilities derived from random matrices PPM_{CS} as $p_h(c|0)$. KLID values for PPM_{CS} were calculated as described in Equation 11. The observed probability $p_n(\xi|m)$ was substituted by nucleotide combination probability $p_h(c)$, and background probability was substituted by random nucleotide combination probability $p_h(c)$. Results were displayed as a combined plot, where bars represent the positional

KLID value ($KLID_h$), and colored points represent the contribution of each of the nucleotide combinations c to the $KLID_h$.

Targeting of repetitive elements

RepeatMasker annotations were obtained from UCSC goldenpath as rmsk.txt tables. BED files were created from the rmsk tables using a custom awk script. The BEDTools intersect tool was used to identify overlaps between IS ranges of length 27 bp and repeat annotations.

Mapping sequences to the *Alu* consensus

An *Alu* repeat consensus was obtained from a previous publication (Price et al. 2004). Bowtie 2 was used to map genomic pre-IS sequences of length 27 bp to the *Alu* consensus sequence. First, Bowtie 2 indexes of the *Alu* consensus sequence were created with the bowtie-build command. Next, two rounds of mapping and filtering were run to select well-mapped IS sequences. In the first round, sequences were mapped to consensus with Bowtie 2 -f -L 5 -N 1 -i S,1,0.2 --score-min L,0,-2 --all command. Only reads mapped to a single site in consensus were selected. In the second round, a Bowtie 2 -f -L 5 -N 1 -i S,1,0.2 --all command was run on multimappers from the first round, and alignments with a single entry were again selected. Selected alignments from both rounds were joined and converted to ranges stored in BED files. Finally, each of the ranges was transformed into a single position representing the center of the range.

Distance to the intra-*Alu* sequence motif

To locate coordinates of sequence motifs, a SeqKit locate -i -r -p CT..G…C..AG command from SeqKit tool (Shen et al. 2016) was used. A BEDTools intersect tool was used to locate motifs positioned inside the *Alu* elements. To calculate the distance from the nearest motif to each IS, both IS and motif ranges were reduced to the center position, and distances were calculated using BEDTools closest tool with a -D option to discriminate upstream and downstream integrations.

Data access

The source code is available at GitHub (https://github.com/dalibormiklik/IS_Motifs.git) and as Supplemental Code. Supporting data are available as Supplemental Material and at Figshare (https://figshare.com/projects/Motifs_in_Retroviral_Integation_Sites/154179).

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank Karen Beemon, Shelby Winans, Rik Gijsbers, and Jonas Demeulemeester for sharing the integration site data sets. We also thank the members of the Laboratory of Viral and Cellular Genetics, especially Tomas Hron, Krystof Stafl, Katerina Trejbalova, and Filip Senigl, for the critical reading and the discussion concerning the manuscript. This work was supported by the Czech Science Foundation (project 19-23407S awarded to J.H.) and Czech Academy of Sciences (Premium Academiae Award to J.H.). D.M., D.E., and J.H. were supported by the project National Institute of Virology and Bacteriology (program EXCELES, no. LX22NPO5103) funded by the European Union–Next Generation EU; we also acknowledge institutional support from the project RVO (68378050).

Author contributions: D.M, J.G., and D.E. conceived the study. D.M. collected and processed integration site data sets, processed EM algorithm output data, created result visualizations, and completed the manuscript. J.G. optimized and ran the EM algorithm and wrote the statistical part of the Methods. D.M., D.E., and J.H. discussed the result interpretations. All authors contributed to the final version of the manuscript.

References

Aiyer S, Rossi P, Malani N, Schneider WM, Chandar A, Bushman FD, Montelione GT, Roth MJ. 2015. Structural and sequencing analysis of local target DNA recognition by MLV integrase. *Nucleic Acids Res* **43:** 5647–5663. doi:10.1093/nar/gkv410

Ballandras-Colas A, Brown M, Cook NJ, Dewdney TG, Demeler B, Cherepanov P, Lyumkis D, Engelman AN. 2016. Cryo-EM reveals a novel octameric integrase structure for betaretroviral intasome function. *Nature* **530:** 358–361. doi:10.1038/nature16955

Ballandras-Colas A, Maskell DP, Serrao E, Locke J, Swuec P, Jónsson SR, Kotecha A, Cook NJ, Pye VE, Taylor IA, et al. 2017. A supramolecular assembly mediates lentiviral DNA integration. *Science* **355:** 93–95. doi:10.1126/science.aah7002

Ballandras-Colas A, Chivukula V, Gruszka DT, Shan Z, Singh PK, Pye VE, McLean RK, Bedwell GJ, Li W, Nans A, et al. 2022. Multivalent interactions essential for lentiviral integrase function. *Nat Commun* **13:** 2416. doi:10.1038/s41467-022-29928-8

Benleulmi MS, Matysiak J, Henriquez DR, Vaillant C, Lesbats P, Calmels C, Naughtin M, Leon O, Skalka AM, Ruff M, et al. 2015. Intasome architecture and chromatin density modulate retroviral integration into nucleosome. *Retrovirology* **12:** 13. doi:10.1186/s12977-015-0145-9

Bhatt V, Shi K, Salamango DJ, Moeller NH, Pandey KK, Bera S, Bohl HO, Kurniawan F, Orellana K, Zhang W, et al. 2020. Structural basis of host protein hijacking in human T-cell leukemia virus integration. *Nat Commun* **11:** 3121. doi:10.1038/s41467-020-16963-6

Carreira-Perpiñán MA, Renals S. 2000. Practical identifiability of finite mixtures of multivariate Bernoulli distributions. *Neural Comput* **12:** 141–152. doi:10.1162/089976600300015925

Carteau S, Hoffmann C, Bushman F. 1998. Chromosome structure and human immunodeficiency virus type 1 cDNA integration: centromeric alphoid repeats are a disfavored target. *J Virol* **72:** 4005–4014. doi:10.1128/JVI.72.5.4005-4014.1998

Ciuffi A, Llano M, Poeschla E, Hoffmann C, Leipzig J, Shinn P, Ecker JR, Bushman F. 2005. A role for LEDGF/p75 in targeting HIV DNA integration. *Nat Med* **11:** 1287–1289. doi:10.1038/nm1329

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. *GigaScience* **10:** giab008. doi:10.1093/gigascience/giab008

Datta RR, Rister J. 2022. The power of the (imperfect) palindrome: sequence-specific roles of palindromic motifs in gene regulation. *Bioessays* **44:** e2100191. doi:10.1002/bies.202100191

Demeulemeester J, Vets S, Schrijvers R, Madlala P, De Maeyer M, De Rijck J, Ndung'u T, Debyser Z, Gijsbers R. 2014. HIV-1 integrase variants retarget viral integration and are associated with disease progression in a chronic infection cohort. *Cell Host Microbe* **16:** 651–662. doi:10.1016/j.chom.2014.09.016

De Ravin SS, Su L, Theobald N, Choi U, Macpherson JL, Poidinger M, Symonds G, Pond SM, Ferris AL, Hughes SH, et al. 2014. Enhancers are major targets for murine leukemia virus vector integration. *J Virol* **88:** 4504–4513. doi:10.1128/JVI.00011-14

De Rijck J, de Kogel C, Demeulemeester J, Vets S, El Ashkar S, Malani N, Bushman FD, Landuyt B, Husson SJ, Busschots K, et al. 2013. The BET family of proteins targets Moloney murine leukemia virus integration near transcription start sites. *Cell Rep* **5:** 886–894. doi:10.1016/j.celrep.2013.09.040

Derse D, Crise B, Li Y, Princler G, Lum N, Stewart C, McGrath CF, Hughes SH, Munroe DJ, Wu X. 2007. Human T-cell leukemia virus type 1 integration target sites in the human genome: comparison with those of other retroviruses. *J Virol* **81:** 6731–6741. doi:10.1128/JVI.02752-06

Elleder D, Pavlíček A, Pačes J, Hejnar J. 2002. Preferential integration of human immunodeficiency virus type 1 into genes, cytogenetic R bands and GC-rich DNA regions: insight from the human genome sequence. *FEBS Lett* **517:** 285–286. doi:10.1016/S0014-5793(02)02612-1

Fitzgerald ML, Vora AC, Zeh WG, Grandgenett DP. 1992. Concerted integration of viral DNA termini by purified avian myeloblastosis virus integrase. *J Virol* **66:** 6257–6263. doi:10.1128/jvi.66.11.6257-6263.1992

Gangadharan S, Mularoni L, Fain-Thornton J, Wheelan SJ, Craig NL. 2010. DNA transposon *Hermes* inserts into DNA in nucleosome-free regions in vivo. *Proc Natl Acad Sci* **107:** 21966–21972. doi:10.1073/pnas.1016382107

Grim J. 2006. EM cluster analysis for categorical data. In *Lecture notes in computer science* (ed. Yeung DY, et al.), pp. 640–648. Springer, Berlin.

Grim J. 2017. Approximation of unknown multivariate probability distributions by using mixtures of product components: a tutorial. *Intern J Pattern Recogniti Artif Intell* **31:** 1750028. doi:10.1142/s0218001417500288

Gupta SS, Maetzig T, Maertens GN, Sharif A, Rothe M, Weidner-Glunde M, Galla M, Schambach A, Cherepanov P, Schulz TF. 2013. Bromo- and extraterminal domain chromatin regulators serve as cofactors for murine leukemia virus integration. *J Virol* **87:** 12721–12736. doi:10.1128/JVI.01942-13

Gyllenberg M, Koski T, Reilink E, Verlaan M. 1994. Non-uniqueness in probabilistic numerical identification of bacteria. *J Appl Probab* **31:** 542–548. doi:10.2307/3215044

Hare S, Maertens GN, Cherepanov P. 2012. 3′-Processing and strand transfer catalysed by retroviral integrase *in crystallo*. *EMBO J* **31:** 3020–3028. doi:10.1038/emboj.2012.118

Harper AL, Skinner LM, Sudol M, Katzman M. 2001. Use of patient-derived human immunodeficiency virus type 1 integrases to identify a protein residue that affects target site selection. *J Virol* **75:** 7756–7762. doi:10.1128/JVI.75.16.7756-7762.2001

Harper AL, Sudol M, Katzman M. 2003. An amino acid in the central catalytic domain of three retroviral integrases that affects target site selection in nonviral DNA. *J Virol* **77:** 3838–3845. doi:10.1128/JVI.77.6.3838-3845.2003

Holman AG, Coffin JM. 2005. Symmetrical base preferences surrounding HIV-1, avian sarcoma/leukosis virus, and murine leukemia virus integration sites. *Proc Natl Acad Sci* **102:** 6103–6107. doi:10.1073/pnas.0501646102

Jóźwik IK, Li W, Zhang D-W, Wong D, Grawenhoff J, Ballandras-Colas A, Aiyer S, Cherepanov P, Engelman AN, Lyumkis D. 2022. B-to-A transition in target DNA during retroviral integration. *Nucleic Acids Res* **50:** 8898–8918. doi:10.1093/nar/gkac644

Kirk PDW, Huvet M, Melamed A, Maertens GN, Bangham CRM. 2016. Retroviruses integrate into a shared, non-palindromic DNA motif. *Nat Microbiol* **2:** 16212. doi:10.1038/nmicrobiol.2016.212

LaFave MC, Varshney GK, Gildea DE, Wolfsberg TG, Baxevanis AD, Burgess SM. 2014. MLV integration site selection is driven by strong enhancers and active promoters. *Nucleic Acids Res* **42:** 4257–4269. doi:10.1093/nar/gkt1399

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9:** 357–359. doi:10.1038/nmeth.1923

Leavitt AD, Rose RB, Varmus HE. 1992. Both substrate and target oligonucleotide sequences affect in vitro integration mediated by human immunodeficiency virus type 1 integrase protein produced in *Saccharomyces cerevisiae*. *J Virol* **66:** 2359–2368. doi:10.1128/jvi.66.4.2359-2368.1992

Lesbats P, Serrao E, Maskell DP, Pye VE, O'Reilly N, Lindemann D, Engelman AN, Cherepanov P. 2017. Structural basis for spumavirus GAG tethering to chromatin. *Proc Natl Acad Sci* **114:** 5509–5514. doi:10.1073/pnas.1621159114

Linheiro RS, Bergman CM. 2008. Testing the palindromic target site model for DNA transposon insertion using the *Drosophila* melanogaster P-element. *Nucleic Acids Res* **36:** 6199–6208. doi:10.1093/nar/gkn563

Maertens GN, Hare S, Cherepanov P. 2010. The mechanism of retroviral integration from X-ray structures of its key intermediates. *Nature* **468:** 326–329. doi:10.1038/nature09517

Malhotra S, Winans S, Lam G, Justice J, Morgan R, Beemon K. 2017. Selection for avian leukosis virus integration sites determines the clonal progression of B-cell lymphomas. *PLoS Pathog* **13:** e1006708. doi:10.1371/journal.ppat.1006708

Melamed A, Fitzgerald TW, Wang Y, Ma J, Birney E, Bangham CRM. 2022. Selective clonal persistence of human retroviruses in vivo: radial chromatin organization, integration site, and host transcription. *Sci Adv* **8:** eabm6210. doi:10.1126/sciadv.abm6210

Michieletto D, Lusic M, Marenduzzo D, Orlandini E. 2019. Physical principles of retroviral integration in the human genome. *Nat Commun* **10:** 575. doi:10.1038/s41467-019-08333-8

Mitchell RS, Beitzel BF, Schroder ARW, Shinn P, Chen H, Berry CC, Ecker JR, Bushman FD. 2004. Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol* **2:** E234. doi:10.1371/journal.pbio.0020234

Miyao A, Tanaka K, Murata K, Sawaki H, Takeda S, Abe K, Shinozuka Y, Onosato K, Hirochika H. 2003. Target site specificity of the *Tos17* retrotransposon shows a preference for insertion within genes and against insertion in retrotransposon-rich regions of the genome. *Plant Cell* **15:** 1771–1780. doi:10.1105/tpc.012559

Moiani A, Suerth JD, Gandolfi F, Rizzi E, Severgnini M, De Bellis G, Schambach A, Mavilio F. 2014. Genome-wide analysis of alpharetroviral integration in human hematopoietic stem/progenitor cells. *Genes (Basel)* **5:** 415–429. doi:10.3390/genes5020415

Mularoni L, Zhou Y, Bowen T, Gangadharan S, Wheelan SJ, Boeke JD. 2012. Retrotransposon Ty1 integration targets specifically positioned asymmetric nucleosomal DNA segments in tRNA hotspots. *Genome Res* **22:** 693–703. doi:10.1101/gr.129460.111

Naughtin M, Haftek-Terreau Z, Xavier J, Meyer S, Silvain M, Jaszczyszyn Y, Levy N, Miele V, Benleulmi MS, Ruff M, et al. 2015. DNA physical properties and nucleosome positions are major determinants of HIV-1 integrase selectivity. *PLoS One* **10:** e0129427. doi:10.1371/journal.pone.0129427

Pandey KK, Bera S, Shi K, Rau MJ, Oleru AV, Fitzpatrick JAJ, Engelman AN, Aihara H, Grandgenett DP. 2021. Cryo-EM structure of the Rous sarcoma virus octameric cleaved synaptic complex intasome. *Commun Biol* **4:** 330. doi:10.1038/s42003-021-01855-2

Passos DO, Li M, Yang R, Rebensburg SV, Ghirlando R, Jeon Y, Shkriabai N, Kvaratskhelia M, Craigie R, Lyumkis D. 2017. Cryo-EM structures and atomic model of the HIV-1 strand transfer complex intasome. *Science* **355:** 89–92. doi:10.1126/science.aah5163

Price AL, Eskin E, Pevzner PA. 2004. Whole-genome analysis of *Alu* repeat elements reveals complex evolutionary history. *Genome Res* **14:** 2245–2252. doi:10.1101/gr.2693004

Pryciak PM, Varmus HE. 1992. Nucleosomes, DNA-binding proteins, and DNA sequence modulate retroviral integration target site selection. *Cell* **69:** 769–780. doi:10.1016/0092-8674(92)90289-O

Pryciak PM, Sil A, Varmus HE. 1992. Retroviral integration into minichromosomes in vitro. *EMBO J* **11:** 291–303. doi:10.1002/j.1460-2075.1992.tb05052.x

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842. doi:10.1093/bioinformatics/btq033

R Core Team. 2023. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Riggs P, Blundell-Hunter G, Hagelberger J, Ren G, Ettwiller L, Berkmen M. 2021. Insertion specificity of the hATx-6 transposase of *Hydra magnipapillata*. *Front Mol Biosci* **8:** 734154. doi:10.3389/fmolb.2021.734154

Schröder ARW, Shinn P, Chen H, Berry C, Ecker JR, Bushman F. 2002. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110:** 521–529. doi:10.1016/S0092-8674(02)00864-4

Serrao E, Krishnan L, Shun M-C, Li X, Cherepanov P, Engelman A, Maertens GN. 2014. Integrase residues that determine nucleotide preferences at sites of HIV-1 integration: implications for the mechanism of target DNA binding. *Nucleic Acids Res* **42:** 5164–5176. doi:10.1093/nar/gku136

Sharma A, Larue RC, Plumb MR, Malani N, Male F, Slaughter A, Kessl JJ, Shkriabai N, Coward E, Aiyer SS, et al. 2013. BET proteins promote efficient murine leukemia virus integration at transcription start sites. *Proc Natl Acad Sci* **110:** 12036–12041. doi:10.1073/pnas.1307157110

Shen W, Le S, Li Y, Hu F. 2016. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One* **11:** e0163962. doi:10.1371/journal.pone.0163962

Sowd GA, Serrao E, Wang H, Wang W, Fadel HJ, Poeschla EM, Engelman AN. 2016. A critical role for alternative polyadenylation factor CPSF6 in targeting HIV-1 integration to transcriptionally active chromatin. *Proc Natl Acad Sci* **113:** E1054–E1063. doi:10.1073/pnas.1524213113

Stevens SW, Griffith JD. 1996. Sequence analysis of the human DNA flanking sites of human immunodeficiency virus type 1 integration. *J Virol* **70:** 6459–6462. doi:10.1128/jvi.70.9.6459-6462.1996

Trobridge GD, Miller DG, Jacobs MA, Allen JM, Kiem H-P, Kaul R, Russell DW. 2006. Foamy virus vector integration sites in normal human cells. *Proc Natl Acad Sci* **103:** 1498–1503. doi:10.1073/pnas.0510046103

Vansant G, Chen H-C, Zorita E, Trejbalová K, Miklík D, Filion G, Debyser Z. 2020. The chromatin landscape at the HIV-1 provirus integration site determines viral expression. *Nucleic Acids Res* **48:** 7801–7817. doi:10.1093/nar/gkaa536

Vigdal TJ, Kaufman CD, Izsvák Z, Voytas DF, Ivics Z. 2002. Common physical properties of DNA affecting target site selection of *Sleeping Beauty* and other Tc1/*mariner* transposable elements. *J Mol Biol* **323:** 441–452. doi:10.1016/S0022-2836(02)00991-9

Wagih O. 2017. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* **33:** 3645–3647. doi:10.1093/bioinformatics/btx469

Wilson MD, Renault L, Maskell DP, Ghoneim M, Pye VE, Nans A, Rueda DS, Cherepanov P, Costa A. 2019. Retroviral integration into nucleosomes through DNA looping and sliding along the histone octamer. *Nat Commun* **10:** 4189. doi:10.1038/s41467-019-12007-w

Winans S, Larue RC, Abraham CM, Shkriabai N, Skopp A, Winkler D, Kvaratskhelia M, Beemon KL. 2017. The FACT complex promotes avian leukosis virus DNA integration. *J Virol* **91:** e00082-17. doi:10.1128/JVI.00082-17

Wu X, Li Y, Crise B, Burgess SM, Munroe DJ. 2005. Weak palindromic consensus sequences are a common feature found at the integration target sites of many retroviruses. *J Virol* **79:** 5211–5214. doi:10.1128/JVI.79.8.5211-5214.2005

Zhyvoloup A, Melamed A, Anderson I, Planas D, Lee C-H, Kriston-Vizi J, Ketteler R, Merritt A, Routy J-P, Ancuta P, et al. 2017. Digoxin reveals a functional connection between HIV-1 integration preference and T-cell activation. *PLoS Pathog* **13:** e1006460. doi:10.1371/journal.ppat.1006460

Received January 13, 2023; accepted in revised form July 12, 2023.