Supplemental Information for

Unraveling the Palindromic and Non-Palindromic Motifs of Retroviral Integration Sites by Statistical Mixture Models

SUPPLEMENTAL TABLES

SUPPLEMENTAL FIGURES

SUPPLEMENTAL METHODS

SUPPLEMENTAL TABLES

Virus	Variant	Cells	genome	#IS	Publication	Source
HTLV-1	**wt**	**Jurkat**	**hg19**	**4521**	**Kirk et al. 2016**	**publication**
HIV-1	**wt**	**Jurkat**	**hg38**	**161470**	**Zhyvoloup et al. 2017**	**publication**
MLV	**wt**	**Human CD34+**	**hg38**	**57991**	**De Ravin et al. 2014**	**SRR1145702**
MVV	**wt**	**HEK293T**	**hg38**	**411721**	**Ballandras-Colas et al 2022**	**GSE196040**
PFV	**wt**	**HT1080**	**hg38**	**55384**	**Lesbats et al. 2017**	**GSM2584469**
ASLV	**RCASC**	**CEF**	**galGal4**	**2096**	**Malhotra et al. 2017**	**upon request**
HTLV-1	wt	patient PBMC	hg38	162760	Melamed et al. 2022	publication
	wt	cell culture (various)	hg38	235150		publication
HIV-1	wt	patient PBMC	hg38	44414	Melamed et al. 2022	publication
	wt	cell culture (various)	hg38	65924		publication
	N74D	Jurkat	hg38	63156	Zhyvoloup et al. 2017	publication
	wt	SupT1	hg19	16803	Vansant et al. 2020	GSE135295
	wt (LEDGIN+)			15683		
	wt (LEDGF-KD)			9069		
	wt	SupT1	hg18	6530	Demeulemeester et al. 2014	upon request
	IN variants	SupT1/HelaP4	hg18	27502 *		
PFV	wt	in vitro	hg38	4226638	Lesbats et al. 2017	GSM2584468
ASLV	ALVJ	CEF	galGal4	1956	Malhotra et al. 2017	upon request
ASLV	LR9			1210		
ASLV	RCASC	HeLa	hg19	837		
ASLV	RSV	Human CD34+	galGal4	1239	Moiani et al. 2014	SRR1282019

Supplemental Table 1. List of the data sets used in the study and sources of the data. The bold text marks the data used for the presentation of mixture model components in the main text. Data where the source is stated as “upon request” were requested and gained directly from the authors of the study.
* number represents the sum of all IS of all variants analyzed in the present study. Numbers for individual variants can be found in Supplemental Table 2.

start	**end**	**Mapped sequences**
248	275	585
82	109	235
31	58	39
166	193	27
248	274	13
248	276	9
248	273	8
61	77	5
178	193	4
248	269	4
82	107	4
166	189	3
248	272	3
82	101	3
82	110	3
247	275	2
248	280	2
251	272	2
251	275	2
74	95	2

Supplemental Table 2. Twenty ranges in the *Alu* consensus sequence with the most HIV IS sequences mapped. In total, 1,587 27 bp long sequences associated with component C07 of the M08 HIV mixture were mapped to Alu consensus of which 971 sequences were identified as mapping. A full table of targeted ranges can be found in supplemental files as: *HIV_Zhyvoloup_M08_PPM07_in_hg38_rmsk_Alu_to_consensus_PosFreq.txt*

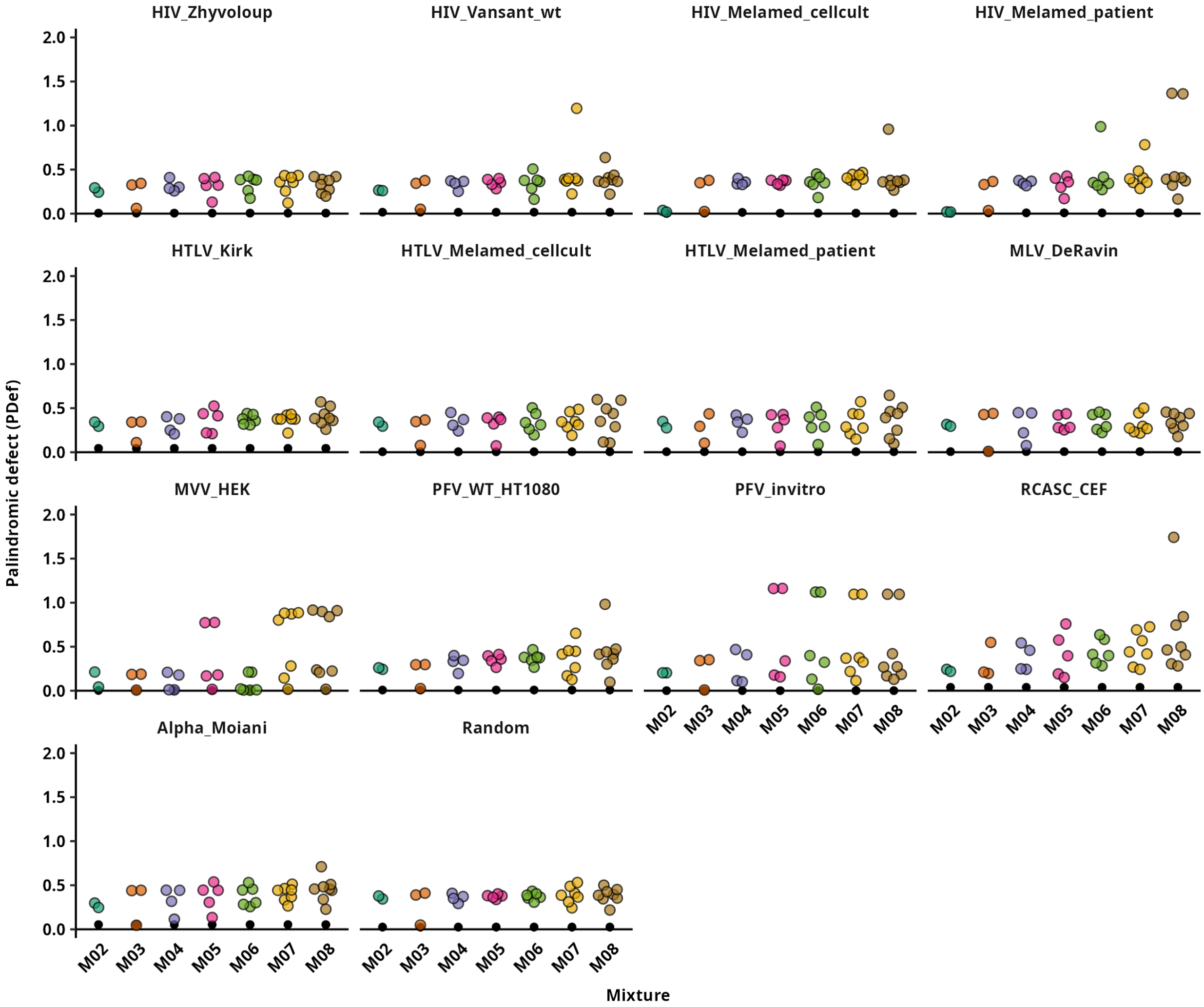
Study	Variant	IS count	intra *Alu* IS;Shuffle	CT..G...C..AG IS;Shuffle
Zhyvoloup	**WT**	**161533**	**20690;16927**	**930;26.5**
Zhyvoloup	N74D	63156	3108;6568	153;5
Vansant	**WT**	**16803**	**2262;1855**	**109;1**
Vansant	LEDGIN	15683	1652;1664	87;3
Vansant	LEDGF-KD	9069	952;897	44;1
Demeulemeester	**WT**	**6530**	**697;652**	**33;0.5**
Demeulemeester	S119A	1549	144;156	1;0.5
Demeulemeester	S119T	2066	194;208	2;0
Demeulemeester	S119T_R231G	2467	140;249	7;0
Demeulemeester	R231G	4897	357;531	14;1
Demeulemeester	R231K	1098	110;125	10;0
Demeulemeester	R231L	2314	184;239	3;0
Demeulemeester	R231Q	2438	216;273	6;0.5
Demeulemeester	S119P	1369	143;144	0;NA
Demeulemeester	S119G	1225	172;117	0;NA
Demeulemeester	S119I	1092	132;100	1;0
Demeulemeester	S119K	4682	436;469	9;1
Demeulemeester	S119R	1132	119;116	0;NA
Demeulemeester	S119V	1173	107;112	0;NA

Supplemental Table 3. Numbers of HIV-1 total IS and IS overlapping with *Alu* elements. The last column displays the number of IS found inside *Alu* and inside the palindromic motif. Intra *Alu* and CT..G…C..AG columns contain two values; the first shows the IS count, while the second shows the number of shuffled control IS. NA - frequencies were not calculated if the sample count was zero.

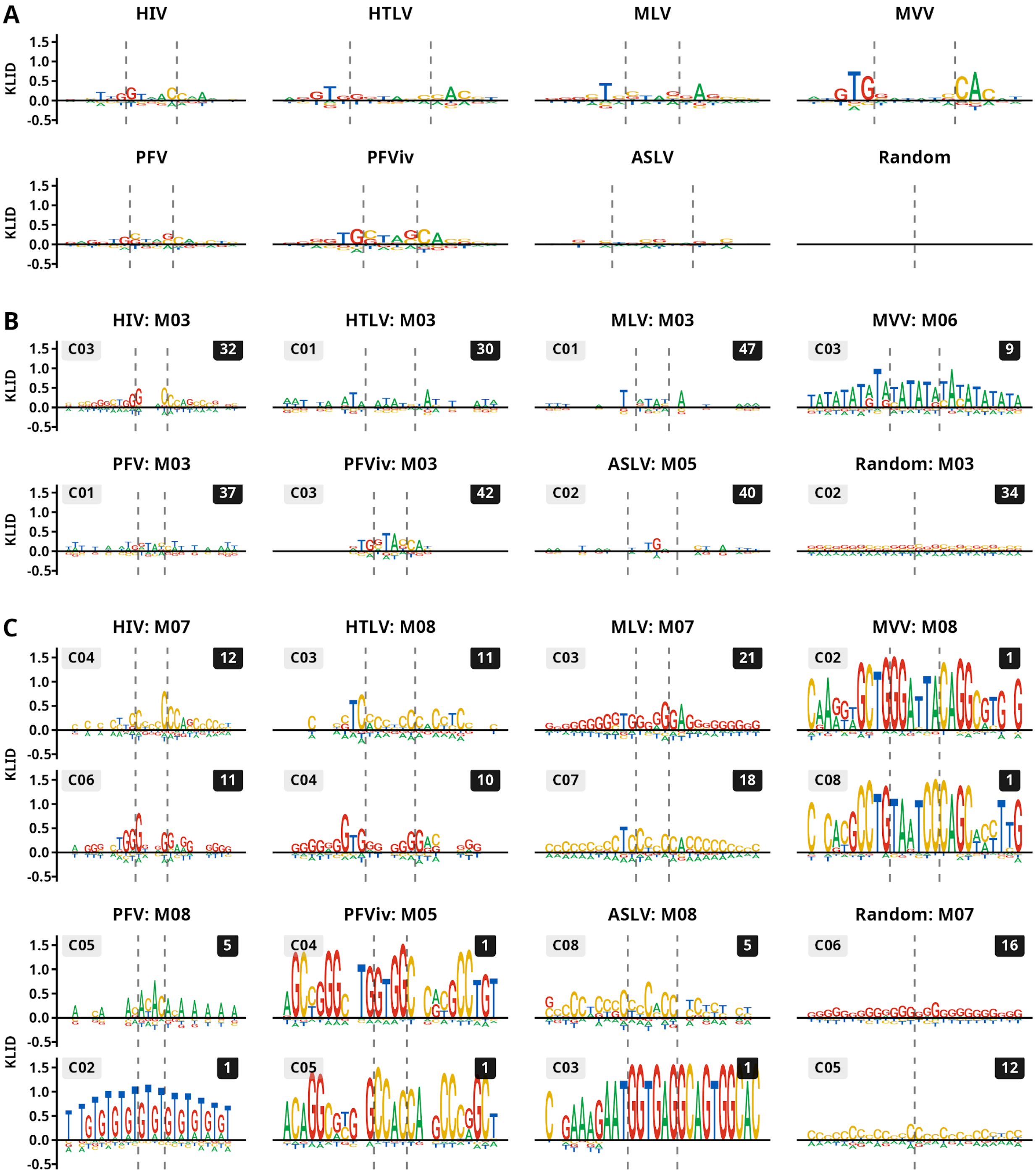
Random set	A	C	G	T
Ran9k_hg19_is26	0.296790192775594	0.204915522433771	0.202801456451091	0.295492828339544
Ran9k_hg19_is27	0.294440458157272	0.204239637867956	0.205076758174103	0.296243145800668
Ran10k_galGal4_is26	0.294465384615385	0.204642307692308	0.207746153846154	0.293146153846154

Supplemental Table 4. Expected nucleotide probabilities. Nucleotide frequencies were derived from $9x10^6$ (Ran9k) or 10^7 (Ran10k) sequences of randomly selected 26 or 27 bp long ranges from human (hg19) and chicken (galGal4) reference genomes.

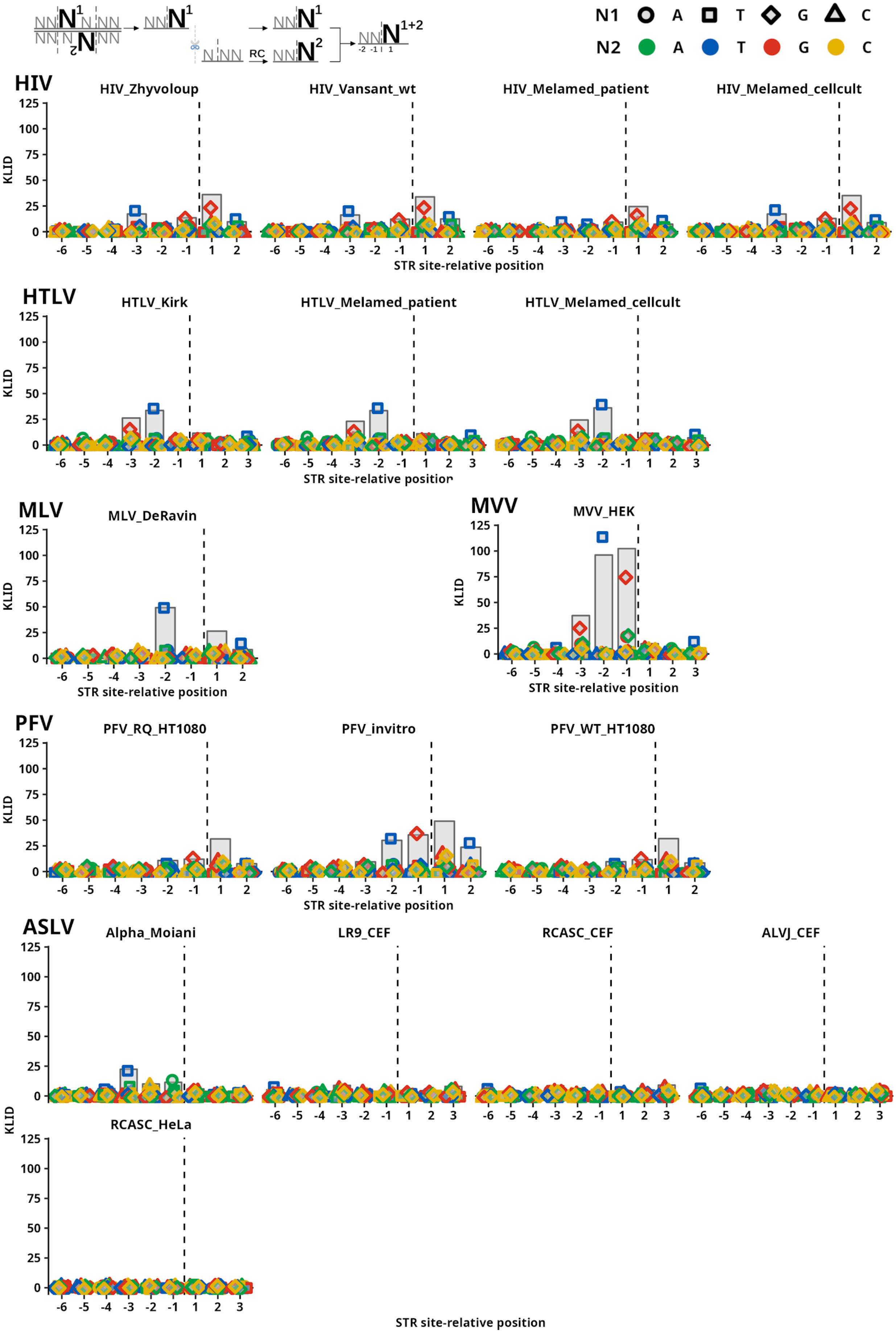
SUPPLEMENTAL FIGURES



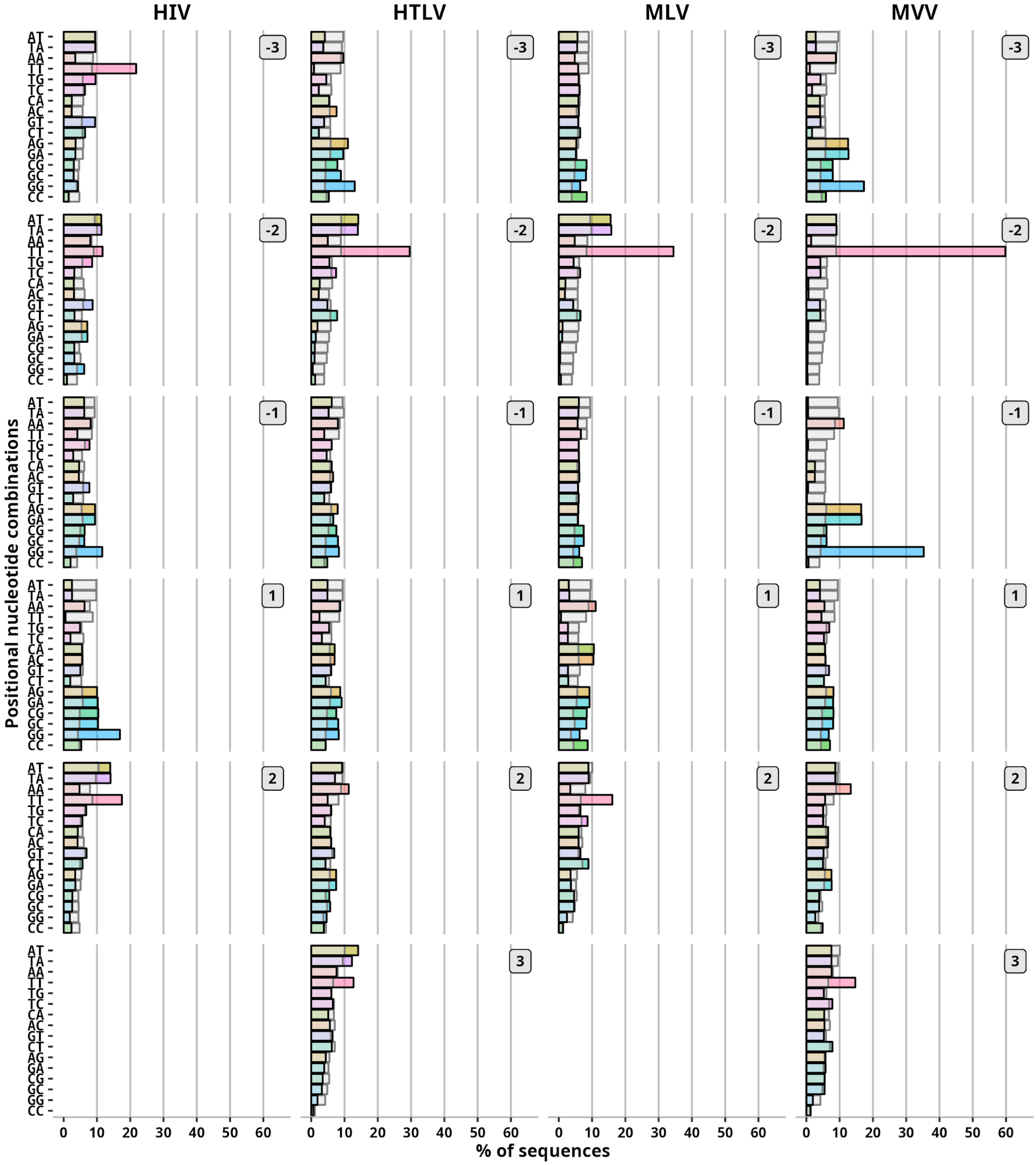
Supplemental Figure 1. Representation of palindromic defect (PDef) of component PPMs. The PDef corresponds to the distance from self-reverse-complementarity and is equal to zero for palindromic PPMs. All PPMs for each retroviral IS component mixture are represented. On the x-axis, mixtures are ordered by the number of components from two-component (M02) to eight-component (M08) mixture. The PDef of the complete IS set (PPM0, logos in Fig. 1A) is included as a dark circle in each column.

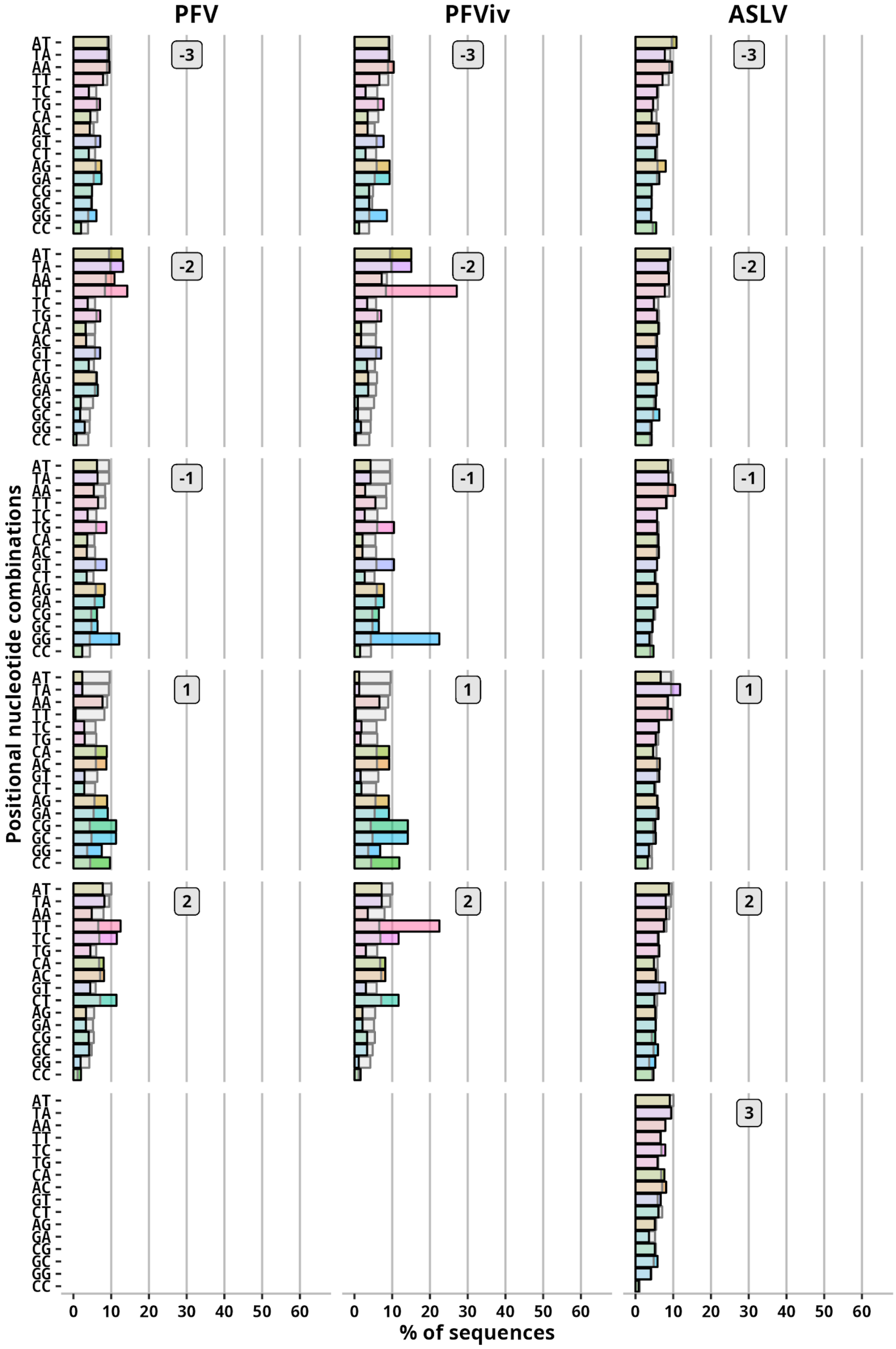


Supplemental Figure 2. Sequence logos of selected mixture components. A) Sequence logos derived from PPMs representing complete sets of retroviral IS and set of random genomic sequences. Panels B) and C) show sequence logos representing B) component PPMs with the lowest observed PDef across all retroviral IS mixtures (i.e. the most palindromic components) and C) component PPMs with the highest observed PDef across all retroviral IS mixtures (i.e. the most reverse-complement asymmetric components; on top) together with the PPMs displaying the lowest reverse-complementary distance to the asymmetric PPMs. Each sequence logo contains the name of the mixture on the top, the component name in the top left corner, and the component weight multiplied by 100 in the top right corner. Logos represent IS sequences spanning 13 nucleotides to each side from the center of the sequences. Vertical dashed lines mark sites where strand transfer reaction takes place.

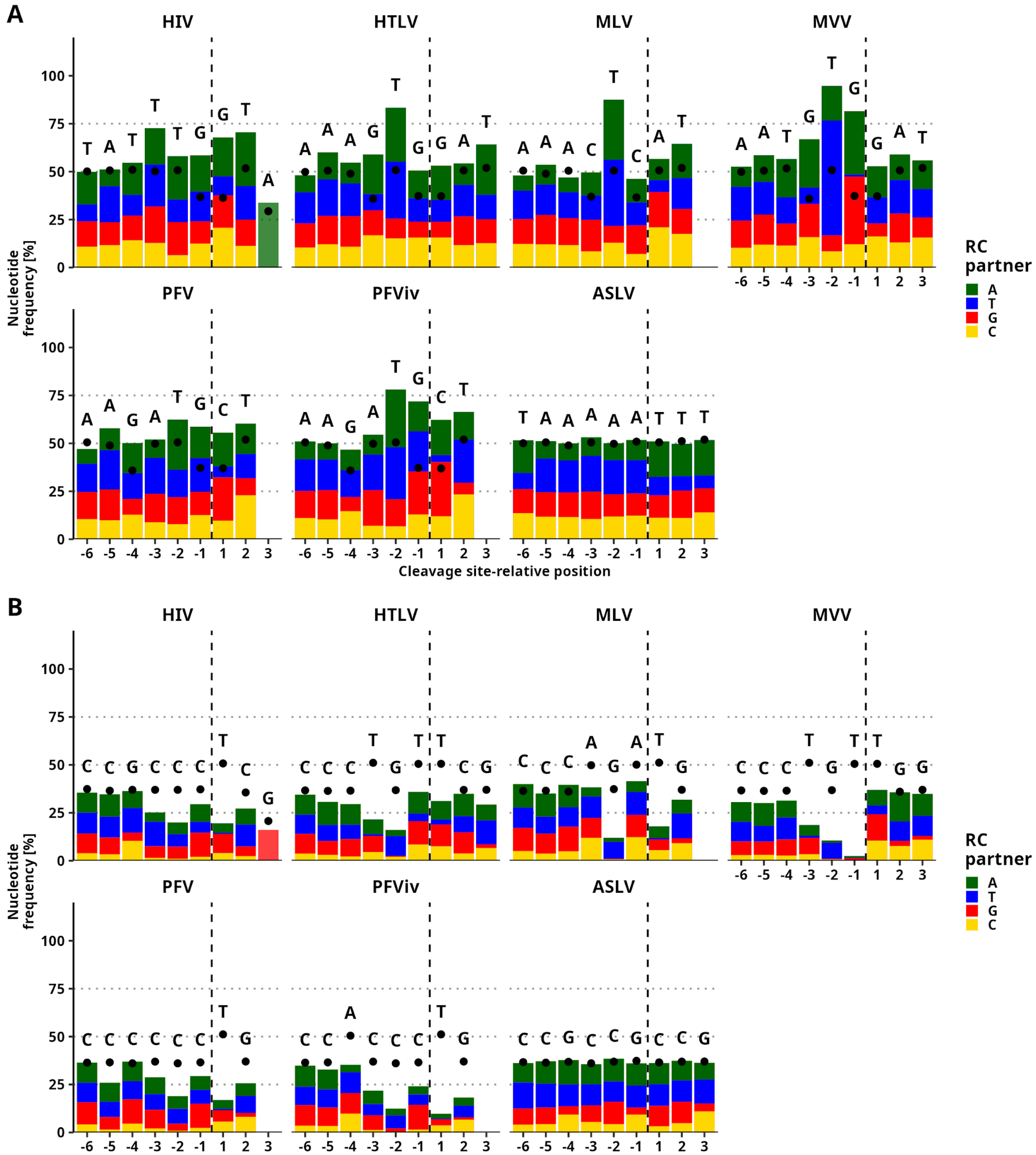

NN N1 N NN
NN N NN N2
NN N1
N NN
RC
NN N1
NN N2
NN N1+2
-2 -1 1
N1 A T G C
N2 A T G C
HIV
HIV_Zhyvoloup
HIV_Vansant_wt
HIV_Melamed_patient
HIV_Melamed_cellcult
HTLV
HTLV_Kirk
HTLV_Melamed_patient
HTLV_Melamed_cellcult
MLV
MLV_DeRavin
MVV
MVV_HEK
PFV
PFV_RQ_HT1080
PFV_invitro
PFV_WT_HT1080
ASLV
Alpha_Moiani
LR9_CEF
RCASC_CEF
ALVJ_CEF
RCASC_HeLa
KLID
125
100
75
50
25
0
-6 -5 -4 -3 -2 -1 1 2 3
STR site-relative position

Supplemental Figure 3. KLID representation of enrichment for positional nucleotide combinations. KLID score of dinucleotide combinations at complementary sites marked by position related to the STR site. The position of the STR site is marked by the vertical dashed line. Positions upstream to the STR 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.

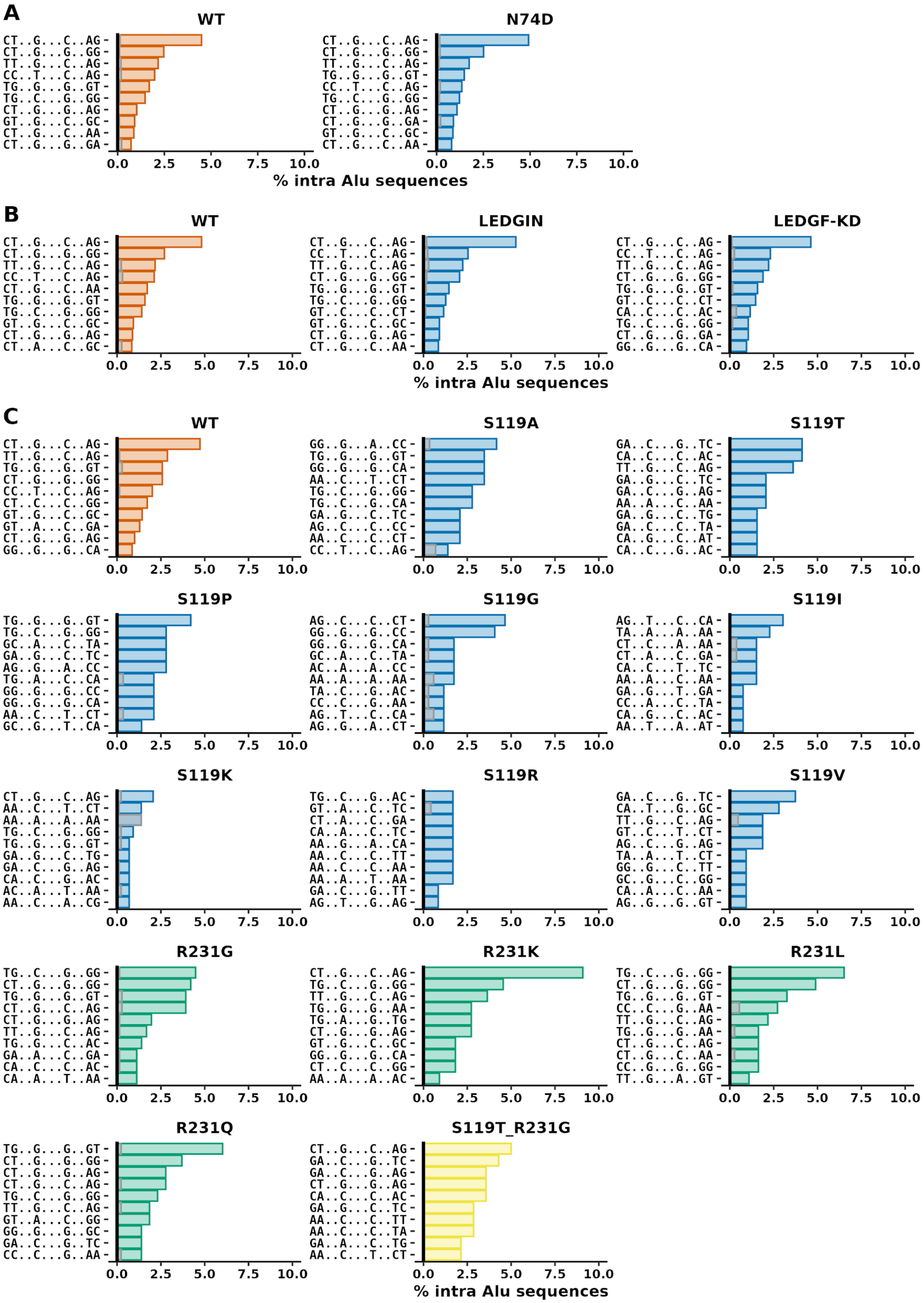

HIV
HTLV
MLV
MVV
AT
TA
AA
TT
TG
TC
CA
AC
GT
CT
AG
GA
CG
GC
GG
CC
-3
-2
-1
1
2
3
Positional nucleotide combinations
0
10
20
30
40
50
60
% of sequences



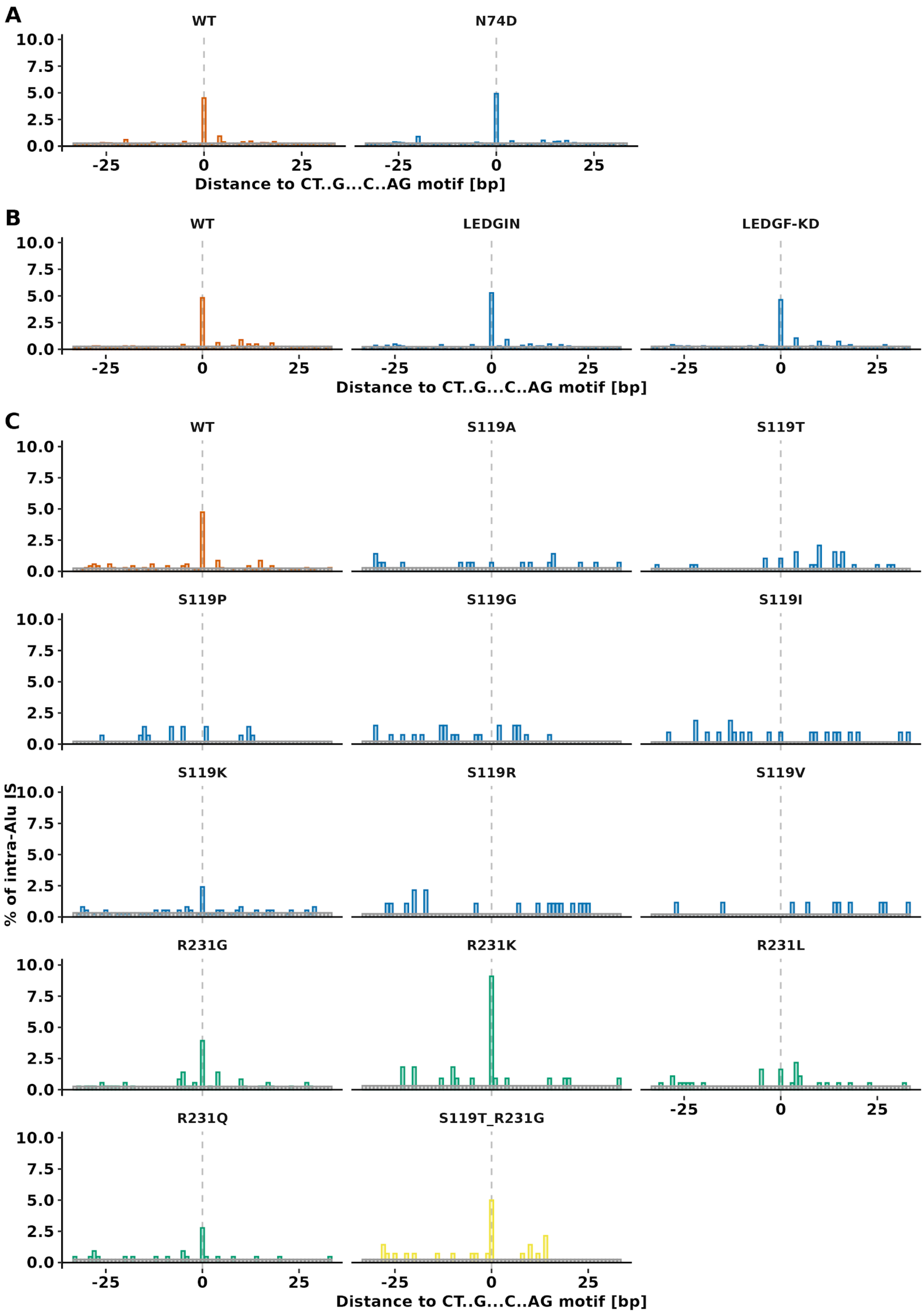
Supplemental Figure 4. Frequency of positional combinations. Frequency of sequences with marked positional nucleotide combination. Gray bars represent the frequency observed in a control (random) set of sequences. The numbers right of the bars show the position relative to the cleavage site. Positions upstream to the STR site are marked with negative values.



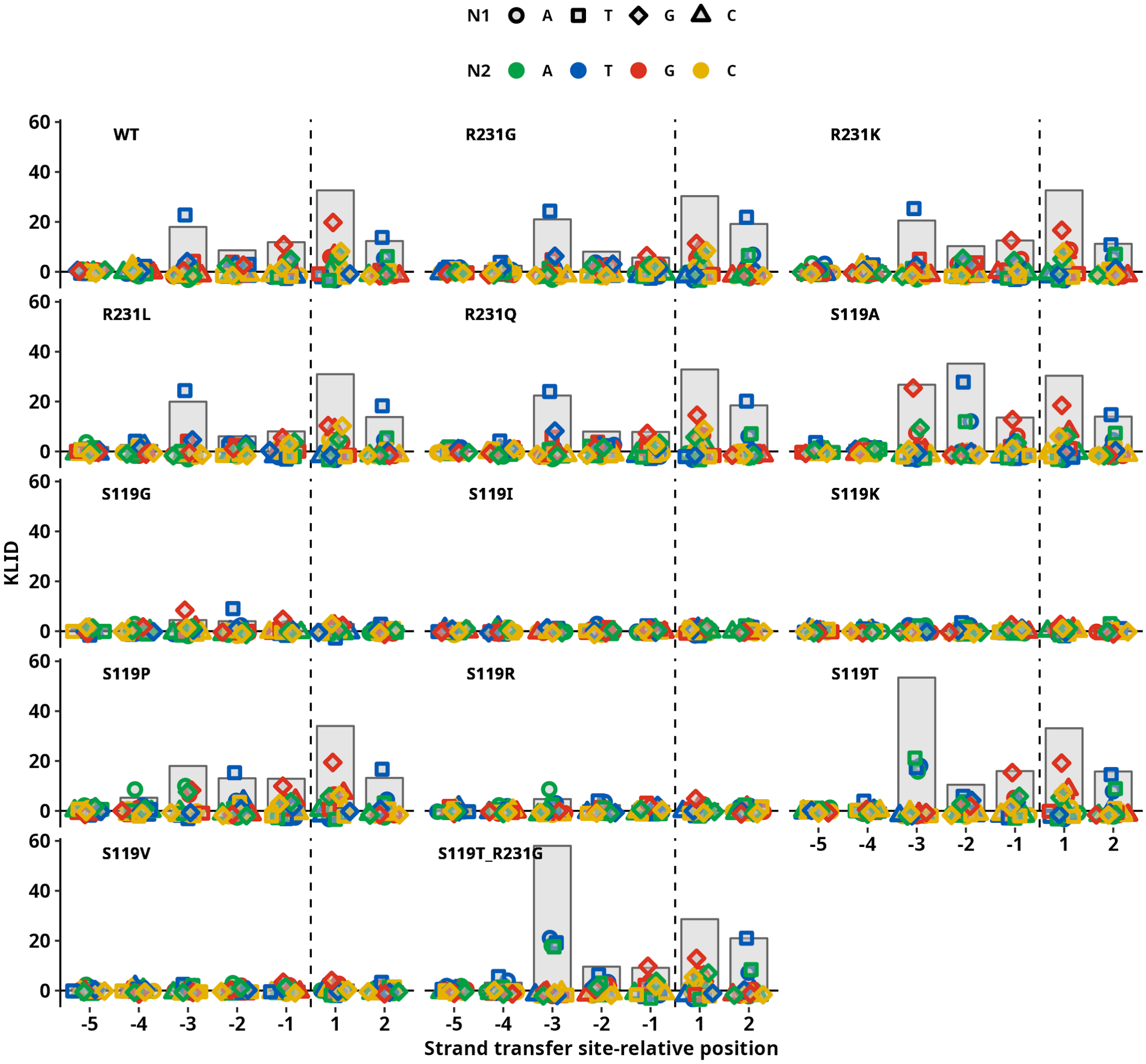
Supplemental Figure 5. The most and least frequent nucleotides at STR site-relative positions. A) The most frequent nucleotide. B) Least frequent nucleotides. Dots represent expected frequencies at the position. The height of colored bars represents the frequency of nucleotide present at the complementary site.


A
WT
CT..G...C..AG
CT..G...G..GG
TT..G...C..AG
CC..T...C..AG
TG..G...G..GT
TG..C...G..GG
CT..G...G..AG
GT..G...C..GC
CT..G...C..AA
CT..G...G..GA
N74D
CT..G...C..AG
CT..G...G..GG
TT..G...C..AG
TG..G...G..GT
CC..T...C..AG
TG..C...G..GG
CT..G...G..AG
CT..G...G..GA
GT..G...C..GC
CT..G...C..AA
0.0
2.5
5.0
7.5
10.0
% intra Alu sequences
B
WT
CT..G...C..AG
CT..G...G..GG
TT..G...C..AG
CC..T...C..AG
CT..G...C..AA
TG..G...G..GT
TG..C...G..GG
GT..G...C..GC
CT..G...G..AG
CT..A...C..GC
LEDGIN
CT..G...C..AG
CC..T...C..AG
TT..G...C..AG
CT..G...G..GG
TG..G...G..GT
TG..C...G..GG
GT..C...C..CT
GT..G...C..GC
CT..G...G..AG
CT..G...C..AA
LEDGF-KD
CT..G...C..AG
CC..T...C..AG
TT..G...C..AG
CT..G...G..GG
TG..G...G..GT
GT..C...C..CT
CA..C...C..AC
TG..C...G..GG
CT..G...G..GA
GG..G...G..CA
0.0
2.5
5.0
7.5
10.0
% intra Alu sequences
C
WT
CT..G...C..AG
TT..G...C..AG
TG..G...G..GT
CT..G...G..GG
CC..T...C..AG
CT..C...C..GG
GT..G...C..GC
GT..A...C..GA
CT..G...G..AG
GG..G...G..CA
S119A
GG..G...A..CC
TG..G...G..GT
GG..G...G..CA
AA..C...T..CT
TG..C...G..GG
TG..C...G..CA
GA..G...C..TC
AG..C...C..CC
AA..C...C..CT
CC..T...C..AG
S119T
GA..C...G..TC
CA..C...C..AC
TT..G...C..AG
GA..G...C..TC
GA..C...G..AG
AA..A...C..AA
GA..G...C..TG
GA..C...C..TA
CA..G...C..AT
CA..C...G..AC
S119P
TG..G...G..GT
TG..C...G..GG
GC..A...C..TA
GA..G...C..TC
AG..G...A..CC
TG..A...C..CA
GG..G...G..CC
GG..G...G..CA
AA..C...T..CT
GC..G...T..CA
S119G
AG..C...C..CT
GG..G...G..CC
GG..G...G..CA
GC..A...C..TA
AC..A...A..CC
AA..A...A..AA
TA..C...G..AC
CC..C...G..AA
AG..T...C..CA
AG..G...A..CT
S119I
AG..T...C..CA
TA..A...A..AA
CT..C...A..AA
CT..A...C..GA
CA..C...T..TC
AA..A...C..AA
GA..G...T..GA
CC..A...C..TA
CA..G...C..AC
AA..T...A..AT
S119K
CT..G...C..AG
AA..C...T..CT
AA..A...A..AA
TG..C...G..GG
TG..G...G..GT
GA..G...C..TG
GA..C...G..AG
CA..C...G..AC
AC..A...T..AA
AA..C...A..CG
S119R
TG..C...G..AC
GT..A...C..TC
CT..A...C..GA
CA..A...C..TC
AA..G...A..CA
AA..C...C..TT
AA..C...C..AA
AA..A...T..AA
GA..C...G..TT
AG..T...G..AG
S119V
GA..C...G..TC
CA..T...G..GC
TT..G...C..AG
GT..C...T..CT
AG..C...G..AG
TA..A...T..CT
GG..G...C..TT
GC..G...C..GG
CA..A...C..AA
AG..G...G..GT
R231G
TG..C...G..GG
CT..G...G..GG
TG..G...G..GT
CT..G...C..AG
CT..G...G..AG
TT..G...C..AG
TG..G...C..AC
GA..A...C..GA
CA..C...C..AC
CA..A...T..AA
R231K
CT..G...C..AG
TG..C...G..GG
TT..G...C..AG
TG..G...G..AA
TG..A...G..TG
CT..G...G..AG
GT..G...C..GC
GG..G...G..CA
CT..C...C..GG
AA..A...A..AC
R231L
TG..C...G..GG
CT..G...G..GG
TG..G...G..GT
CC..G...G..AA
TT..G...C..AG
TG..G...G..AA
CT..G...C..AG
CT..G...C..AA
CC..G...G..GG
TT..G...A..GT
R231Q
TG..G...G..GT
CT..G...G..GG
CT..G...G..AG
CT..G...C..AG
TG..C...G..GG
TT..G...C..AG
GT..A...C..GG
GG..G...G..GC
GA..C...G..TC
CC..C...G..AA
S119T_R231G
CT..G...C..AG
GA..C...G..TC
GA..C...G..AG
CT..G...G..AG
CA..C...C..AC
GA..G...C..TC
AA..C...C..TT
AA..C...C..TA
GA..A...C..TG
AA..C...T..CT
0.0
2.5
5.0
7.5
10.0
% intra Alu sequences

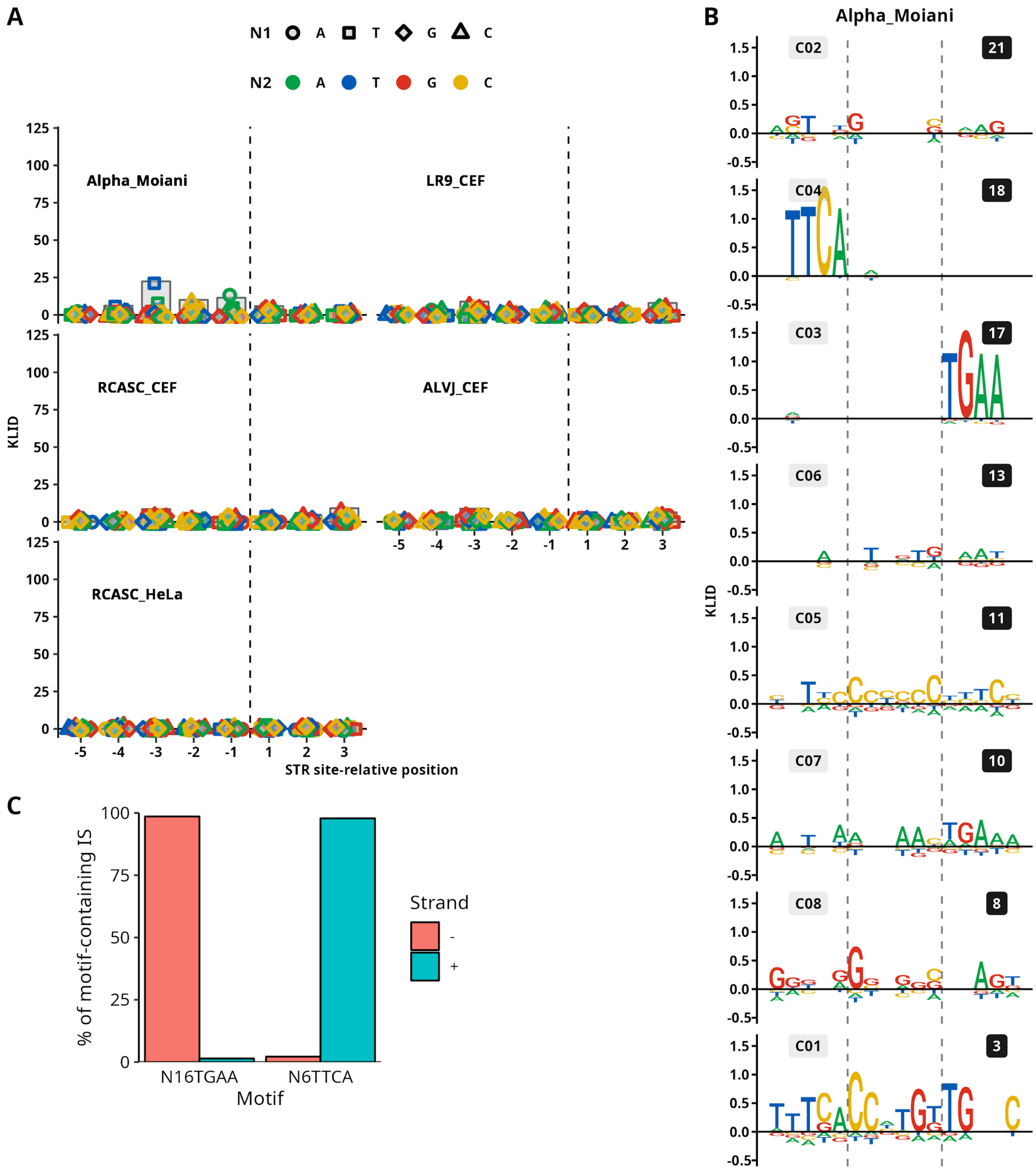
Supplemental Figure 6. Frequency of motifs at intra-Alu IS of HIV integrations sites. Frequency of the sequence motifs among intra-Alu IS. Gray bars represent the mean frequency of 100 iterations of shuffled controls created for each IS set. The ten most frequent sequence motifs are shown. Panels A) and B) show data from IS sets where integration retargeting was achieved by interruption of capsid-CPSF6 A) or IN-LEDGEF/p75 B) interactions. Panel C) shows data from IS sets where mutations of S119 of R231 were introduced to HIV-1 IN. Each set contains its own set of control (wt) with unaffected integration.


A
WT
N74D
10.0
7.5
5.0
2.5
0.0
-25
0
25
Distance to CT..G...C..AG motif [bp]
B
WT
LEDGIN
LEDGF-KD
Distance to CT..G...C..AG motif [bp]
C
WT
S119A
S119T
S119P
S119G
S119I
S119K
S119R
S119V
R231G
R231K
R231L
R231Q
S119T_R231G
% of intra-Alu IS
Distance to CT..G...C..AG motif [bp]

Supplemental Figure 7. Frequency of HIV intra-Alu IS in palindromic CT..G…C..AG motif. Frequency intra-Alu IS inside and in close proximity to CT..G…C..AG palindromic motif. Gray bars represent the mean frequency of 100 iterations of shuffled controls created for each IS set. Bar plots show the frequency of intra-Alu IS 33 bp down- and upstream to the motif relative to *Alu* repeat orientation. Frequencies are binned by 1 bp. Panels A) and B) show data from IS sets where integration retargeting was achieved by interruption of capsid-CPSF6 A) or IN-LEDGEF/p75 B) interactions. Panel C) shows data from IS sets where mutations of S119 of R231 were introduced to HIV-1 IN. Each set contains its own set of control (wt) with unaffected integration.



Supplemental Figure 8. KLID representation of enrichment for positional nucleotide combinations of HIV integrase variants. KLID score of dinucleotide combinations at complementary sites marked by position related to the STR site. The position of the STR site is marked by the vertical dashed line. Positions upstream to the STR 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. Data from IS sets where mutations of S119 of R231 were introduced to HIV-1 IN.



Supplemental Figure 9. Representations of IS sequence preferences of other ASLVs. A) Nucleotide combinations as positional KLID values. The data set named Alpha_Moiani comes from the publication by Moiani et al. 2014. Other data sets are from the work of Malhotra et al. 2017. Data sets represent distinct retroviruses or retroviral vectors transducing different cells. B) KLID logo representations of the M08 mixture components from the “Alpha_Moiani " data set. Components C04 and C03 form a reverse-complementary asymmetric pair with high KLID values upstream to the STR site. C) Sequences containing the “N16TGAA” or “N6TTCA” motif proximal to the STR site were identified in the IS data set (number represents the number of nucleotides preceding the motif). The coordinates of the motif-containing IS were identified. Surprisingly, the majority of IS of each group motif group was mapped to either strand of the reference genome. As this finding might result from some artifact introduced during the data processing, the IS set was omitted from further analysis.

SUPPLEMENTAL METHODS

Obtaining coordinates of integration sites (IS)

Retroviral integration sites were obtained from different resources in different formats. The following text describes the procedure leading from the original file to the number-encoded IS sequence file used as input for the EM algorithm.

HTLV

Sequences of pre-integration genomic sequences were obtained directly from supplemental data of the publication.

HIV

Zhyvoloup et al. 2017
IS coordinates were obtained from a publication supplemental table. Replicates of DMSO-treated samples of wt and capsid N74D mutant were joined to create wt and N74D IS sets. Coordinates of both IS sets were joined to create a single file.

Vansant et al. 2020
Coordinates were obtained from the GSE135295_ledgins1_integration_features table. Mock, LEDGIN+, and LEDGF-KD data sets were created joining bulk, GFP+, and GFP- samples from the same treatment group marked by the key present in field 8 of the feature table as follows: wt (S168, S169, S178), LEDGIN+ (S170, S171, S179), LEDGF-KD (S174, S175, S181).

Demeulemeester et al. 2016
Coordinates of IS in the form of text files were obtained upon request from the authors of the study. Data obtained by transduction of different cell types with identical variants were mixed. The IS coordinates of the IN variants were transformed to BED format with a custom script.

MLV

Raw reads were obtained from Sequence Read Archive using the fastq-dump --split-e command.
Sequences were trimmed using cutadapt with the following sequences to be trimmed:
LTR3nest=TGACTACCCGTCAGCGGGGGTC, LTR3rest=TTTCA,
LTR5nest=CAAACCTACAGGTGGGGTCTTTC, LTR5rest=A, Adaptor=AGTCCCTTAAGCGGAGCCCT.
First 5’/3’ LTR sequences from primer (LTR*nest) were trimmed with
cutadapt -m 11 -O 20 -e 0.1 -j 0 --trim-n .
The rest of each LTR (LTR*rest) was trimmed using
cutadapt -m X -O 20 -e 0 -j 0 ,
where the X in -m option equals to LTR*rest sequence length. The adapter was removed from sequences using
cutadapt -m 11 -O 10 -e 0.1 -j 0 .
Finally, sequences containing inner proviral sequences were removed using
cutadapt -m 11 -O 10 -e 0.1 -j 0 .
The resulting LTR5_F_full.fastq and LTR3_F_full.fastq were mapped to the hg38 human genome assembly using bowtie2 with *-p 20 -q --no-unal -x hg38* options. Next, alignments of mapped reads were filtered to start at position 0 with *grep -v "MD:Z:0"*. Alignments with a single hit in the genome were then sorted with the *grep -v "XS:i:"* command. The final BED file was created with samtools view, bedtools bamtobed and sort commands.

MVV

BED formatted IS coordinates retrieved from MVV vector-infected HEK293T cells were downloaded from Gene Expression Omnibus (study accession GSE196042).

PFV

BED formatted IS coordinates were downloaded from Gene Expression Omnibus (study accession GSE97973).

ASLV

Malhotra *et* al. 2017

SAM formatted alignments were obtained from the authors of the study upon request. First, data from experiments with identical vectors and cells were joined (48h and 120h collection time points). For compatibility with downstream tools, SAM headers from hg19 (HeLa)/galGal4 (CEF) were added to alignments. Next, alignments of mapped reads were filtered to start at position 0 with *grep -v "MD:Z:0"*. Alignments with a single hit in the genome were then sorted with the *grep -v "XS:i:"* command. The final BED file was created with samtools view, bedtools bamtobed, and sort commands. BED ranges were converted to contain single LTR-proximal position with custom awk script:

```
awk '{if ($6 == "+") print $1"\t"$1":"$2"_"$6; else print $1"\t"$1":"$3"_"$6;}'
```

If the mapped strand was “+”, the start coordinate was used, otherwise the end coordinate was used. The awk command was followed by:

```
sort | uniq -c | awk '{print $3"\t"$1}'
```

to count occurrences of each IS. Only IS coordinates with more than one occurrence were selected for further analysis.

Moiani *et* al. 2014

Raw reads were obtained from Sequence Read Archive using the fastq-dump --split-e command. Sequences were trimmed using cutadapt with the following sequences to be trimmed:

```
LTR=TTGGTGTGCACCTGGGTTGATGGCCGGACCGTTGATTCCCTGACGACTACGAGCACCTGCATGAAGCAGAAGG
LTRend=^CTTCA
ADAPT=GTCCCTTAAC
MseADAPT=TTAGTCC
```

First, the majority of LTR sequence was trimmed:

```
cutadapt -g $LTR -m 20 -O 30 -e 0.1 -j 0 --trim-n .
```

Next, the end of LTR was trimmed if the sequence was exactly matching the LTRend:

```
cutadapt -g $LTRend -m 15 -O 5 -e 0 -j 0 .
```

Finally, the adaptor sequence was removed from the reads:

```
cutadapt -a $MseADAPT -m 15 -O 7 -e 0.1 -j 0 .
```

In adapter-containing sequences, the end of the read was modified to contain MseI site instead of the MseI-Adapter junction using sed:

```
sed 's/TTAGTCC$/TTAA/'  .
```

The reads were mapped to hg38 genome assembly using bowtie2 with *-p 20* parameter set.
In the next step, alignments were filtered using custom code. First, the header of the SAM file was removed with *grep -v "^@"* and only alignments starting at position 0 were selected using *grep -v "MD:Z:0"* and only reads with single alignment were further selected with *grep -v "XS:i:"*. The final BED file was created with SAMtools view, BEDtools bamtobed, and sort commands.
To calculate the frequency of each IS in the IS set, first, the awk was used to convert IS into ISIDs:

```
awk '{if ($6 == "+") print $1"\t"$1":"$2+1"_"$6; else print $1"\t"$1":"$3"_"$6;}'
```

and then counted with:

```
sort | uniq -c | awk '{print $3"\t"$1}' .
```

IS with at least 5 occurrences were selected and the BED coordinates were transformed to represent the central position of the target site with awk:

```
awk -v tsd=6 'BEGIN{half=tsd/2} {if ($4 =="+") print $1"\t"$2+half-1"\t"$2+half"\t"$4"\t"$5"\t"$6; else print
$1"\t"$2-half"\t"$2-half+1"\t"$4"\t"$5"\t"$6}' .
```

Creating ranges for IS sequences

BED-formatted files containing coordinates of LTR proximal nucleotides were transformed into BED-formatted ranges covering 13 bp to each side from the center of IS (center of the inter-STR area). For this purpose, a custom awk script was used. Owing to the different origin of each IS set, the custom codes may differ and thus individual codes for each IS set is presented. Generally, *tsd* is the length of the target site duplication for a given retrovirus (or floored half of the length for HIV and MLV IS), *dist* is the final length of the sequence and *xtr* is 0 if the length of the target site duplication is even and 1 if the length of the target site duplication is odd. ISFILE is an input file with IS coordinates and output is saved into *OUTFILE.bed* file and sorted using *sort -k 1,1 -k2,2n ${OUTFILE}.bed* for compatibility with downstream tools.

<u>HIV</u>

Zhyvoloup et al. 2017

INVAR is either wt or n74d. This variable selects the virus variant.

```
awk -v tsd=2 -v dist=26 -v xtr=1 -v invar=$INVAR 'BEGIN{start=(dist/2)+xtr-tsd; end=tsd+(dist/2)} $5 ==
invar && $3 == "+" {print $1"\t"$2-start"\t"$2+end"\t"$4"\t""1""\t"$4}' ${ISFILE} > ${OUTFILE}.bed
awk -v tsd=2 -v dist=26 -v xtr=1 -v invar=$INVAR 'BEGIN{start=(dist/2)+xtr+tsd; end=(dist/2)-tsd} $5 ==
invar && $3 == "-" {print $1"\t"$2-start"\t"$2+end"\t"$4"\t""2""\t"$4}' ${ISFILE} >> ${OUTFILE}.bed
```

Vansant et al. 2020

```
awk -v tsd=2 -v dist=26 -v xtr=1 'BEGIN{start=(dist/2)-xtr-tsd+1+1; end=(dist/2)+xtr+tsd-1} $6 == "-"
{print $1"\t"$2-start"\t"$2+end"\t"$4"\t"$5"\t"$6}' ${ISFILE} > ${OUTFILE}.bed
awk -v tsd=2 -v dist=26 -v xtr=1 'BEGIN{start=(dist/2)+xtr+tsd+1; end=(dist/2)-tsd-xtr} $6 == "+" {print
$1"\t"$2-start"\t"$2+end"\t"$4"\t"$5"\t"$6}' ${ISFILE} >> ${OUTFILE}.bed
```

Demeulemeester et al. 2016

```
awk -v tsd=2 -v dist=26 -v xtr=1 'BEGIN{start=(dist/2)+xtr-tsd; end=tsd+(dist/2)} $6 == "+" {print
$1"\t"$2-start+1"\t"$2+end+1"\t"$4"\t""1""\t"$4}' ${ISFILE} > ${OUTFILE}.bed
 awk -v tsd=2 -v dist=26 -v xtr=1 'BEGIN{start=(dist/2)+xtr+tsd; end=(dist/2)-tsd} $6 == "-" {print
$1"\t"$2-start"\t"$2+end"\t"$4"\t""2""\t"$4}' ${ISFILE} >> ${OUTFILE}.bed
```

<u>MLV</u>

```
awk -v tsd=2 -v dist=26 -v xtr=0 'BEGIN{start=(dist/2)+xtr-tsd; end=(dist/2)-tsd} {print
$1"\t"$2-start"\t"$3+end"\t"$4"\t""1""\t"$4}' ${ISFILE} > ${OUTFILE}.bed
```

<u>MVV</u>

```
awk -v dist=26 'BEGIN{start=(dist/2)-1; end=(dist/2)+1} {print $1"\t"$2-start"\t"$2+end"\t"$4"\t""1""\t"$6}'
${ISFILE} > ${OUTFILE}.bed
```

<u>PFV</u>

```
awk -v tsd=4 -v dist=26 -v xtr=0 'BEGIN{start=(dist/2)+xtr; end=(dist/2)} {print
$1"\t"$2-start"\t"$3+end-1"\t"$5"\t""1""\t"$4}' ${ISFILE} > ${OUTFILE}.bed
```

ASLV

Malhotra *et* al. 2017

```
awk -v tsd=6 -v dist=26 -v xtr=0 'BEGIN{start=(dist/2)+xtr-tsd; end=tsd+(dist/2)} $4 == "+" {print
$1"\t"$2-start"\t"$2+end"\t"$5"\t""1""\t"$5}' ${ISFILE} > ${OUTFILE}.bed
awk -v tsd=6 -v dist=26 -v xtr=0  'BEGIN{start=(dist/2)+xtr+tsd; end=(dist/2)-tsd} $4 == "-" {print
$1"\t"$2-start"\t"$2+end"\t"$5"\t""2""\t"$5}' ${ISFILE} >> ${OUTFILE}.bed
```

Moiani *et* al. 2014

```
awk -v tsd=6 -v dist=26 -v xtr=0 'BEGIN{start=(dist/2)+xtr; end=(dist/2)} {print
$1"\t"$2-start"\t"$3+end-1"\t"$5"\t""1""\t"$4}' ${ISFILE} > ${OUTFILE}.bed
```

Obtaining genomic sequences from IS ranges

BEDtools getfasta tool was used to retrieve genomic sequences. IS BED ranges and FASTA files with sequences of chromosomes served as input. In the resulting FASTA files, all nucleotides were converted to capitals with

```
cat ${ISSEQ}.fa | sed 's/a/A/g' | sed 's/c/C/g' | sed 's/g/G/g' | sed 's/t/T/g' > ${ISSEQ}_CAP.fa ,
```

Where ISSEQ is an output file from the getfasta tool.

The right alignment of the sequences was inspected through the classical sequence logo produced by the ggseqlogo R package.

Converting sequences to numeric strings

To use sequences as input for the EM algorithm, nucleotide sequences needed to be converted to a string of numbers. Thus, each nucleotide was given a numeric code. Sequences in the FASTA file were first converted to a table, where sequences are placed into a single column and each sequence occupies a single row:

```
grep -v ">" ${ISSEQ}_CAP.fa > ${ISSEQ}_is26.txt .
```

Finally, the sequences were converted to numeric strings:

```
sed 's/[Aa]/1/g' ${ISSEQ}_is26.txt | sed 's/[Cc]/2/g' | sed 's/[Gg]/3/g' | sed 's/[Tt]/4/g' | sed 's/[Nn]/5/g'
${ISSEQ}_is26_num.txt
```