*5164–5176* *Nucleic Acids Research, 2014, Vol. 42, No. 8* *Published online 11 February 2014*
*doi:10.1093/nar/gku136*

# Integrase residues that determine nucleotide preferences at sites of HIV-1 integration: implications for the mechanism of target DNA binding

**Erik Serrao[1,†], Lavanya Krishnan[1,†], Ming-Chieh Shun[1], Xiang Li[1], Peter Cherepanov[2,3], Alan Engelman[1,*] and Goedele N. Maertens[2,*]**

[1]Department of Cancer Immunology and AIDS, Dana-Farber Cancer Institute and Department of Medicine, Harvard Medical School, Boston, MA 02215, USA, [2]Division of Infectious Diseases, Imperial College London, St-Mary's Campus, Norfolk Place, London W2 1PG, UK and [3]Clare Hall Laboratories, London Research Institute, Cancer Research UK, Hertfordshire EN6 3LD, UK

Received November 24, 2013; Revised January 19, 2014; Accepted January 22, 2014

## ABSTRACT

**Retroviruses favor target-DNA (tDNA) distortion and particular bases at sites of integration, but the mechanism underlying HIV-1 selectivity is unknown. Crystal structures revealed a network of prototype foamy virus (PFV) integrase residues that distort tDNA: Ala188 and Arg329 interact with tDNA bases, while Arg362 contacts the phosphodiester backbone. HIV-1 integrase residues Ser119, Arg231, and Lys258 were identified here as analogs of PFV integrase residues Ala188, Arg329 and Arg362, respectively. Thirteen integrase mutations were analyzed for effects on integrase activity *in vitro* and during virus infection, yielding a total of 1610 unique HIV-1 integration sites. Purine (R)/pyrimidine (Y) dinucleotide sequence analysis revealed HIV-1 prefers the tDNA signature (0)RYXRY(4), which accordingly favors overlapping flexible dinucleotides at the center of the integration site. Consistent with roles for Arg231 and Lys258 in sequence specific and non-specific binding, respectively, the R231E mutation altered integration site nucleotide preferences while K258E had no effect. S119A and S119T integrase mutations significantly altered base preferences at positions −3 and 7 from the site of viral DNA joining. The S119A preference moreover mimicked wild-type PFV selectivity at these positions. We conclude that HIV-1 IN residue Ser119 and PFV IN residue Ala188 contact analogous tDNA bases to effect virus integration.**

## INTRODUCTION

Retroviral integrase (IN) enzymes catalyze the insertion of reverse-transcribed viral DNA (vDNA) into host chromosomal or target DNA (tDNA) as an essential step toward productive virus infection. The multistep integration process initiates with the formation of the stable synaptic complex or intasome, which is comprised of an IN tetramer and the two ends of linear vDNA (1–3). IN processes the vDNA ends adjacent to conserved CA sequences, which liberates a $pGT_{OH}$ dinucleotide from each 3′-end of HIV-1 DNA (4,5). The target capture complex (TCC) subsequently forms in the nucleus when the intasome engages tDNA (3). IN catalyzes the concerted joining of the $CA_{OH}$ ends to the 5′-phosphates of a staggered double stranded cut in tDNA (3,6,7). Repair of the single-stranded gaps at the vDNA–tDNA junctions yields the flanking duplication of the tDNA cut sequence, which varies from 4 to 6 bp among integrated retroviruses.

Although integration can occur throughout most of the animal cell genome (8), it is not random (9,10). There are seven retroviral genera (α through ε, lenti and spuma), and the different viruses differentially target chromatin features during integration. Lentiviruses such as HIV-1 prefer the bodies of active genes within gene-dense regions of chromosomes (11), whereas Moloney murine leukemia virus (MLV), a prototypical γ-retrovirus,

*To whom correspondence should be addressed. Tel: +44 20 7594 3655; Fax: +44 20 7594 3906; Email: g.maertens@imperial.ac.uk
Correspondence may also be addressed to Alan Engelman. Tel: +1 6176 324 361; Fax: +1 6176 324 338; Email: alan_engelman@dfci.harvard.edu
Present address:
Ming-Chieh Shun, Department of Molecular Biology and Microbiology, Case Western Reserve University, Cleveland, Ohio, USA.

†These authors contributed equally to the paper as first authors.

favors gene promoter regions (12). IN-binding host factors dictate these targeting preferences: bromodomain and extraterminal domain (BET) proteins were shown recently to mediate promoter proximal integration by MLV (13–15), while lens epithelium-derived growth factor (LEDGF)/p75 in large part dictates the lentiviral preference for active genes (16–18). Retroviruses also prefer particular nucleotides at sites of integration as evident by weakly conserved palindromic sequences that center on the tDNA cut (9,19–21). Integration site nucleotide preferences of lentiviruses are notably independent of cellular LEDGF/p75 content (17,18).

The X-ray crystal structure of the prototype foamy virus (PFV) TCC revealed that the intasome accommodates tDNA in a severely bent conformation (7). As predicted by the relatively weak nature of palindrome conservation at sites of retroviral integration, the majority of IN–tDNA contacts in the TCC were mediated through the phosphodiester backbone (7). IN is comprised of separate protein domains that include the N-terminal domain, catalytic core domain (CCD) and C-terminal domain (CTD) (22), and main chain amide groups of several CCD residues as well as the side chain of CTD residue Arg362 mediated interactions with the tDNA backbone. The side chains of two key PFV IN amino acids, Ala188 and Arg329, in contrast made contacts with tDNA bases. Consequently, recombinant Ala188 and Arg329 IN mutant proteins displayed DNA-strand-transfer defects and selected for novel nucleotide preferences at sites of PFV integration *in vitro* (7). Based on these observations, we hypothesized that HIV-1 IN amino acids that interact with tDNA bases could be identified by comparing integration sites of mutant IN enzymes to the canonical integration sequence (–3)TDG↓*(G/V)TWA(C/B)*CHA(7) (written using standard International Union of Biochemistry base codes; the vertical arrow marks the position of vDNA plus-strand joining and the underline highlights the tDNA duplication, which is 5 bp for HIV-1) (20,21). Structure-based IN amino acid sequence alignments were perused to identify HIV-1 IN amino acids analogous to PFV IN residues Ala188, Arg329 and Arg362, and 13 mutations targeting these as well as nearby residues were tested for their effects on IN enzyme function, HIV-1 infection and nucleotide site preferences at sites of integration *in vitro* and in virally infected cells.

## MATERIALS AND METHODS

### Plasmids and protein purification

Hexahistidine ($His_6$)-tagged HIV-$1_{HXB2}$ IN was expressed in bacteria from pCPH6P-HIV1-IN (23). LEDGF/p75 was expressed in bacteria using pFT-1-LEDGF, which also yields N-terminal $His_6$-tagged protein (24). The single-round HIV-luciferase (Luc) reporter construct was pNLX.Luc(R-)ΔAvrII (25) whereas pCG-VSV-G was used to express vesicular stomatitis virus G (VSV-G) glycoprotein (17). Mutations introduced by PCR using *Pfu* Ultra DNA polymerase (Agilent Technologies, Inc.) were verified by DNA sequencing. Plasmid pGEM-3 or pGEM9zf(-) served as tDNA in *in vitro* concerted integration reactions (23,26).

IN and LEDGF/p75 were expressed and purified from bacteria essentially as previously described (24,27) and the $His_6$ tags were removed by proteolysis with human rhinovirus 3C protease (GE Healthcare). Purified MuA transposase protein was a kind gift from Dr Michiyo Mizuuchi, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health (NIH).

### IN activity assays and integration product sequencing

*In vitro* assays for quantification of $Mg^{2+}$-dependent HIV-1 IN 3′-processing, DNA-strand transfer and concerted integration activities were performed as previously described (26–28). Concerted vDNA strand transfer reaction products were isolated, sub-cloned and sequenced essentially as previously reported (7).

### Cells, viruses and infections

HEK293T and SupT1 cells were propagated in Dulbecco's modified Eagle medium and RPMI 1640 (Gibco—Life Technologies), respectively, supplemented to contain 10% fetal bovine serum, 100 IU/ml penicillin and 100 μg/ml streptomycin. HEK293T cells were co-transfected with pNLX.Luc(R-)ΔAvrII and pCG-VSV-G at the mass ratio of 10:1 to produce single-round HIV-Luc pseudotypes. Viral production was monitored using a p24 antigen capture immunoassay (ABL, Inc.), and SupT1 cells ($4 \times 10^5$) were infected with 5 ng/ml p24 of wild-type (WT) or IN mutant virus in triplicate in 96-well plates. Luc values, expressed as percent WT relative light units, were determined 48 h post-infection.

### Viral integration site cloning

SupT1 cells ($5 \times 10^6$) in 6-well plates were spinoculated with WT or mutant virus preparations at 150 ng/ml p24 for 2 h, incubated for an additional 4 h and then washed, resuspended in 75 $cm^2$ flasks and cultured for 48 h. DNA was extracted with the DNeasy Blood and Tissue Kit (Qiagen), and integration sites were amplified using either restriction enzyme digestion (17) or bacteriophage Mu transposition-based (29) protocols essentially as described previously. Genomic DNA was digested at 37°C overnight with 100 U each of AvrII, SpeI and NheI, purified with the QIAquick PCR Purification Kit (Qiagen) and ligated to a double-stranded linker consisting of AE5237 (5′-[$PO_4^-$]CTAGGCAGCCCG[AmC7-Q]) and AE5238 (5′-GTAATACGACTCACTATAGGGCA CGCGTGGTCGACGGCCCGGGCTGC) (30). The DNA was PCR-amplified using primers AE5239 (5′-GA GGGATCTCTAGTTACCAGAGTCACA) and AE5240 (5′-GACTCACTATAGGGCACGCGT), diluted 1:200, and subjected to a second PCR round using primers AE5241 (5′-AGCCAGAGAGCTCCCAGGCTCAGATC) and AE5242 (5′-GTCGACGGCCCGGGCTGCCTA). Alternatively, annealed Mu right-end adaptors AE4455 (5′-GTAATACGACTCACTATAGGGCTCCGC TTAAGGGACTGTTTTCGCATTTATCGTGAAACGC TTTCGCGTTTTTCGTGCGCCGCTTCA) and AE4456

(5′-TCGGATGAAGCGGCGCACGAAAAACGCGAAA GCGTTTCACGATAAATGCGAAAACA[AmC7-Q]) were incubated with MuA transposase (440 ng) and 250 ng XhoI-digested genomic DNA at 30°C for 2 h in buffer (12.5 µl) containing 25.8 mM Tris–HCl, pH 8.0, 68 mM NaCl, 1 mg/ml bovine serum albumin, 10 mM $MgCl_2$, 0.08 mM EDTA, 0.05% Triton X-100 and 15% glycerol. The DNA (2 µl) was PCR-amplified using AE4392 (5′-GTA ATACGACTCACTATAGGGC) and AE4395 (5′-GCAC CATCCAAAGGTCAGTGGATATCTG), diluted as above, and re-amplified using AE4393 (5′-AGGGCTCCG CTTAAGGGAC) and AE4394 (5′-GTGTGTGGTAGAT CCACAGATCAAGG). Purified second round PCR products (500 ng) from both protocols were incubated with 10 ng pCR4-TOPO (Life Technologies) for 30 min, followed by transformation of competent Top10 bacteria. Individual colonies seeded in 96-well plates in LB medium containing 100 µg/ml kanamycin were sequenced at Beckman Coulter using the T3 reverse primer or viral U3-specific AE4396 (5′-CCACAGATCAAGGATATCTT GTC).

### Quantitative PCR analysis of vDNA

SupT1 cells ($1 \times 10^6$/well of a 12-well plate) were spinoculated with 100 ng/ml p24 of DNase-treated virus for 2 h. Cells were washed and reseeded into 48-well plates at $2.5 \times 10^5$ cells/well. The concentration of cellular DNA extracted at 8, 24 and 48 h post-infection using the DNeasy Blood and Tissue Kit was measured by spectrophotometry, and normalized DNA levels were analyzed by quantitative PCR (qPCR). Primers and probes for quantification of late reverse transcription (LRT) products and integrated proviruses were as described (31). Plasmid pNLX.Luc(R-)ΔAvrII diluted in uninfected genomic DNA generated the LRT standard curve, whereas dilutions of DNA recovered from cells infected for 48 h with HIV-Luc served as the integration standard curve. DNA was prepared from parallel infections conducted in the presence of 10 µM efavirenz (NIH AIDS Research and Reference Reagent Program) to account for residual transfected plasmid DNA in the qPCRs, and these background values, which varied from 0.4 to 1.3%, were subtracted from experimental samples.

### Bioinformatic analysis of integration sites

Sample sizes required for statistically significant comparisons between the WT and random, or between WT and mutant IN-integration-site sequences, were calculated using a Cohen's *d* value of 0.8, desired statistical power level of 0.9, and probability level or *P*-value of 0.05 (32), which yielded 34 as the minimal number of unique sites needed.

Data derived from infected cells was processed as described (11,17) to remove all U3, linker- and vector-derived sequences, duplicate sequences and sequences that did not contain the processed 5′-TTAGCCCTT CCA U3 terminus. Matches to human DNA were identified using BLAT (UCSC Human Genome Project, February 2009 GRCh37/hg19 assembly) and judged acceptable if they contained >98% average identity over the entire length of the sequence and also yielded a unique best hit in BLAT ranking. Positions −5 to 4 were experimentally determined, while positions 5–9 were assumed from genomic sequences upstream of the mapped integration site.

Consensus nucleotide sequences were visualized using the WebLogo program (33). Differences in integration sites from random, which was calculated relative to the pGEM9zf(-) plasmid sequence for *in vitro* integration site analysis and relative to 10 000 computer-generated sites for cellular DNA analysis, were determined by chi-square as described (17,34). Nucleotide preferences of IN mutants were also compared to the WT sequences using chi-square analysis.

Purine (R)/pyrimidine (Y) dinucleotide content was calculated by counting the number of the four kinds of sequences (RY, YR, RR and YY) in bins of dinucleotides from positions −10 to 14; IN mutant analyses were confined to the same windows as the nucleotide analyses (dinucleotide bins −5 to 8). Dinucleotide frequencies were normalized to the total number of WT or IN mutant integration sequences and also to the dinucleotide content of pGEM-9Zf(-), which was calculated as 23.9% RY, 23.9% YR and 52.2% RR/YY from $10^6$ computer-generated integration sites. WT HIV-1 IN dinucleotide frequency was compared to these randomly generated *in silico* integration sites using chi square analysis, and the dinucleotide preferences of IN mutants were compared to the WT also using chi-square analysis.

## RESULTS

### Experimental strategy

The X-ray crystal structure of the PFV TCC revealed a network of protein–tDNA interactions mediated through IN main chain and side chain atoms. The subset of IN CCD residues that contacted tDNA through polypeptide backbone amides (7) were not considered here due to potential complications of interpreting the effects of side chain substitutions on the function of main chain protein atoms.

Previous structure-based PFV/HIV-1 amino acid sequence alignments (2,27) were analyzed to identify potential functional analogs of PFV IN residues Ala188, Arg329 and Arg362. Ala188 forms part of the short CCD α2 helix that additionally harbors Ala189, Phe190 and Thr191 (Figure 1A). Phe and Thr are conserved in the analogous HIV-1 IN α2 helix, where residues Ser119 and Asn120 align with PFV IN residues Ala188 and Ala189, respectively. Ser119 and Asn120 were accordingly targeted for mutagenesis. Although the CTD is the least conserved domain among retroviral IN proteins (22), Arg362 aligned with HIV-1 IN residue Lys258 at the same relative position within CTD β4 (Figure 1B). Arg329 forms part of the loop that connects CTD β1 and β2 strands, which is four residues longer in PFV IN than the analogous loop in HIV-1 IN (2,27) (Figure 1B). Although HIV-1 IN residue Arg231 could be aligned with Arg329, adjacent residues Asp229, Ser230 and Asn232 were additionally targeted due to potential ambiguity in this region of the sequence

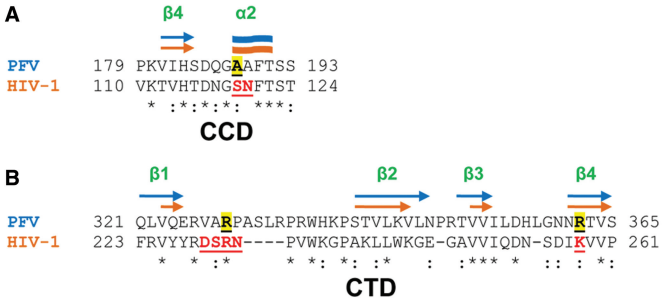


**Figure 1.** PFV and HIV-1 IN sequence alignments. (**A**) Structure-based amino acid sequence alignment of IN CCDs with secondary elements (α and β represent α helix and β strand, respectively) noted atop the sequences. PFV assignments (upper) are from protein database (PDB) code 3OY9, whereas the HIV-1 elements (lower) are from reference (27). (**B**) Sequence alignment of IN CTDs. PFV IN residues Ala188, Arg329 and Arg362 in panels (A) and (B) are highlighted in yellow and underlined. HIV-1 IN residues targeted for mutagenesis are underlined and highlighted in red. Positions of amino acid identity are marked by asterisk, while colons mark positions of chemical similarity.

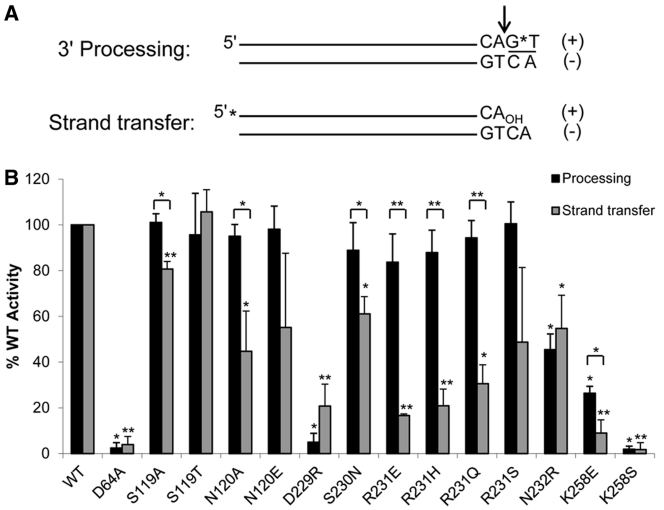


**Figure 2.** HIV-1 IN 3′-processing and DNA-strand-transfer substrate design and enzyme activities. (**A**) Substrate cartoons are shown to highlight the different U5 DNA +strand termini and positions of radiolabel (*). The vertical arrow and underline in the 3′-processing substrate mark the scissile phosphodiester bond and cleaved $pGT_{OH}$ dinucleotide, respectively. (**B**) IN mutant 3′-processing (black) and strand-transfer (gray) activities expressed as percentages of WT IN function. Results are averages ± standard deviation (SD) for three experimental replicates. Paired analyses revealed mutant activities that differed significantly from the WT ($*P < 0.05$; $**P < 0.01$) as well as individual mutant IN strand-transfer activities that differed significantly from the level of 3′-processing activity (asterisks above brackets).

alignment. The following mutations were engineered into a bacterial expression vector to assess effects on recombinant HIV-1 IN enzyme activity and integration site sequence preferences *in vitro*: S119A, S119T, N120A, N120E, D229R, S230N, R231E, R231H, R231Q, R231S, N232R, K258E and K258S.

### Biochemical activities of recombinant IN proteins

$His_6$-tagged IN proteins were purified from bacterial extracts using $Ni^{2+}$-nitrilotriacetate chromatography, and the $His_6$ tag was removed by site-specific proteolysis prior to IN activity assays. IN 3′-processing and DNA-strand-transfer activities were initially analyzed using relatively short (21 and 30 bp, respectively) mimics of the viral U5 end. The blunt ended 3′-processing substrate was labeled at the internal phosphate of the $pGT_{OH}$ dinucleotide to afford quantitation of 3′-processing activity independent from DNA strand transfer activity, and the DNA-strand-transfer substrate was pre-processed to allow this activity to be analyzed independent of 3′-processing activity (Figure 2A). Separate 30-bp molecules serve as vDNA and tDNA substrates under these strand transfer reaction conditions (35).

IN containing the D64A active-site mutation was expressed and purified for use as a negative control in enzyme activity assays (27). Because the profile of K258S IN mutant 3′-processing and DNA-strand-transfer activities mirrored those of D64A IN, K258S IN was at best minimally active (Figure 2B). In contrast, S119T IN supported the WT level of the IN 3′-processing and DNA-strand-transfer activities. Whereas numerous additional INs, including S119A, N120A, N120E, S230N, R231E, R231H, R231Q and R231S, also supported the WT level of 3′-processing activity, each mutant displayed between a 20% and 80% reduction in DNA-strand-transfer activity. Both the 3′-processing and DNA-strand-transfer activities of N232R IN were reduced ~2-fold from the WT, whereas mutants D229R and K258E were more defective, supporting between ~5 and 25% of the levels of WT IN activity (Figure 2B).

### Sequence analysis of *in vitro* integration products

We next sought to determine tDNA nucleotide preferences for the mutant enzymes that supported appreciable levels of IN-strand-transfer activity. Although the prior sequencing-gel-based assay supports strand-transfer activity, the relatively short tDNA oligonucleotide provides a suboptimal degree of sequence heterogeneity to query all four nucleotides at multiple positions of vDNA joining. We accordingly included supercoiled plasmid DNA as an integration target in the reaction mixture, which additionally distinguishes products that form through the integration of a single vDNA end from those that form by the concerted integration of two vDNA ends. For reasons that are not entirely clear, HIV-1 IN preferentially integrates single oligonucleotide vDNAs into tDNA, which yields nicked plasmid circles that co-migrate through agarose gels with open circular plasmids isolated from *Escherichia coli*. Concerted integration of two vDNA ends by contrast yields linear plasmid DNA products (Figure 3A). The system therefore affords analysis of concerted integration reaction products that harbor proper 5-bp tDNA duplications (28,36,37). As we previously established that the addition of LEDGF/p75 protein significantly enhanced the concerted integration activity of HIV-1 IN *in vitro* (28), mutant IN activities were compared to the WT in both the presence and absence of the integration cofactor.

In the absence of LEDGF/p75, WT IN displayed a basal level of half-site integration activity (Figure 3A, compare lanes 2 to 1). As expected (28), LEDGF/p75 boosted the formation of half-site and concerted vDNA

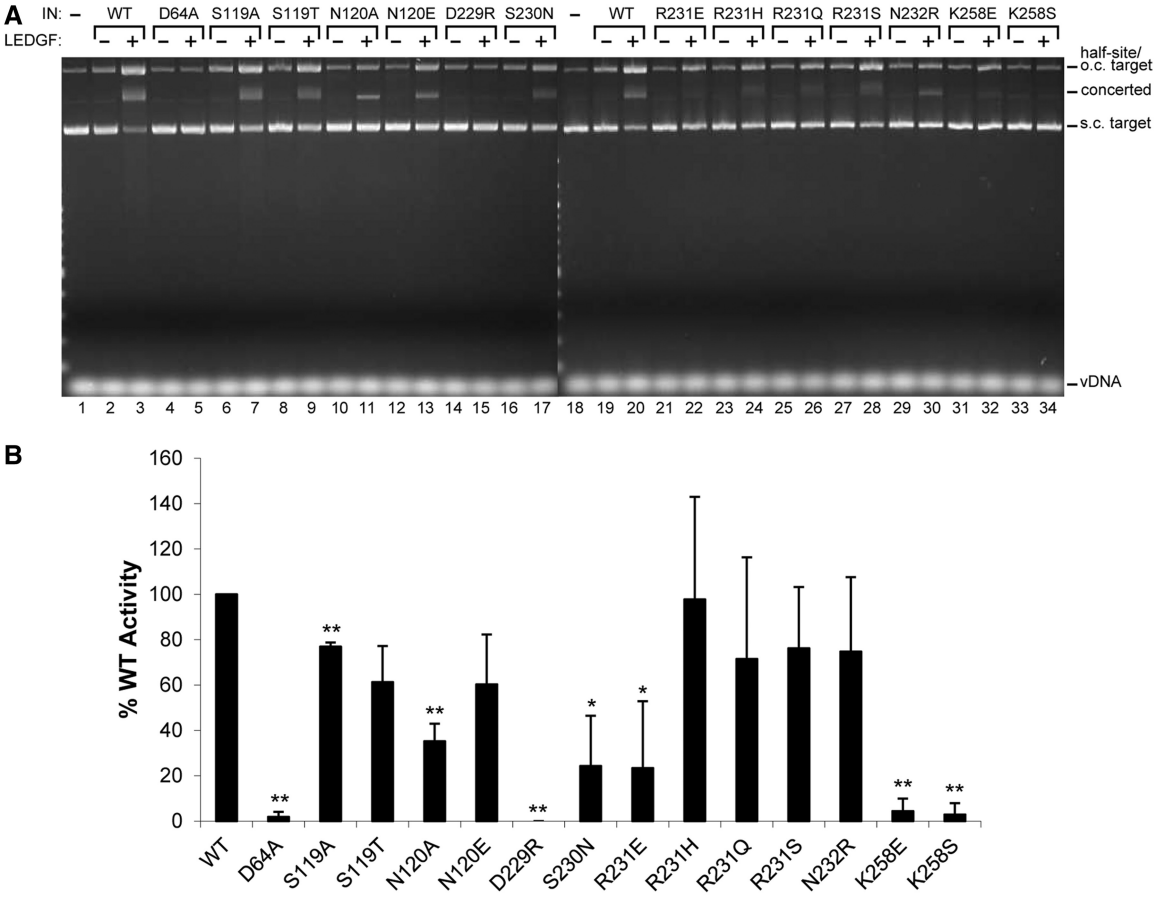


**Figure 3.** Concerted vDNA integration activities of HIV-1 IN proteins. (**A**) The agarose gel image highlights migration positions of the pre-processed 32-bp vDNA substrate, supercoiled (s.c.) and open circular (o.c.) forms of pGEM-3 tDNA, as well as products of half-site and concerted vDNA integration. The reactions loaded in lanes 1 and 18 omitted IN protein; LEDGF/p75 was included in each set of reactions as indicated. (**B**) Results (average ± SD for $n = 3$ experiments) of HIV-1 IN mutant concerted integration activities normalized to WT, which was set at 100%. Asterisks highlight significant differences from the WT as defined in Figure 2.

integration products significantly (compare lanes 3 to 2). Similar to the WT enzyme, the formation of IN mutant concerted integration products was LEDGF/p75-dependent (Figure 3A). Relative levels of IN mutant concerted integration activities in large part mirrored the DNA-strand-transfer activity levels observed in the sequencing gel-based assay (compare Figure 3B to Figure 2B, grey bars).

LEDGF/p75-dependent integration reactions were scaled up 30- to 90-fold (for the minimally active K258E mutant), and linear DNA products isolated from agarose gels were ligated to a kanamycin resistance cassette as previously described (7). Due to their relatively low levels of strand transfer activity (Figures 2 and 3), IN mutants D229R and K258S were omitted from this analysis. Plasmids isolated from individual bacterial colonies that released an insert of expected size upon restriction enzyme digestion were subjected to dideoxy sequencing using primers that faced outward from the kanamycin cassette. The total number of sequences that contained two vDNA ends varied from a low of 60 for K258E IN to a high of 170 for the R231S mutant (Table 1). About 83% of the WT sequences harbored 5-bp duplications, while ~9% and 7% harbored deletions or duplications other than 5 bp, respectively (Table 1). Most of the mutant enzymes yielded 5-bp duplications at frequencies similar to the WT, with notable exceptions of S119A and S119T INs. The frequencies of 5-bp duplications for these enzymes hovered ~60%, with concomitant increases in the number of product DNAs that harbored deletions and aberrant duplications of plasmid sequences (Table 1).

Site preferences were tabulated from concerted vDNA integration products that contained unique 5-bp duplications, which included 1090 WT and IN mutant sequences (Table 1). Observed nucleotides were compared to the frequency expected at each position based on the sequence of the tDNA plasmid, and *P*-values were calculated by $\psi^2$-analysis. Our dataset recapitulated the preference of WT IN for TDG↓*(G/V)TWA(C/B)*CHA, with the observed frequency at each nucleotide position differing significantly from random (Figure 4 and Supplementary Figure S1). The integration sites of the mutant enzymes were additionally compared to the base preference of the WT enzyme at each position. Each mutant that contained an alteration of a CCD α2 residue notably displayed a novel tDNA sequence preference (Figure 4 and Supplementary Figure S1). S119A and S119T IN each selected for novel

**Table 1.** WT and IN mutant *in vitro* integration products

| IN | Concerted integration products | Five-bp duplication (%) | Unique 5-bp duplication (%) | Deletions (%) | Other duplications (%) |
|---|---|---|---|---|---|
| WT | 163 | 136 (83.4) | 122 (74.8) | 15 (9.2) | 12 (7.4) |
| S119A | 168 | 97 (57.7) | 92 (54.8) | 44 (26.2) | 27 (16.1) |
| S119T | 159 | 94 (59.1) | 87 (54.7) | 40 (25.2) | 25 (15.7) |
| N120A | 168 | 149 (88.7) | 131 (77.9) | 8 (4.8) | 11 (6.5) |
| N120E | 163 | 131 (80.4) | 120 (73.6) | 15 (9.2) | 17 (10.4) |
| S230N | 86 | 70 (81.4) | 65 (75.6) | 4 (4.7) | 12 (14.0) |
| R231S | 170 | 131 (77.1) | 119 (70.0) | 23 (13.5) | 16 (9.4) |
| R231E | 129 | 124 (96.1) | 112 (86.8) | 2 (1.6) | 3 (2.3) |
| R231Q | 82 | 66 (80.5) | 65 (79.3) | 7 (8.5) | 9 (11.0) |
| R231H | 80 | 69 (86.3) | 66 (82.3) | 7 (8.6) | 4 (5.0) |
| N232R | 83 | 64 (77.1) | 62 (74.7) | 7 (8.4) | 12 (14.5) |
| K258E | 60 | 58 (92.1) | 49 (77.8) | 0 (0) | 2 (3.3) |

nucleotides at position 7: S119A IN preferred cytosine over adenosine ($P = 2.5 \times 10^{-8}$) whereas S119T IN preferred thymidine with a bias against guanosine ($P = 1.3 \times 10^{-10}$). Compared to the WT, S119A IN additionally favored adenosine and disfavored cytosine at position 6 ($P = 1.6 \times 10^{-4}$). N120A IN revealed a bias for thymidine at position 6 ($P = 0.003$), while N120E IN displayed a marginal preference for guanosine at position 4 ($P = 0.04$). Two of the mutant enzymes with changes in the loop region between CTD β1 and β2, R231E and R231H, also displayed modest preferences for guanosine at position 4 (*P*-value differences of 0.013 and 0.04 versus the WT, respectively). In contrast, the nucleotide sequence preferences of S230N, R231Q, R231S, N232R and K258E INs did not differ significantly from the WT (Figure 4 and Supplementary Figure S1).

The X-ray structure of the PFV TCC demonstrated that tDNA is severely bent to accommodate the scissile phosphodiester bonds at the IN active sites. This distortion is enabled by the unstacking of the two central base pairs at positions 1 and 2 from the site of vDNA joining (7). Pyrimidine (Y)-purine (R) dinucleotides display lower base-stacking properties than RR and YY, or the most rigid RY dinucleotide (38), and YR dinucleotides are accordingly favored at positions 1 and 2 during PFV integration (7). PFV integration yields a 4-bp duplication of tDNA sequences (39). Because HIV-1 integration yields 5-bp duplications, we reasoned the mechanism of tDNA bending might very well differ from that of PFV. The 25 nucleotides that span positions −10 to 14 of the WT HIV-1-integration sites were grouped into 24 dinucleotide bins (Figure 5A and B). The frequencies of RR and YY dinucleotides generally hovered around the combined random average of 52.2% (calculated from one million computer-generated integration sites), with points of significant difference at bins −2 and 5. Greater frequency alterations were however noted for the rigid RY signature at bins 0 and 3 surrounding the central base pair at the site of integration. Thus, ∼47% of the bin 0 and bin 3 dinucleotides were RY, practically doubling the unbiased frequency of 23.9% (Figure 5B; replotted as fractional RY usage in panel C). Concomitantly, a significant decrease to ∼5% of the most flexible YR dinucleotide was observed at these positions. RY and YR dinucleotide frequencies settled toward the random 23.9% value further away from the site of integration. Bin 0 and bin 3 RY dinucleotides notably increase the frequency of flexible YR dinucleotides at nucleotide positions 1 and 2 and at positions 2 and 3, as the (0)RYXRY(4) signature gives rise to either YR or YY nucleotides at positions 1 and 2, which translates to either RR or YR dinucleotides at positions 2 and 3.

Analysis of IN mutant protein fractional RY signatures revealed similar preferences for RYXRY at the integration site, with the notable exception of the R231H mutant (Supplementary Figure S2). In this case the frequency of RY at bin positions 0 and 3 was actually lower than the frequencies observed at other positions (bins −5, −1, 4 and 8; Supplementary Figure S2A). One other Arg231 mutant protein, R231Q, also revealed a significant fluctuation from the WT signature, with a novel switch in preference for RY sequences at bin positions −2 and 5 outside of the central RYXRY motif. The N120A mutation, like R231H, significantly reduced the frequency of RY at central bin positions 0 and 3, yet in this case the bin 0 and 3 RY frequency remained greatest across the integration site (Supplementary Figure S2B). IN mutant S119A yielded the largest alterations from the WT, in this case significantly increasing the RY frequency at bins −3 and 6 (Supplementary Figure S2C). The marked preferences for GT at nucleotide positions −3 and −2 and for AC at positions 6 and 7, respectively (Figure 4), account for this unique signature. The other IN mutant proteins did not reveal significant RY frequency differences from WT IN (Supplementary Figure S2).

### Infectivities and DNA analyses of HIV-1 IN mutant viruses

SupT1 cells were infected with normalized amounts of single-round WT and IN mutant viruses that carried and expressed the Luc reporter gene. Two days post-infection, cells were harvested and mutant viral Luc activities were calculated as percentage of WT activity. As the K258S mutation abrogated IN activity under all assay conditions *in vitro*, it was omitted from the virus study. Each tested mutation significantly reduced HIV-1 infectivity, with the extent of the infection defect ranging from ∼25% for the S119A IN mutant virus to >100-fold for the N120E and K258E mutant viruses (Figure 6).

Integration site sequences were determined for the WT virus and five mutant viruses (S119A/T, N120A and

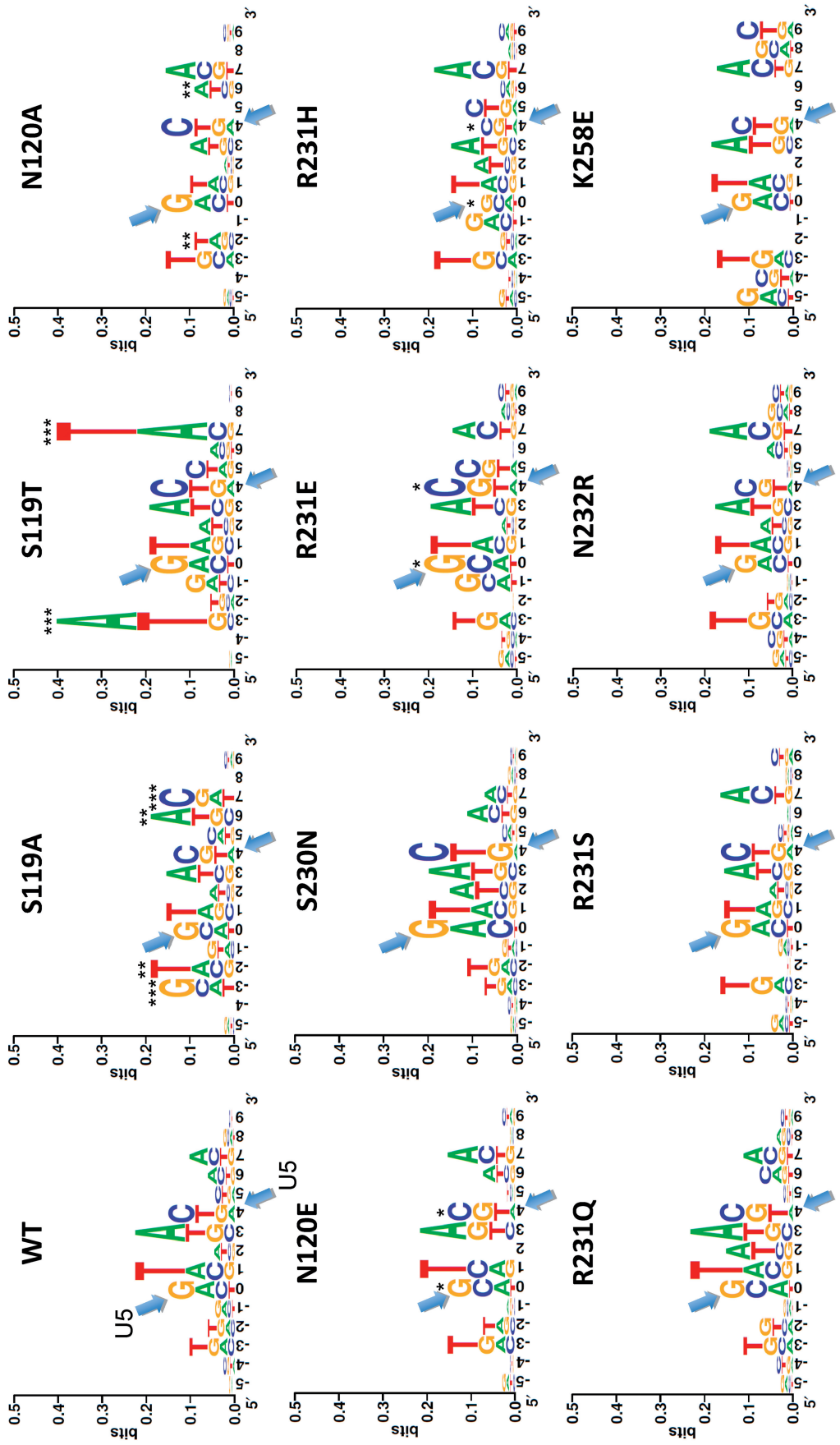


**Figure 4.** Nucleotide preferences at sites of WT and IN mutant enzyme integration *in vitro*. Data is presented using WebLogo (33), wherein the height of each base logo within a given stack is proportional to the frequency of the corresponding nucleotide within the alignment, and the height of each stack of logos reflects the level of conservation at each position. Arrows denote the boundaries of 5-bp duplicated tDNA sequence. Asterisks denote statistically significant differences from the WT signature at the indicated nucleotide position ($*P < 0.05$; $**P < 0.01$; $***P < 10^{-5}$).

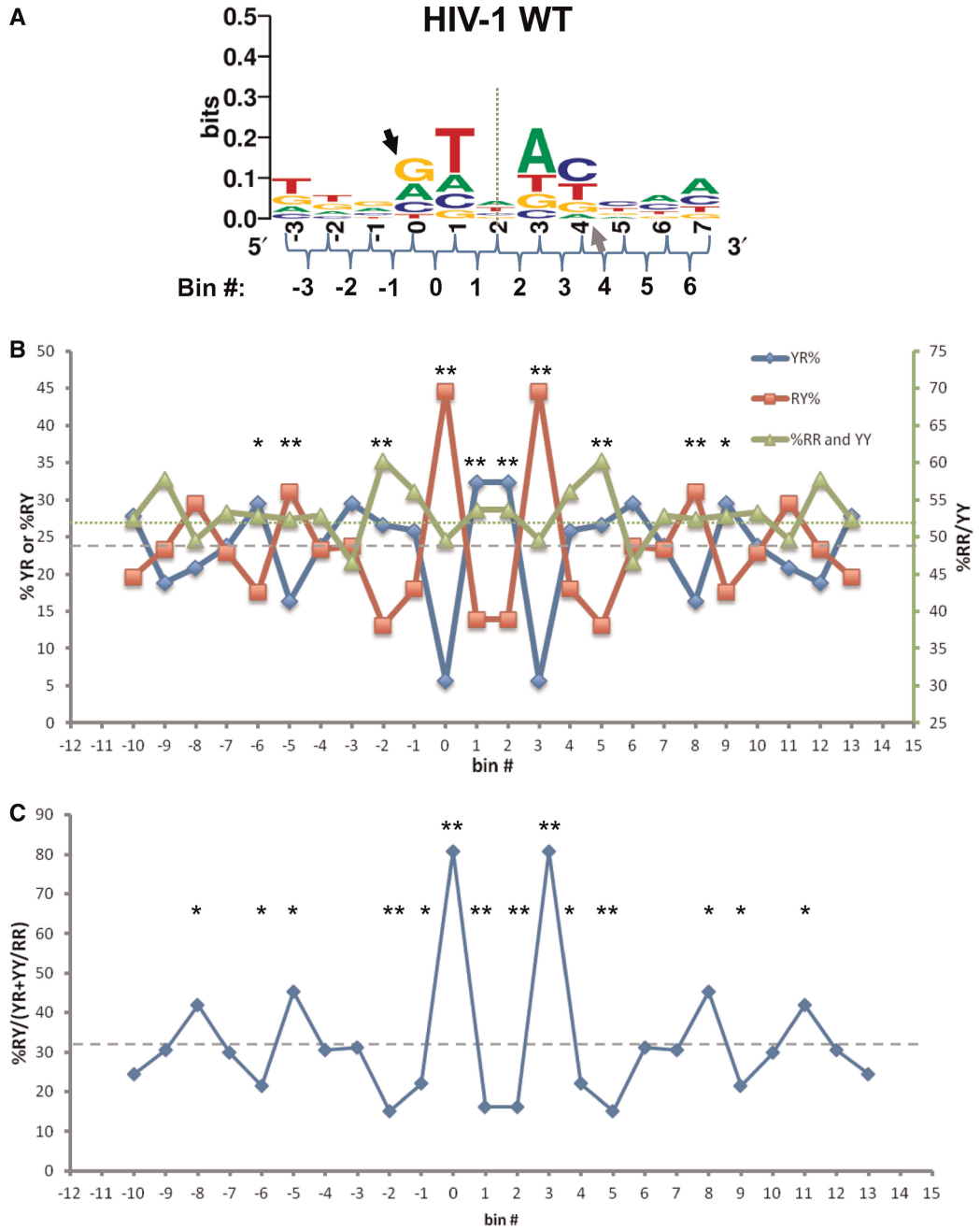


**Figure 5.** Dinucleotide content analysis of WT HIV-1 IN *in vitro* concerted integration sites. (**A**) Dinucleotide bin positions are defined using the WebLogo consensus sequence for WT HIV-1 IN from the upper left panel in Figure 4 as guide. (**B**) Frequencies of RY and YR (left *y*-axis) and RR and YY content (right *y*-axis) at and surrounding the integration sites. (**C**) Graph showing the frequency of rigid versus flexible dinucleotides (% RY/(YR+YY/RR)) for WT HIV-1 IN across the integration site. The dashed lines (B) indicate the expected frequency of RY and YR (23.9%) or RR/YY (52.2%) dinucleotides and (C) the expected RY/(YR+YY/RR) frequency (31.4%). *P*-values (B) reflect the significance of the change in dinucleotide preference of WT HIV-1 IN versus randomly generated sequences and (C) significance of the change in preference of rigid dinucleotides over flexible dinucleotides. $*P \leq 0.05$; $**P \leq 0.005$.

R231H/S) that supported at least 10% of the levels of WT infectivity and integration (Figure 6 and Supplementary Figure S3). Genomic DNA from infected cells was digested with restriction endonucleases and ligated to asymmetric linkers or utilized as a template for *in vitro* bacteriophage Mu transposition as described (17,29). The modified DNAs were then amplified by two rounds of PCR, cloned and sequenced. Duplicated sequences as well as sequences that did not match the processed U3 end of vDNA, cellular genome and linker DNAs at >98% identity were omitted from the analysis. Cellular sequences upstream from the point of U3 vDNA joining were compiled from the draft human genome. In total, 520 unique integration sites were determined for the six viruses. Observed nucleotides at the sites of vDNA joining were compared to those expected based on the sequence of

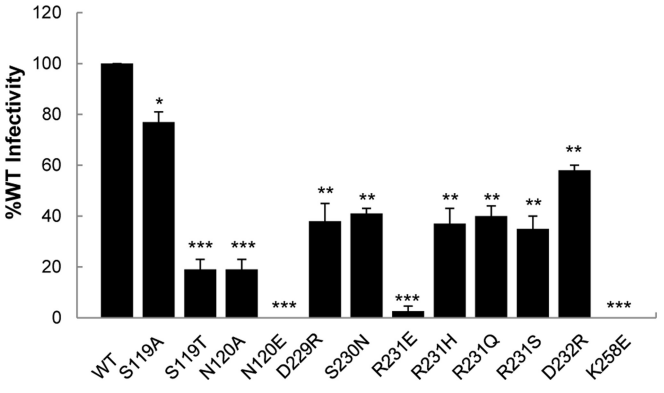


**Figure 6.** Single-cycle HIV-1 infections. The infectivity of each indicated IN mutant virus is normalized to the level of WT virus infection, which was set to 100%. Results are average ± SD for two independent experiments, each performed in duplicate. Asterisks denote statistically relevant differences from the WT ($*P < 0.05$; $**P < 0.01$; $***P < 10^{-4}$). Residue 232, which is a known polymorphic site in IN (40), is Asn and Asp in our HIV-$1_{HXB2}$-based bacterial and HIV-$1_{NL4-3}$-based viral expression vectors, respectively.

human DNA, which was calculated from 10 000 computer-generated coordinates.

Although our results in large part recapitulated the WT preference for the TDG↓*(G/V)TWA(C/B)*CHA consensus sequence, we noted lack of palindromic symmetry among the virus-derived integration sites (Figure 7 and Supplementary Figure S4). Viral integration datasets in the vast majority of cases are derived from cellular sequences that abut one of the two integrated vDNA ends, and the inclusion of events that generated duplications other than 5 bp or deletions during integration will skew palindrome symmetry. Despite this limitation, the mutations that altered the preferences of purified IN enzymes *in vitro* imparted similar novel nucleotide preferences during HIV-1 infection. Accordingly, the S119A mutant virus disfavored adenosine and preferred cytosine at position 7 ($P = 7.9 \times 10^{-4}$), with recognizable alterations in complementary T/G utilization at position −3. Though the viral data failed to recapitulate the statistically significant preference for G/T bases at position −2 that was observed *in vitro*, this trend was nevertheless evident. The S119A IN mutant virus also disfavored guanosine at position 9 and favored adenosine at position 0. The S119T IN mutant virus behaved quite similar to its purified enzyme counterpart, in that similar novel base preferences were selected at positions −3 and 7 (compare Figures 4 and 7). The N120A mutant virus revealed marginal preferences for adenosine and cytosine at positions 0 and 3, respectively. The Arg231 mutant viruses also yielded nucleotide preference patterns similar to those observed with mutant enzymes *in vitro*. Compared to the WT, the R231H mutant preferred adenosine and guanosine at positions 0 and 4, respectively, as well as guanosine at position 7. The R231S IN mutant viral integration preference in large part mirrored the WT pattern, with a marginal shift at position 4 toward unbiased frequencies of A/C utilization ($P = 0.02$; Supplementary Figure S4).

## DISCUSSION

The crystal structure of the PFV TCC revealed unprecedented details on the mechanism of retroviral DNA integration (7). Two inner monomers of an IN tetramer interact with both vDNA and tDNA and donate the active sites required to integrate the vDNA ends. The intasome accommodates tDNA in a severely bent conformation, which is accomplished by the enzyme wrenching down on a preferentially bendable substrate. The unstacking of the two base pairs at the center of the integration site is accordingly facilitated by the marked preference for flexible YR dinucleotides. IN CCD and CTD residues furthermore contact the tDNA at numerous positions outside this central base pair: the amide groups of numerous CCD residues as well as the Arg362 side chain interact with the tDNA backbone, whereas Ala188 and Arg329 make base-specific contacts (7). HIV-1 IN residues Ser119, Arg231 and Lys258 were identified here as analogues of PFV IN residues Ala188, Arg329 and Arg362, respectively (Figure 1). WT and IN mutant enzyme activities and nucleotide site preferences *in vitro* and in virus-infected cells were analyzed to assess amino acid residue roles in tDNA recognition during HIV-1 integration.

### The role of CCD α2 residues in tDNA binding and HIV-1 integration

Retroviral IN amino acid analogs of PFV IN residues Ala188 and Ala189 have previously been implicated in tDNA recognition, primarily through novel banding patterns of DNA-strand-transfer reaction products in sequencing gels (41–44). Our results fine-tune the analysis by narrowing which tDNA bases are likely contacted by HIV-1 IN residue Ser119 during integration.

The methyl group of PFV IN residue Ala188 mediates a van der Waals interaction with the $O_2$ atom of cytosine 6, and PFV accordingly favors cytosine and guanosine at positions 6 and −3, respectively, during integration (7). PFV IN mutant A188S by contrast favored adenosine and thymidine at positions 6 and −3, respectively (Supplementary Figure S5). HIV-1 IN, which harbors Ser at the position analogous to Ala188 in PFV IN, accordingly favors adenosine and thymidine at nucleotide positions 7 and −3, respectively (positions 5–7 in the HIV-1 integration site are analogous to positions 4–6 in the PFV site due to the 5- and 4-bp tDNA cuts made by HIV-1 and PFV IN, respectively). Moreover, HIV-1 IN mutant S119A favored cytosine and guanosine at positions 7 and −3, respectively (Figure 4), in a sense recapitulating the WT PFV IN preferences at these positions (Supplementary Figure S5). Based on these observations, we conjecture that Ser119 in HIV-1 IN and Ala188 in PFV IN interact with tDNA similarly during vDNA integration. Accordingly, the methyl group of the HIV-1 IN mutant S119A side chain might preferentially form a van der Waals interaction with cytosine at position 7 during integration. Consistent with this hypothesis, rhesus macaque simian immunodeficiency virus, which harbors alanine at the analogous position in IN (45), favors cytosine at position 7 (21). Rous sarcoma virus (RSV), an α-retrovirus that produces a 6-bp

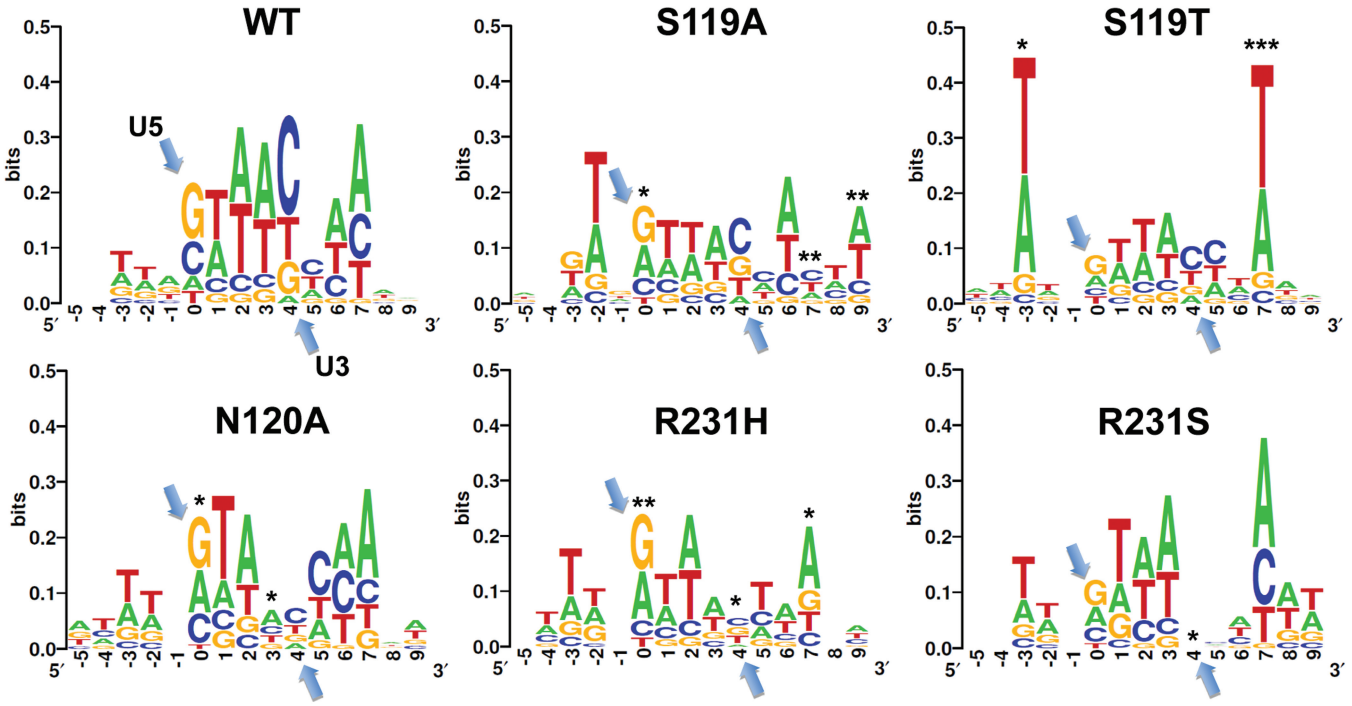


**Figure 7.** Integration site preferences of WT and IN mutant viruses. The legend to Figure 4 explains the heights of the different base logos within a stack and the heights of the different stacks along the *x*-axis. Asterisks denote statistically significant differences from the WT signature at the indicated nucleotide position (*$P < 0.05$; **$P < 0.01$; ***$P < 10^{-5}$).

target-site duplication, harbors serine at the analogous IN position 124 (43) and favors adenosine at position 8 of its consensus integration site (21). Furthermore, the lentivirus equine infectious anemia virus (EIAV), which carries threonine at this position in IN (45), reveals remarkable preference for thymidine and adenosine at position 7 (18,46), virtually identical to the selectivity of our HIV-1 S119T IN mutant (Supplementary Figure S5). It seems unmistakable that this position in IN is dedicated to interacting with the bases that lay three residues of either side of the tDNA cut.

The conservation of a compact amino acid at positions analogous to Ala188 in PFV IN is likely important for tDNA recognition across retroviral INs (7). Accordingly, introducing bulky electronegative aspartate or glutamate residues for Ser124 in RSV IN abrogated mutant enzyme strand-transfer activity (47). It is unclear whether Asn120 in HIV-1 IN might interact directly with tDNA, or whether the relatively modest novel preferences for nucleotides at sites of N120E and N120A IN mutant integration are due to indirect effects on the neighboring Ser119 side chain. A search of the Los Alamos HIV Sequence Database revealed that Asn120 is completely conserved across HIV-1/SIVcpz strains (48). In contrast, proline, which predominates at Ser119 analogous positions across retroviral IN proteins (45), is oftentimes found in HIV-1/SIVcpz IN.

### The HIV-1 IN CTD and tDNA binding

Arg329 in PFV IN hydrogen bonded with tDNA bases guanine 3, guanine −1, and thymine −2 in the TCC crystal structure, while Arg362 interacted with the tDNA backbone. The R329E mutation in PFV IN severely reduced DNA-strand-transfer activity, and the mutant integration sites revealed a significant novel preference for cytosine at position −1 (7). R231E and K258E substitutions in HIV-1 IN also yielded significant reductions in DNA strand transfer activity (Figures 2 and 3). R231E IN displayed novel preferences for A/C and T/G at positions 0 and 4, respectively, whereas K258E IN did not select for significant nucleotide differences from the WT integration sequence (Figure 4 and Supplementary Figure S1). These results are consistent with our hypothesis from structure-based sequence alignment that Arg231 and Lys258 in HIV-1 IN mediate tDNA base and backbone contacts, respectively. While the R231E mutant enzyme retained the WT level of 3′-processing activity, K258E IN displayed ∼3 fold 3′-processing defect (Figure 2). Although the significant further reduction in K258E IN DNA-strand-transfer activity is consistent with a role for Lys258 in tDNA binding, the associated 3′-processing defect suggests that Lys258 might play more than one role in HIV-1 integration. The K258A IN mutation (49) like K258E (Figure 6) reduced HIV-1 single-round infectivity >100-fold.

The isolated HIV-1 IN CTD binds DNA non-specifically (50–53), and exposure of the Arg231 side chain on a saddle-shaped groove in an NMR structure of a CTD dimer originally implicated this residue in tDNA binding (54). While our results are consistent with a role for Arg231 in tDNA binding, the absence of an analogous dimer in the PFV IN-DNA co-crystal structures has since questioned the biological relevance of the isolated CTD multimeric form (2,7).

Although a subset of Arg231 mutants selected for novel integration site preferences *in vitro* and in cells (Figures 4 and 7), the magnitude of these effects are significantly less than the preference of PFV IN mutant R329E for cytosine at position −1 (7). Different possibilities were considered to account for this outcome. First and foremost, the HIV-1 IN CTD may not possess a single residue that is functionally analogous to Arg329 in PFV IN, which imparts significant distortion through contacting multiple nucleotides that surround a central, flexible YR dinucleotide. Our dinucleotide analysis revealed a marked preference for RYXRY by HIV-1 (Figure 5), which enforces flexibility at the two dinucleotide positions that overlap the central base pair. Thus, the inherent asymmetric flexibility instilled through selective YY or YR nucleotides at positions 1 and 2 and concurrent YR or RR nucleotides at positions 2 and 3 may very well necessitate an asymmetric recognition mechanism through more than one IN amino acid residue. It seems possible that a single HIV-1 IN residue may also be unable to span sufficient distance to contact numerous nucleotides that are minimally separated by an additional base pair as compared to PFV. Alternatively, an HIV-1 IN residue that is functionally analogous to Arg329 in PFV IN exists, but it was overlooked in this study. As Arg231 mutants were selectively defective for DNA strand transfer activity (Figure 2) and exhibited altered overall nucleotide and RY dinucleotide preferences during integration (Figures 4, 7 and Supplementary Figure S2), we nevertheless conclude it is likely to play a role in tDNA distortion. Due to the different lengths of HIV-1 and PFV CTD β1-β2 loop regions, residues abutting Arg231 in HIV-1 IN, including Asp229, Ser230 and Asn232, were mutagenized (Figure 1B). The *in vitro* nucleotide site preferences of S230N and N232R mutant INs did not vary from the WT, indicating that neither of these residues is likely to contact tDNA bases during integration. The virtual lack of D229R IN concerted integration activity precluded its site analysis.

A model for the HIV-1 TCC was built by overlying our previous HIV-1 IN-vDNA intasome model with the PFV TCC structure to further investigate the functionality of CTD arginine residues. Arg231 in HIV-1 IN expectedly aligned with PFV IN residue Arg329 in the model (Supplementary Figure 6A). Our intasome model was assembled step-wise from separate HIV-1 IN 2-domain structures, and the Arg231 side chain in the model and the 2-domain CCD–CTD structure on which it was built (55) positions away from the tDNA. Superposition of two available CTD NMR structures (54,56) revealed considerable flexibility among Arg231 side chain positions (Supplementary Figure S6A), indicating that Arg231 is in theory positioned to interact with tDNA during HIV-1 integration. Arg228, the nearest arginine to Arg231 in the primary HIV-1 IN sequence (Figure 1B), in contrast aligned with vDNA (Supplementary Figure 6B) (27).

We note that one HIV-1 intasome model, in particular, yielded a shift in the register of CTD β strands, such that residues $E_{246}$GAVVIQ situated between β1 and β2 (57). It could be informative to assess integration site preferences of IN mutant enzymes containing changes of some of these residues.

## CONCLUSIONS

Although we did not assess the IN–DNA-binding affinities of mutant enzymes in this study, we expect such measures would in the majority of cases be uninformative. The S124D mutation, which abrogated RSV IN strand transfer activity, instilled a relatively mild 2-fold defect in sequence non-specific DNA binding (47). We accordingly suspect that S119A and S119T mutant enzymes, which retained >50% of strand-transfer activity and showed the greatest variation among integration site sequence preferences, would support normal levels of tDNA binding under similar reaction conditions.

Numerous factors likely contribute to the subtle differences observed between *in vitro* and virus-derived integration-site datasets (compare Figures 4 and 7). As mentioned above, direct sequencing of only one of two viral-cellular DNA joints distorted palindromic symmetry, a trend that is largely overcome by analyzing several thousands of integration sites (20,58). Inherent differences between the *in vitro* and live cell tDNA template also likely influenced the outcome. Whereas purified LEDGF/p75 protein binds DNA in a sequence non-specific manner (59), chromatin binding is accomplished through the additional engagement of trimethylated Lys36 on histone H3 (H3K36me3) (60,61), an epigenetic mark that typically associates with actively transcribed genes (reviewed in 62). Biochemical reactions that utilized nucleosomes as the source of tDNA first established the preference of HIV-1 IN for tDNA distortion (63,64). LEDGF/p75 accordingly targets integration to distorted nucleosomal DNA that exists in cells in an inherently different structural conformation than the naked plasmid DNA used in our *in vitro* reactions. DNA remodeling enzymes and members of the RNA polymerase transcription machinery that associate with H3K36me3 chromatin may also contribute to tDNA distortion at sites of integration. Despite these limitations, similarly skewed tDNA nucleotide preferences among the subset of IN mutants that were studied as purified enzymes and viruses (Figures 4 and 7) indicate that IN is the primary determinant responsible for nucleotide selection at sites of vDNA integration.

The work presented here importantly uncovers the mechanistic basis for tDNA distortion during HIV-1 integration. First, we clarify that distortion is spread over two inherently flexible dinucleotides (at positions 1 and 2, and at positions 2 and 3), which contrasts with the distortion of a single, central dinucleotide during PFV integration. The spreading of tDNA distortion over two dinucleotides is likely to put less overall strain on the DNA molecule; in this vein it is not surprising that we did not pinpoint a single amino acid, similar to Arg329 in PFV IN, that contributed significantly to alleviating the penalty of tDNA distortion. Our results moreover clarify that retroviral IN residues analogous to Ala188 in PFV and Ser119 in HIV-1 interact with bases at three positions upstream and downstream from the sites of vDNA joining to help impart the tDNA distortion necessary for concerted vDNA integration.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

We thank Michiyo Mizuuchi for the generous gift of MuA protein and Alex Holman for technical advice on bioinformatic analysis of integration sites.

## FUNDING

United States National Institutes of Health (NIH) [AI039394 and AI070042 to A.E.]; NIH [training grant T32 AI007245 to E.S.]. Medical Research Council UK [G0900116 to P.C.]. Funding for open access charge: Imperial College London.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Li,M., Mizuuchi,M., Burke,T.R. Jr and Craigie,R. (2006) Retroviral DNA integration: reaction pathway and critical intermediates. *EMBO J.*, **25**, 1295–1304.
2. Hare,S., Gupta,S.S., Valkov,E., Engelman,A. and Cherepanov,P. (2010) Retroviral intasome assembly and inhibition of DNA strand transfer. *Nature*, **464**, 232–236.
3. Hare,S., Maertens,G.N. and Cherepanov,P. (2012) 3′-processing and strand transfer catalysed by retroviral integrase in crystallo. *EMBO J.*, **31**, 3020–3028.
4. Pauza,C.D. (1990) Two bases are deleted from the termini of HIV-1 linear DNA during integrative recombination. *Virology*, **179**, 886–889.
5. Sherman,P.A. and Fyfe,J.A. (1990) Human immunodeficiency virus integration protein expressed in *Escherichia coli* possesses selective DNA cleaving activity. *Proc. Natl Acad. Sci. USA*, **87**, 5119–5123.
6. Engelman,A., Mizuuchi,K. and Craigie,R. (1991) HIV-1 DNA integration: mechanism of viral DNA cleavage and DNA strand transfer. *Cell*, **67**, 1211–1221.
7. Maertens,G.N., Hare,S. and Cherepanov,P. (2010) The mechanism of retroviral integration from X-ray structures of its key intermediates. *Nature*, **468**, 326–329.
8. Engelman,A. (1994) Most of the avian genome appears available for retroviral DNA integration. *Bioessays*, **16**, 797–799.
9. Carteau,S., Hoffmann,C. and 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.
10. Bushman,F., Lewinski,M., Ciuffi,A., Barr,S., Leipzig,J., Hannenhalli,S. and Hoffmann,C. (2005) Genome-wide analysis of retroviral DNA integration. *Nat. Rev. Microbiol.*, **3**, 848–858.
11. Schroder,A.R., Shinn,P., Chen,H., Berry,C., Ecker,J.R. and Bushman,F. (2002) HIV-1 integration in the human genome favors active genes and local hotspots. *Cell*, **110**, 521–529.
12. Wu,X., Li,Y., Crise,B. and Burgess,S.M. (2003) Transcription start regions in the human genome are favored targets for MLV integration. *Science*, **300**, 1749–1751.
13. Sharma,A., Larue,R.C., Plumb,M.R., Malani,N., Male,F., Slaughter,A., Kessl,J.J., Shkriabai,N., Coward,E., Aiyer,S.S. *et al.* (2013) BET proteins promote efficient murine leukemia virus integration at transcription start sites. *Proc. Natl Acad. Sci. USA*, **110**, 12036–12041.
14. Gupta,S.S., Maetzig,T., Maertens,G.N., Sharif,A., Rothe,M., Weidner-Glunde,M., Galla,M., Schambach,A., Cherepanov,P. and Schulz,T.F. (2013) Bromo and ET domain (BET) chromatin regulators serve as co-factors for murine leukemia virus integration. *J. Virol.*, **87**, 12721–12736.
15. De Rijck,J., de Kogel,C., Demeulemeester,J., Vets,S., El Ashkar,S., Malani,N., Bushman,F.D., Landuyt,B., Husson,S.J., 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.
16. Ciuffi,A., Llano,M., Poeschla,E., Hoffmann,C., Leipzig,J., Shinn,P., Ecker,J.R. and Bushman,F. (2005) A role for LEDGF/p75 in targeting HIV DNA integration. *Nat. Med.*, **11**, 1287–1289.
17. Shun,M.C., Raghavendra,N.K., Vandegraaff,N., Daigle,J.E., Hughes,S., Kellam,P., Cherepanov,P. and Engelman,A. (2007) LEDGF/p75 functions downstream from preintegration complex formation to effect gene-specific HIV-1 integration. *Genes Dev.*, **21**, 1767–1778.
18. Marshall,H.M., Ronen,K., Berry,C., Llano,M., Sutherland,H., Saenz,D., Bickmore,W., Poeschla,E. and Bushman,F.D. (2007) Role of PSIP1/LEDGF/p75 in lentiviral infectivity and integration targeting. *PLoS One*, **2**, e1340.
19. Stevens,S.W. and Griffith,J.D. (1996) Sequence analysis of the human DNA flanking sites of human immunodeficiency virus type 1 integration. *J. Virol.*, **70**, 6459–6462.
20. Holman,A.G. and Coffin,J.M. (2005) Symmetrical base preferences surrounding HIV-1, avian sarcoma/leukosis virus, and murine leukemia virus integration sites. *Proc. Natl Acad. Sci. USA*, **102**, 6103–6107.
21. Wu,X., Li,Y., Crise,B., Burgess,S.M. and Munroe,D.J. (2005) Weak palindromic consensus sequences are a common feature found at the integration target sites of many retroviruses. *J. Virol.*, **79**, 5211–5214.
22. Li,X., Krishnan,L., Cherepanov,P. and Engelman,A. (2011) Structural biology of retroviral DNA integration. *Virology*, **411**, 194–205.
23. Cherepanov,P. (2007) LEDGF/p75 interacts with divergent lentiviral integrases and modulates their enzymatic activity in vitro. *Nucleic Acids Res.*, **35**, 113–124.
24. Vandegraaff,N., Devroe,E., Turlure,F., Silver,P.A. and Engelman,A. (2006) Biochemical and genetic analyses of integrase-interacting proteins lens epithelium-derived growth factor (LEDGF)/p75 and hepatoma-derived growth factor related protein 2 (HRP2) in preintegration complex function and HIV-1 replication. *Virology*, **346**, 415–426.
25. Koh,Y., Wu,X., Ferris,A.L., Matreyek,K.A., Smith,S.J., Lee,K., KewalRamani,V.N., Hughes,S.H. and Engelman,A. (2013) Differential effects of human immunodeficiency virus type 1 capsid and cellular factors nucleoporin 153 and LEDGF/p75 on the efficiency and specificity of viral DNA integration. *J. Virol.*, **87**, 648–658.
26. Li,X., Koh,Y. and Engelman,A. (2012) Correlation of recombinant integrase activity and functional preintegration complex formation during acute infection by replication-defective integrase mutant human immunodeficiency virus. *J. Virol.*, **86**, 3861–3879.
27. Krishnan,L., Li,X., Naraharisetty,H.L., Hare,S., Cherepanov,P. and Engelman,A. (2010) Structure-based modeling of the functional HIV-1 intasome and its inhibition. *Proc. Natl Acad. Sci. USA*, **107**, 15910–15915.
28. Hare,S., Shun,M.C., Gupta,S.S., Valkov,E., Engelman,A. and Cherepanov,P. (2009) A novel co-crystal structure affords the design of gain-of-function lentiviral integrase mutants in the presence of modified PSIP1/LEDGF/p75. *PLoS Pathog.*, **5**, e1000259.
29. Brady,T., Roth,S.L., Malani,N., Wang,G.P., Berry,C.C., Leboulch,P., Hacein-Bey-Abina,S., Cavazzana-Calvo,M., Papapetrou,E.P., Sadelain,M. *et al.* (2011) A method to sequence and quantify DNA integration for monitoring outcome in gene therapy. *Nucleic Acids Res.*, **39**, e72.
30. Ciuffi,A., Ronen,K., Brady,T., Malani,N., Wang,G., Berry,C.C. and Bushman,F.D. (2009) Methods for integration site distribution analyses in animal cell genomes. *Methods*, **47**, 261–268.
31. Matreyek,K.A. and Engelman,A. (2011) The requirement for nucleoporin NUP153 during human immunodeficiency virus type 1 infection is determined by the viral capsid. *J. Virol.*, **85**, 7818–7827.

32. Cohen,J. (1988) *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. Lawrence Erlbaum Associates, Hillsdale, N.J.
33. Crooks,G.E., Hon,G., Chandonia,J.-M. and Brenner,S.E. (2004) WebLogo: A sequence logo generator. *Genome Res.*, **14**, 1188–1190.
34. Ding,D., Lou,X., Hua,D., Yu,W., Li,L., Wang,J., Gao,F., Zhao,N., Ren,G. and Lin,B. (2012) Recurrent targeted genes of hepatitis B virus in the liver cancer genomes identified by a next-generation sequencing-based approach. *PLoS Genet.*, **8**, e1003065.
35. Bushman,F.D. and Craigie,R. (1991) Activities of human immunodeficiency virus (HIV) integration protein in vitro: specific cleavage and integration of HIV DNA. *Proc. Natl Acad. Sci. USA*, **88**, 1339–1343.
36. Li,M. and Craigie,R. (2005) Processing of viral DNA ends channels the HIV-1 integration reaction to concerted integration. *J. Biol. Chem.*, **280**, 29334–29339.
37. Sinha,S. and Grandgenett,D.P. (2005) Recombinant human immunodeficiency virus type 1 integrase exhibits a capacity for full-site integration in vitro that Is comparable to that of purified preintegration complexes from virus-infected cells. *J. Virol.*, **79**, 8208–8216.
38. Johnson,R.C., Stella,S. and Heiss,J.K. (2008) Bending and compaction of DNA by proteins. In: Rice,P.A. and Correll,C.C. (eds), *Protein–Nucleic Acid Interactions: Structural Biology*. RSC Publishing, London, pp. 176–220.
39. Valkov,E., Gupta,S.S., Hare,S., Helander,A., Roversi,P., McClure,M. and Cherepanov,P. (2009) Functional and structural characterization of the integrase from the prototype foamy virus. *Nucleic Acids Res.*, **37**, 243–255.
40. Low,A., Prada,N., Topper,M., Vaida,F., Castor,D., Mohri,H., Hazuda,D., Muesing,M. and Markowitz,M. (2009) Natural polymorphisms of human immunodeficiency virus type 1 integrase and inherent susceptibilities to a panel of integrase inhibitors. *Antimicrob. Agents Chemother.*, **53**, 4275–4282.
41. van Gent,D.C., Groeneger,A.A. and Plasterk,R.H. (1992) Mutational analysis of the integrase protein of human immunodeficiency virus type 2. *Proc. Natl Acad. Sci. USA*, **89**, 9598–9602.
42. Harper,A.L., Skinner,L.M., Sudol,M. and 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.
43. Harper,A.L., Sudol,M. and 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.
44. Nowak,M.G., Sudol,M., Lee,N.E., Konsavage,W.M.J. and Katzman,M. (2009) Identifying amino acid residues that contribute to the cellular-DNA binding site on retroviral integrase. *Virology*, **389**, 141–148.
45. Engelman,A. and Craigie,R. (1992) Identification of conserved amino acid residues critical for human immunodeficiency virus type 1 integrase function in vitro. *J. Virol.*, **66**, 6361–6369.
46. Hacker,C.V., Vink,C.A., Wardell,T.W., Lee,S., Treasure,P., Kingsman,S.M., Mitrophanous,K.A. and Miskin,J.E. (2006) The integration profile of EIAV-based vectors. *Mol. Ther.*, **14**, 536–545.
47. Konsavage,W.M.J., Sudol,M., Lee,N.E. and Katzman,M. (2007) Retroviral integrases that are improved for processing but impaired for joining. *Virus Res.*, **125**, 198–210.
48. Foley,B., Leitner,T., Apetrei,C., Hahn,B., Mizrachi,I., Mullins,J., Rambaut,A., Wolinsky,S. and Korber,B. (2013) *HIV Sequence Compendium 2013*. Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, NM.
49. Lu,R., Limon,A., Ghory,H.Z. and Engelman,A. (2005) Genetic analyses of DNA-binding mutants in the catalytic core domain of human immunodeficiency virus type 1 integrase. *J. Virol.*, **79**, 2493–2505.
50. Woerner,A.M., Klutch,M., Levin,J.G. and Marcus-Sekura,C.J. (1992) Localization of DNA binding activity of HIV-1 integrase to the C-terminal half of the protein. *AIDS Res. Hum. Retrovir.*, **8**, 297–304.
51. Woerner,A.M. and Marcus-Sekura,C.J. (1993) Characterization of a DNA binding domain in the C-terminus of HIV-1 integrase by deletion mutagenesis. *Nucleic Acids Res.*, **21**, 3507–3511.
52. Vink,C., Groeneger,A.A.M.O. and Plasterk,R.H.A. (1993) Identification of the catalytic and DNA-binding region of the human immunodeficiency virus type I integrase protein. *Nucleic Acids Res.*, **21**, 1419–1425.
53. Engelman,A., Hickman,A.B. and Craigie,R. (1994) The core and carboxyl-terminal domains of the integrase protein of human immunodeficiency virus type 1 each contribute to nonspecific DNA binding. *J. Virol.*, **68**, 5911–5917.
54. Lodi,P.J., Ernst,J.A., Kuszewski,J., Hickman,A.B., Engelman,A., Craigie,R., Clore,G.M. and Gronenborn,A.M. (1995) Solution structure of the DNA binding domain of HIV-1 integrase. *Biochemistry*, **34**, 9826–9833.
55. Chen,J.C.-H., Krucinski,J., Miercke,L.J.W., Finer-Moore,J.S., Tang,A.H., Leavitt,A.D. and Stroud,R.M. (2000) Crystal structure of the HIV-1 integrase catalytic core and C-terminal domains: a model for viral DNA binding. *Proc. Natl Acad. Sci. USA*, **97**, 8233–8238.
56. Eijkelenboom,A.P., Sprangers,R., Hård,K., Puras Lutzke,R.A., Plasterk,R.H., Boelens,R. and Kaptein,R. (1999) Refined solution structure of the C-terminal DNA-binding domain of human immunovirus-1 integrase. *Proteins*, **36**, 556–564.
57. Johnson,B.C., Métifiot,M., Ferris,A., Pommier,Y. and Hughes,S.H. (2013) A homology model of HIV-1 integrase and analysis of mutations designed to test the model. *J. Mol. Biol.*, **425**, 2133–2146.
58. Wang,G.P., Ciuffi,A., Leipzig,J., Berry,C.C. and Bushman,F.D. (2007) HIV integration site selection: Analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Res.*, **17**, 1186–1194.
59. Turlure,F., Maertens,G., Rahman,S., Cherepanov,P. and Engelman,A. (2006) A tripartite DNA-binding element, comprised of the nuclear localization signal and two AT-hook motifs, mediates the association of LEDGF/p75 with chromatin in vivo. *Nucleic Acids Res.*, **34**, 1653–1665.
60. Pradeepa,M.M., Sutherland,H.G., Ule,J., Grimes,G.R. and Bickmore,W.A. (2012) Psip1/Ledgf p52 binds methylated histone H3K36 and splicing factors and contributes to the regulation of alternative splicing. *PLoS Genet.*, **8**, e1002717.
61. Eidahl,J.O., Crowe,B.L., North,J.A., McKee,C.J., Shkriabai,N., Feng,L., Plumb,M., Graham,R.L., Gorelick,R.J., Hess,S. *et al.* (2013) Structural basis for high-affinity binding of LEDGF PWWP to mononucleosomes. *Nucleic Acids Res.*, **41**, 3924–3936.
62. Kouzarides,T. (2007) Chromatin modifications and their function. *Cell*, **128**, 693–705.
63. Pruss,D., Bushman,F.D. and Wolffe,A.P. (1994) Human immunodeficiency virus integrase directs integration to sites of severe DNA distortion within the nucleosome core. *Proc. Natl Acad. Sci. USA*, **91**, 5913–5917.
64. Pruss,D., Reeves,R., Bushman,F.D. and Wolffe,A.P. (1994) The influence of DNA and nucleosome structure on integration events directed by HIV integrase. *J. Biol. Chem.*, **269**, 25031–25041.