Sequence specificity of viral end DNA binding by HIV-1 integrase reveals critical regions for protein–DNA interaction

Dominic Esposito and Robert Craigie¹

Laboratory of Molecular Biology, NIDDK, National Institutes of Health, 5 Center Drive MSC0560, Bethesda, MD 20892, USA

¹Corresponding author
e-mail: bobc@helix.nih.gov

HIV-1 integrase specifically recognizes and cleaves viral end DNA during the initial step of retroviral integration. The protein and DNA determinants of the specificity of viral end DNA binding have not been clearly identified. We have used mutational analysis of the viral end LTR sequence, *in vitro* selection of optimal viral end sequences, and specific photocrosslinking to identify regions of integrase that interact with specific bases in the LTR termini. The results highlight the involvement of the disordered loop of the integrase core domain, specifically residues Q148 and Y143, in binding to the terminal portion of the viral DNA ends. Additionally, we have identified positions upstream in the LTR termini which interact with the C-terminal domain of integrase, providing evidence for the role of that domain in stabilization of viral DNA binding. Finally, we have located a region centered 12 bases from the viral DNA terminus which appears essential for viral end DNA binding in the presence of magnesium, but not in the presence of manganese, suggesting a differential effect of divalent cations on sequence-specific binding. These results help to define important regions of contact between integrase and viral DNA, and assist in the formulation of a molecular model of this vital interaction.

Keywords: HIV-1/integrase/photocrosslinking/protein–DNA interactions/specific DNA binding

Introduction

Integration of a DNA copy of the human immunodeficiency virus (HIV) genome into the host genome is a vital step in HIV replication. After reverse transcription of the viral genomic RNA, two nucleotides are cleaved from each 3' end of the viral DNA adjacent to a conserved CA dinucleotide. Following this 3' processing reaction, the exposed 3' hydroxyls are used in a direct transesterification reaction called strand transfer, in which the 3' viral ends are joined to 5' phosphates in the host DNA to produce the integration intermediate. After integration, the remaining unpaired nucleotides at the 5' ends of the viral DNA are removed, and the single-stranded region of DNA between the integrated virus and the host DNA is repaired; both of these steps are presumably accomplished by host enzymes. For recent reviews on retroviral integration, see Katz and Skalka (1994), and Vink and Plasterk (1993).

Human immunodeficiency virus type 1 (HIV-1) inte-

grase can catalyze both the 3' processing and strand-transfer reactions *in vitro* using oligonucleotide substrates which mimic the viral DNA ends (Sherman and Fyfe, 1990; Bushman and Craigie, 1991; LaFemina *et al.*, 1991). Integrase can also carry out an apparent reversal of strand transfer which has been termed disintegration, in which a branched substrate which mimics the product of strand transfer can be resolved into viral and target DNA segments (Chow *et al.*, 1992). Mutagenesis and proteolysis experiments have shown that integrase consists of three functional domains (Engelman *et al.*, 1993); structural information is now available for all three pieces of the protein. The N-terminal domain contains a conserved HHCC motif which binds zinc with a stoichiometry of 1:1. Binding of zinc to the N-terminal domain enhances the multimerization of the full-length protein and increases the activity of the protein (Burke *et al.*, 1992; Haugan *et al.*, 1995; Zheng *et al.*, 1996). The structure of a dimer of the N-terminal domain was solved by heteronuclear NMR (Cai *et al.*, 1997). Each monomer consists of a helix–turn–helix motif, in which the helix normally used by this class of proteins to recognize specific DNA sequences is instead used to form protein–protein contacts that create the dimer interface. The catalytic core domain contains three acidic residues, D64, D116 and E152 which are essential for catalytic activity (Engelman and Craigie, 1992; Kulkosky *et al.*, 1992). These residues constitute a D,D-35,E motif which is conserved among all retroviral integrases, retrotransposons and some bacterial transposases (Rowland and Dyke, 1990; Polard and Chandler, 1995). The crystal structure of the HIV-1 core domain has been determined by X-ray crystallography, and has been shown to be a member of a superfamily of polynucleotidyl transferases, including Mu transposase, *Escherichia coli* RuvC, and *E.coli* RNase H, which are proposed to use these three catalytic acidic residues to coordinate metal ions during catalysis (Dyda *et al.*, 1994; Yang and Steitz, 1995; Bujacz *et al.*, 1996; Rice *et al.*, 1996). The C-terminal domain, which displays a much lower level of sequence homology among retroviruses, has been shown to bind DNA in a non-specific fashion (Vink *et al.*, 1993; Woerner and Marcus-Sekura, 1993; Engelman *et al.*, 1994). The NMR structure of this domain reveals a dimer with an SH3-like fold, which produces a large cleft that has been proposed to bind DNA (Eijkelenboom *et al.*, 1995; Lodi *et al.*, 1995).

In order to catalyze cleavage and strand-transfer reactions, integrase must first distinguish between substrate and target DNA; that is, between specific viral end DNA, and non-specific host DNA. Stable complexes of integrase and viral DNA ends have been observed in the presence of divalent metal ions, and require the intact integrase protein (Ellison and Brown, 1994). These complexes are stable in the presence of competitor DNA, can be formed

on either the U3 or U5 DNA substrates, and require <15 bp of DNA at the viral termini (Ellison and Brown, 1994; Vink *et al.*, 1994). Numerous studies have been carried out in which viral DNA end sequences were altered, and effects on processing and integration were monitored, but the results of these studies are contradictory. Several investigations have suggested that there is limited sequence specificity required, with only the well conserved CA dinucleotide being needed for processing and strand transfer (LaFemina *et al.*, 1991). Other studies have suggested the importance of 2–7 base pairs upstream of the conserved CA, with varying effects on the reaction (Vink *et al.*, 1991; van den Ent *et al.*, 1994; Yoshinaga and Fujiwara, 1995; Katzman and Sudol, 1996a; Balakrishnan and Jonsson, 1997). *In vivo* studies have not significantly clarified these results, although the majority of them seem to indicate that more than just the conserved CA residue are needed (Murphy *et al.*, 1993; Reicin *et al.*, 1995). There are several possible reasons for the contradictory results. First, many of these experiments failed to examine all possible base substitutions in the LTR region, which could lead to erroneous assumptions about specificity at certain positions. Secondly, all of the *in vitro* studies were carried out with assays using manganese as the divalent metal cofactor. More recently, assay conditions utilizing magnesium as a cofactor have been developed (Engelman and Craigie, 1995). Since magnesium is likely to be the relevant divalent cation used by integrase in the cell, this assay should provide a more realistic way to measure effects on activity. In addition, a number of related enzymes are known to have dramatically altered DNA-binding specificity or activity in the presence of manganese compared with magnesium (Hiom and Gellert, 1997; Junop and Haniford, 1997).

In order to better understand the basis for viral end DNA-binding specificity, we have examined all of the possible viral DNA end mutations for their effect on integrase activity in the presence of either magnesium or manganese. In addition, we have developed a system in which integrase itself can be used to select for sequences that it prefers under different conditions. In a previous study, we used specific photocrosslinking to identify a vital contact between the conserved adenosine residue in the viral DNA end and K159 of integrase (Jenkins *et al.*, 1997). We now extend those studies to examine potential regions of interaction along the viral end DNA. These three methods taken together provide a much clearer picture of how integrase binds the viral end DNA, and should help us to better model the critical interactions of amino acids on the protein with specific regions of the viral DNA.

Results

Processing rates on wildtype substrates

The rate of 3' processing activity of HIV-1 integrase was determined using two separate sets of reaction conditions, and two different metal cofactors. The maximal rate of processing (1.22 ± 0.18 fmol/min) was observed in reactions using the magnesium buffer conditions in the presence of magnesium (Figure 1). With the same buffer, activity was also observed with manganese as cofactor; however, the rate of processing was reduced 2-fold (data

Fig. 1. Comparison of 3' processing rates in Mg or Mn buffers. Time-courses of integrase activity were performed as described in Materials and methods. Aliquots were removed at the indicated times and the amount of product produced was determined by autoradiography. The vertical axis is presented as % product formed, which was calculated as the ratio of radioactivity in the product band to the sum of the radioactivity in the product and substrate bands. The indicated data points are the average of three separate determinations; error bars indicate the standard deviation. Calculated reaction rates were determined by a least squares fit of the first seven timepoints (out to 60 min); in both cases, the r^2 value for the fit is >0.98 .

not shown). Processing in the manganese buffer conditions with manganese as cofactor (0.84 ± 0.11 fmol/min) was ~70% of that seen in magnesium buffer conditions. This same ratio of processing activity between the magnesium and manganese supporting buffers was seen at several different concentrations of integrase, and both reactions showed a similar optimal rate of processing near 1 μ M integrase. As expected, no activity was detectable in either buffer in the absence of any metal ion. The extent of processing in both the magnesium and manganese assays was linear over a period of 60 min, and in both cases, all of the substrate could be completely converted to cleaved product.

Mutational analysis of Long Terminal Repeat (LTR) sequence specificity

In order to identify elements of sequence specificity for the 3' processing reaction, mutations were introduced into the U5 LTR sequence. All possible single mutations were made in the terminal 16 bases of the LTR (Figure 2A); in all cases, double-stranded substrates contained the proper complementary base pairs at the site of mutation. For each of the 64 mutant substrates, integrase reactions were carried out and initial rates of 3' processing were determined. Reactions were carried out at least three times per substrate, and an average rate was calculated. This rate was compared with the rate of reaction of the wildtype U5 LTR substrate to determine a percentage of wildtype value for each mutant. This process was carried out in both the magnesium and manganese assays. The results of this analysis are presented in Figure 2B, C, D and E.

Changes in sequence at several positions showed no effect on 3' processing. These included positions 12, 13 and 14, for which the activity observed in the mutants was similar to that of the wildtype U5 substrate (data not shown). Additionally, the mutants at positions 5, 6 and 7 were cleaved at rates between 90 and 110% of the wildtype levels, suggesting that these sequences had no significant

effect on the processing reaction (Figure 2C and D). These results were observed in both magnesium and manganese assays. Figure 2B shows that mutations at positions 10 and 11 had no significant effect in the manganese assay, but dramatically affected 3' processing in the magnesium assay. In particular, substitutions of guanine at position 11 reduced rates of processing to <10% of the wildtype

substrate, while all three substitutions at position 10 reduced levels of processing by at least 2-fold. Position 9 showed an intermediate effect, in which neither assay was dramatically affected by changes to adenine or guanine, whereas a mutation to cytosine reduced processing levels 2-fold in the manganese assay, and nearly 20-fold in the magnesium assay. Position 8 (Figure 2C) showed a similar effect, with one mutation (cytosine to guanine) producing a significant change in processing rate.

Nucleotides closer to the 3' end of the substrate were also adversely affected by mutations. Specific mutations in positions 3 and 4 (guanine and thymine, respectively) reduced activity by 5- to 15-fold (Figure 2D). Here, a differential effect of magnesium and manganese on the changes in activity was not observed. Position 2 (Figure 2E), a cytosine which is conserved throughout all retroviral LTR sequences, produced the anticipated results; all three mutants were poor substrates for processing, with almost no activity observed in the magnesium assay, and <10% of wildtype activity in manganese. The conserved adenine at position 1 was completely essential under both assay conditions; none of the mutant substrates was able to be processed (data not shown). The -1 and -2 positions also had strong specificity requirements (Figure 2E). Position -1 favored the presence of a guanine or thymine; cytosine and adenine were tolerated in the manganese assay, but rates of processing were reduced 10-fold under magnesium conditions. Position -2 was only marginally affected by any mutations in the manganese assay, but under magnesium conditions, the cytosine mutant was strongly defective for processing.

In vitro selection of sequence specificity

An alternate means to examine substrate specificity is to allow integrase to 'choose' its own preferred substrates. To do this, we carried out a modified ligation mediated PCR (LM-PCR) reaction as described in Materials and methods, and outlined in Figure 3. Four sets of degenerate oligonucleotides were used to analyze the processing specificity. The four substrates each had three degenerate bases; substrate 1 at positions 9-11, substrate 2 at positions 6-8, substrate 3 at positions 3-5, and substrate 4 at positions 1 and 2. After creation of the double-stranded substrates by addition of DNA polymerase, integrase was added to allow 3' processing to occur. Processed substrates

Fig. 2. Relative 3' processing rates of mutated HIV-1 LTR sequences. (A) Sequence of the top strand of the wildtype HIV U5 and U3 LTRs, with the numbering of the positions of the terminal 15 bases. The position of the 3' processing cleavage is shown by the arrow. (B-E) Effects of mutations of the various positions of the LTR on 3' processing rates. The number and letters along the horizontal axis signify the position and type of mutation. The vertical axis expresses the initial rate of processing of the mutant as a percentage of the wildtype initial rate (using the U5 LTR as the wildtype substrate). Initial rates were generally calculated using timepoints out to 60 min; in the few cases where r^2 values for the 60 min points were <0.95, only timepoints out to 45 min were used. One hundred percent signifies that the mutated substrate is processed with the same rate as the wildtype U5 substrate. Each bar corresponds to the average of three separate rate determinations on a given mutant LTR sequence. Black bars represent reactions carried out under Mg conditions, while gray bars indicate reactions carried out under Mn conditions. Percentage of wildtype rates were based on wildtype reactions using the same metal ion as the mutated substrates. Error bars show the standard deviation of the three measurements.

Fig. 3. Schematic diagram of the *in vitro* LTR selection scheme. Only portions of the actual DNA sequences are shown for the sake of clarity. In step 1, a short oligonucleotide is annealed to a degenerate U5 LTR bottom strand. This duplex is extended with DNA polymerase in step 2 to form the substrate, which is reacted with integrase in step 3 in the presence of either Mg or Mn to form the cleaved product with protruding 3' hydroxyl. In step 4, a linker with a 2 bp 5' overhang is ligated to the product; the gray triangle emphasizes that there is no covalent linkage between the 3' hydroxyl of the adenosine and the 5' hydroxyl of the guanosine on the linker top strand. In step 5, PCR primers are added to the reaction, and LM-PCR is carried out to produce the final product.

were then captured by ligating a duplex linker with a 2 bp 5' overhang complementary to the 5'-CA overhang of the cleaved processing product. LM-PCR then amplified the ligated fragments, ensuring that only processing products were analyzed. After subcloning, these products were sequenced to determine which sequences were utilized by integrase. As a control, a template containing a sequence which was not cleavable by integrase was used; no LM-PCR products were detected with this template.

Between 9 and 16 individual clones were analyzed for each substrate, and the same selection procedure was carried out using both the magnesium and manganese assay conditions; a summary of the sequences found at each position is shown in Table I (Mg) or Table II (Mn). Figure 4 graphically compares the results of the selection experiment with the results of the analysis of individual LTR mutations. Regions of the LTR which have effects on processing rates are marked with circles. The intensity of the shading of the circles shows the relative significance of the effect, with open circles being a minor effect (or effects from only one nucleotide mutation), shaded circles signifying regions that are strongly reduced for processing when mutated, and black circles for bases which appear critical for processing. Because of the design of the LM-PCR experiment, the two most 3' positions (-1 and -2) could not be tested in the *in vitro* selection system, as these bases must remain the same for the ligation linker to anneal properly.

Table I. Results of *in vitro* sequence selection under magnesium conditions

	Position										
	11	10	9	8	7	6	5	4	3	2	1
Base frequency											
A	4	10	4	2	3	2	2	9	0	0	9
C	4	2	1	7	6	5	6	2	3	11	0
G	0	2	5	0	4	3	1	0	11	2	0
T	8	2	6	6	2	5	4	3	0	0	0
Consensus	H	A	D	Y	N	N	C	A	G	C	A

Sequence selection was carried out as described in Materials and methods. Four separate substrates were used spanning the region from position 1–11. The frequency of occurrence of each base at a given position is shown. The consensus sequence was determined by taking any single base which was present in >50% of sequences. If this requirement could not be fulfilled, a multiple-base consensus was used if the multiple bases were present in at least 85% of sequences. Otherwise, an 'N' was used to indicate a lack of clear base preference. Standard IUPAC codes were used for degenerate consensus sequences: S=G/C, W=A/T, Y=C/T, H=A/C/T, D=A/G/T.

Table II. Results of *in vitro* sequence selection under manganese conditions

	Position										
	11	10	9	8	7	6	5	4	3	2	1
Base frequency											
A	4	4	3	3	2	3	2	6	1	0	9
C	3	2	2	3	3	3	4	1	3	7	0
G	3	3	2	1	2	2	2	1	7	2	0
T	2	3	5	2	2	1	3	3	0	0	0
Consensus	N	N	N	N	N	N	N	W	G	C	A

Sequence selection was carried out as described in Materials and methods. Four separate substrates were used spanning the region from position 1–11. The frequency of occurrence of each base at a given position is shown. The consensus sequence was determined by taking any single base which was present in >50% of sequences. If this requirement could not be fulfilled, a multiple-base consensus was used if the multiple bases were present in at least 85% of sequences. Otherwise, an 'N' was used to indicate a lack of clear base preference. Standard IUPAC codes were used for degenerate consensus sequences: S=G/C, W=A/T, Y=C/T, H=A/C/T, D=A/G/T.

Specific photocrosslinking of integrase to LTR substrates

To analyze further the regions of both protein and DNA required for viral DNA binding by integrase, we placed specific photocrosslinking agents at various positions along the DNA substrates. Figure 5A shows the location of the 5-iododeoxyuracil (IdU) or 5-iododeoxycytosine (IdC) residues (boxed regions). Each substrate had only one unique position replaced by the photocrosslinker. Figure 5B shows a representative autoradiogram of a gel separating the products of a 30 min crosslinking reaction. In several cases, distinct bands corresponding to the molecular weight expected for a protein–DNA complex were seen. A control reaction, containing DNA without the photocrosslinker, produced no shifted material. The level of crosslinking (defined as the percentage of radioactivity in the product band relative to the total radioactivity in the lane) ranged from <1% [for C(2)] up to nearly 10% [for A(-2)]. The amount of crosslinking at position A(1), which has been previously described and is not shown in

Fig. 4. Comparison of the regions of the viral LTR important for 3' processing under Mg or Mn conditions. The sequence of the wildtype HIV U5 LTR is shown at the top. Circles represent positions which when mutated produce an effect on 3' processing rates. Open circles represent the weakest effect, including small changes (<40%) in activity or effects seen only with specific mutations. Shaded circles represent a moderate effect; mutations in these residues produce dramatic reductions in processing rates and/or reductions are seen with multiple changes at that position. Black circles represent the strongest effects; mutations in these residues nearly eliminate processing activity. The top two lines represent results using the Mg conditions, while the bottom two lines represent the same results in the Mn conditions. The top line in each group presents a summary of the data from the complete set of LTR mutants (Figure 2), while the bottom line summarizes the results seen in the *in vitro* selection assay (Figure 3). Positions -1 and -2 could not be examined in the *in vitro* selection assay.

this figure, was approximately 10% (Jenkins *et al.*, 1997). These crosslinking reactions were carried out under magnesium conditions; under manganese conditions, specific crosslinking was only observed with three substrates: A(-2), C(-1) and A(1), and in all three cases the level of crosslinking was <1% (data not shown).

In order to identify the specific regions of integrase that were interacting to form the crosslinks, integrase-DNA complexes were isolated and proteolyzed, and labeled crosslinked peptides were N-terminally sequenced. Figure 5C identifies the regions of the DNA which were sensitive to crosslinking, and the locations of four separate specific crosslinks. In two cases, we were able to specifically determine the residue(s) which interacted with the crosslinked position. The crosslink to K159 at position A(1) has previously been described (Jenkins *et al.*, 1997). The strong crosslink observed at position A(-2) was mapped to a peptide containing amino acids 139-152. N-terminal sequencing of this peptide produced the data shown in Figure 5D. There is a significant absence of the expected PTH derivative at the position corresponding to tyrosine 143. In addition, there is a decrease in the expected amount of the derivative in the cycle corresponding to Q148. Using the repetitive yield to calculate expected amounts of PTH derivatives, we estimate that 75% of the crosslinking at this position is to Y143, and the other 25% to Q148.

The crosslink to C(-1) could not be isolated to a specific amino acid. The crosslinked peptide began at H51, and based on the cleavage reagents and fragment size, likely ended after position 64. Sequencing produced data up to I60, but no sequence was observed C-terminal of that position. This could be due to crosslinking which somehow disrupted the amino acid chain, but is more likely due to premature elution from the sequencing column. Other data suggest that peptides of less than 4 or 5 amino acids often elute from the column during washing. Therefore, we could not isolate the specific amino acid, but crosslinking was likely to be to W61, Q62 or L63. The minimal peptide that was involved in the T(5) crosslink was identified as 247-270. However, the lower amounts of crosslinked material prohibited assignment of a specific region of crosslinking, as clear Edman data were only seen through position 254. In the other three significant crosslinks, the

level of crosslinking was either too low to permit accurate determination of the location, or crosslinking occurred to multiple regions of the protein (data not shown). Figure 6 graphically compares the results of the three magnesium-based experiments. As in Figure 4, shaded circles are used to represent the relative amount of specificity requirements on a given position of the LTR. Levels of crosslinking were used as an indication of the relative importance of interactions at given positions in the photocrosslinking study.

Discussion

Differential cation effects on specificity

HIV-1 integrase 3' processing and strand transfer occur *in vitro* in the presence of either magnesium or manganese. Early reports suggested that activity was more robust in manganese (Sherman and Fyfe, 1990; van Gent *et al.*, 1992); however, under optimal conditions, we observe activity in magnesium which is slightly higher than that seen in manganese. Several lines of evidence suggest that DNA binding, and probably sequence specificity, are differentially affected by the type of metal ion used. It has been shown that HIV-1 integrase has a non-specific nuclease activity when in the presence of manganese; similar activity is not observed in magnesium (Engelman and Craigie, 1995). In addition, in the presence of manganese, integrase is capable of using alternate nucleophiles to carry out the 3' processing reaction, including alcohols such as glycerol or methanol (Katzman and Sudol, 1996b). In magnesium, water is the predominant nucleophile used by the protein. Specific photocrosslinking is also dramatically reduced in the presence of manganese as compared with magnesium (D.Esposito and R.Craigie, unpublished results). These facts suggest that the active site of the enzyme may be considerably less 'tight' in the presence of manganese, allowing a wider selection of nucleophiles, and potentially reducing the DNA-binding specificity of the 3' processing reaction. This would also help to explain why the disintegration activity observed *in vitro*, which involves the apparent reversal of the strand-transfer reaction, occurs only in the presence of manganese and appears to require no sequence specificity beyond the CA dinucleotide (Chow *et al.*, 1992; Balakrishnan and

Fig. 5. Results of photocrosslinking studies on the HIV U5 LTR. (A) The sequence of the wildtype U5 LTR. Boxes indicate residues which have been replaced uniquely with photocrosslinking reagents in the various substrates. Thymine and adenine residues were replaced with 5-iododeoxyuracil (IdU), while cytosines were replaced with 5-iododeoxycytosine (IdC). (B) The autoradiogram of a 4–12% NuPage MES gel of crosslinked samples. Lanes are identified by the position of the crosslinker in the substrate DNA; for example, T(7) represents the replacement of the thymine at position 7 with an IdU residue. ‘Con’ is a control DNA containing no photocrosslinkable residues. ‘S’ represents the position of the substrate DNA, and ‘X’ represents the position of the protein–DNA crosslinked complex. (C) Results of the crosslinking experiments. Residues with detectable amounts of crosslinking are boxed. The four strongest crosslinking signals are marked with filled circles. The peptides or amino acids which were crosslinked at these positions, as determined by N-terminal sequencing, are indicated. The A(1) crosslink to K159 was described in a previous work, and is not represented in the figure in (B). (D) Plot of the per cycle yield of PTH-amino acid derivatives from N-terminal sequencing of the A(-2) crosslinked peptide. Crosslinked residues are represented by black bars. The diagonal line represents the best-fit linear regression of the cycles excluding the two crosslinked residues. Serine residues naturally fall below the regression line due to hydrolytic loss of derivatives.

Jonsson, 1997). Many other DNA-binding proteins involved in polynucleotide transfer show dramatic effects when manganese is substituted for magnesium, ranging from alterations in DNA-sequence specificity for the eukaryotic RAG proteins (van Gent *et al.*, 1996; Hiom and Gellert, 1997) to effects on catalytic activity or reaction kinetics for the MuA and Tn10 transposases (Savilahti

Fig. 6. Summary of the critical regions of the viral LTR for 3' processing under Mg conditions. The sequence of the wildtype HIV U5 LTR is displayed at the top. Circles represent positions which when mutated produce an effect on 3' processing rates as indicated in the legend to Figure 4. The three sets of data represent results of the mutated LTR study (Mutation), the *in vitro* selection (Selection) and the photocrosslinking (Crosslink). Where known, the regions of the protein which interact with the different parts of the LTR are indicated.

et al., 1995; Junop and Haniford, 1997). One report has also shown a change in sequence specificity of RSV integrase in the presence of magnesium (Katzman *et al.*, 1989). These arguments suggest that many of the effects of sequence changes seen in HIV-1 integrase under the commonly studied manganese conditions may not mimic the effects that are important in the presence of magnesium, likely to be the essential cation *in vivo*. In order to examine this possibility, we constructed LTR substrates which had mutations in every possible position 14 bases in from the 3' end, and analyzed their ability to be processed by integrase under magnesium or manganese conditions. The results clearly show a differential effect of the cation on specific locations along the LTR, and also very strong effects of particular nucleotide replacements on activity. As added confirmation, we used an *in vitro* selection procedure to reproduce very similar results under both cation conditions.

The distal LTR sequence (positions 8–14)

Perhaps the most interesting result of the mutational studies was the apparent importance of residues very distal to the LTR end, at positions 8–11. In the past, it has been reported that regions >6–8 bases from the end of the viral DNA were not required for integration (LaFemina *et al.*, 1991; Vink *et al.*, 1991; van den Ent *et al.*, 1994; Balakrishnan and Jonsson, 1997); these reports using direct assays and methylation interference were all done under conditions in which manganese was the divalent cation. Under these conditions, we saw little effect of these mutations on activity. Only in two specific cases, T9C and C8G, was there any significant effect in the presence of manganese, and even with these mutants the reduction in activity was no more than 2- to 3-fold. In contrast, with magnesium, activity was dramatically affected by mutations at positions 9, 10 and 11 (Figure 2B). It is important to note that the U3 sequence differs from the U5 sequence at positions 9 and 11. In both cases, the nucleotide present in the U3 sequence (T at position 11, G at position 9) shows the least effect on activity in these experiments, as would be expected. The data obtained from the *in vitro* selection experiment (Tables I and II) corroborate the mutational data. Under manganese conditions, no sequence specificity was observed at any position upstream of position 4, while in the magnesium

conditions, very clear selectivity was seen from positions 8–11, with similar base preferences as predicted by the mutational study. The results of the selection experiment also help to suggest that the effects of mutations at various positions appear to be mainly isolated to those positions. Since we are mutating single bases, there is always the possibility of synergistic effects of mutations at different positions. The fact that the selection experiment produces results very similar to those seen with the single mutations suggests that there is little synergy within the LTR. Mutations at positions 12–14 (data not shown) failed to have any effect on the *in vitro* 3' processing activity in the presence of either cation.

The proximal LTR sequence (positions 3–7)

Mutations at positions 6 and 7 had no effect on 3' processing activity (data not shown) and these positions showed no specificity in the *in vitro* selection in the presence of either cation. Position 5 mutations also had no effect on 3' processing activity, and this position did not appear to favor any particular base in the *in vitro* selection under manganese conditions. However, there did appear to be a strong preference in the presence of magnesium for the use of cytosine or thymine at this position. Thymine is the wildtype base in both U5 and U3 LTRs, but the activity of the T5C mutant in the mutational study is actually slightly higher than the wild-type rate, and was in fact the most active of all the mutant substrates. Mutations at positions 3 and 4 did have deleterious effects in both the magnesium and manganese assays, and in most cases there was little differential effect between the two cations. The selection experiments confirmed that A was the most commonly selected base at position 4 with both cations, and G was not observed at this position. The apparent lower activity of U3 substrates *in vitro* may in part be explained by the presence of a thymine at position 4, which has a small but significant effect on activity in the mutational data, and is not nearly as favored in the selection. Likewise, G and C were most commonly seen at position 3, and T was not selected. These results are qualitatively similar to published results using manganese.

The terminal LTR sequence (positions –2 to 2)

The 'invariance' of the CA dinucleotide has been well established in previous studies (Vink *et al.*, 1991; van den Ent *et al.*, 1994; Yoshinaga and Fujiwara, 1995; Balakrishnan and Jonsson, 1997), and the results of this study confirm those data. The effects of mutations on position 2 are dramatic. In the presence of magnesium, only a C2G mutation shows any activity, and that is down >20-fold from wildtype levels. In the presence of manganese, slight activity can be detected with any of the three mutations, but processing is always significantly impaired. Position 1 mutations were not tolerated under any cation conditions; no mutation of position 1 produced any detectable activity, demonstrating the complete requirement for an adenosine at this position. It is interesting to note that replacement of the A–T base pair at position 1 with a mismatched base pair of T–T does permit activity; in magnesium the level is ~20% of wildtype, while in manganese it is slightly higher. Previous results have demonstrated the likelihood that this base (as

well as those at positions –1 and –2) is unpaired during the reaction (Scottoline *et al.*, 1997), suggesting that the increase in activity with the mismatched bases may be a result of easier separation of the two strands.

The two terminal positions demonstrated interesting specificity requirements, with particular substitutions having very deleterious effects. However, it was impossible to confirm these results by means of the *in vitro* selection procedure, since the sequence of these bases was required to be fixed for proper ligation during LM-PCR. These positions are particularly interesting since they vary greatly among the different retroviruses, and may strongly impact on the sequence specificity of the reaction.

Specific photocrosslinking reveals critical contacts

Specific photocrosslinking has been used to reveal important protein–DNA contacts in several systems, including the interaction of HIV-1 integrase with DNA (Willis *et al.*, 1993, 1994; Jenkins *et al.*, 1997). We have chosen 5-iododeoxyuracil and 5-iododeoxycytidine for a number of reasons, including their longer wavelength of crosslinking, which limits photodamage to the protein, and their chemical and steric similarity to natural DNA bases. Most importantly, these reagents are 'zero-distance' crosslinkers, which must be nearly within van der Waals contact to their target in order to form stable crosslinks (Deitz *et al.*, 1987). Therefore, they avoid the complications which arise with linker-arm crosslinking agents which can crosslink 5–10 Å away from the DNA, and should reveal more detailed regions of interaction. Placement of the crosslinkers at various positions along the U5 LTR revealed numerous regions of crosslinking, with varying amounts of crosslinked product being formed. As expected from the results of the mutational studies, many of the strongest crosslinks were observed in regions which were highly sensitive to mutation, including the terminal LTR region. The strongest crosslinks were seen to the A at position 1, but significant levels of crosslinking to the A at the –2 position and the C at –1 were also observed. A region in the proximal LTR at positions 4 and 5 also produced high levels of crosslinking. Finally, the two iodouracil residues placed at positions 10 and 11 showed lower, but clearly detectable levels of crosslinking, suggesting that this region is also involved in contacts with the DNA, as was predicted from the magnesium mutational data. Conspicuously absent were crosslinks to the C(2) position, which might be expected due to the importance of that position to specificity. There are several possible reasons for the lack of crosslinking. First, the residue(s) involved in this interaction may be poorly crosslinkable to iodocytosine. Alternatively, the interaction at this region may involve the positioning of water or metal ions, which may place the actual amino acid outside the distance of the crosslinking agent. Finally, the interaction may take place to a region away from the 5-position of the cytosine, and again, may be outside the optimal crosslinking distance. It is possible that other crosslinking agents may be more successful at identifying interactions at this position. Interestingly, little crosslinking was observed in the presence of manganese at any position other than positions 1, –1 and –2, and these positions produced crosslinking at levels 3- to 5-fold lower than in magnesium. This suggests that in the presence of manganese, the DNA is

in fact less tightly bound to the protein, or that the complex of integrase with DNA is more transient, and thus less likely to crosslink.

Samples which produced significant levels of crosslinked product were analyzed by proteolysis and N-terminal sequencing to identify regions of contact. The results help to clarify some of the mutational data, and identify critical protein–DNA interactions. The terminal bases all seem to be involved in contacts with the catalytic core of the protein, while the only analyzed proximal crosslink is clearly contacting the C-terminal portion of the protein. These data are consistent with crosslinking results seen by Heuer and Brown using disintegration-type substrates involving target and viral DNA (Heuer and Brown, 1997, 1998). They found several peptides involved in crosslinks to the C(2) position, and the 5' end of the bottom strand [adjacent to the A(–2) position]; these peptides were localized to the catalytic core region. In contrast, a crosslink to the T5 position of a disintegration substrate contacted peptides in the C-terminal, in the same 247–270 region as the T5 crosslink described here.

A model for critical interactions of integrase and viral DNA

These data taken together provide clues to the important protein–DNA contacts required for integration. As would be expected, the catalytic core appears to bind the terminal LTR regions, at which the 3' processing and strand-transfer reactions occur. These regions also exhibit the highest level of sensitivity to mutation, and show the strongest protein–DNA crosslinking. The very strong crosslinks at the two terminal bases, coupled with previous data suggesting that these base pairs are likely to be disrupted upon binding suggests that these interactions may be critical for the proper alignment of the DNA in the active site. Gerton *et al.* (1998) recently suggested that Q148 might be involved in interactions with the 5 dinucleotide of the viral DNA. Our crosslinking data seem to confirm that this residue is capable of making close contact to the A(–2) position of the viral DNA. However, crosslinking at this position also occurred to Y143. Neither of these residues are visible in current crystal structures of retroviral integrases; they lie on the disordered loop just C-terminal to the active site glutamate. This loop has been proposed to be involved in a conformational change that may occur upon DNA binding (Dyda *et al.*, 1994). It is likely that once DNA binds to the active site, residues on this loop stabilize the 5' end of the viral DNA. The fact that specific crosslinking to these two residues is seen also argues that the previously disordered loop probably becomes ordered upon DNA binding. It is not clear what the relative importance of the two crosslinked residues are. One caveat of this class of crosslinkers is that they have two separate electronic pathways for crosslinking. Crosslinks to aromatic amino acids occur by a higher energy triplet state, and produce much higher levels of crosslinking (Deitz *et al.*, 1987). Therefore, it is possible that the Q148 interaction is as strong or stronger than that produced by Y143, but that it may produce lower signal intensity due to the preferred aromatic crosslinking pathway. One argument in favor of this is that a mutant in Y143 appeared to have little effect on 3' processing or strand transfer (van den Ent *et al.*, 1998), whereas a

mutant in Q148 lowered levels of processing and strand transfer, and affected the choice of nucleophiles used by integrase (Gerton *et al.*, 1998). However, more thorough mutagenesis is needed to rule out the involvement of either of these amino acids.

Crosslinking of the C(–1) position was isolated to a peptide spanning the region from 51–64. PTH derivative data were obtained for residues 51–60, but thereafter, signal was lost dramatically. This is most likely to be due to loss of the small peptide from the sequencing column matrix. Thus, the point of crosslinking is likely to be in one of the last few residues. W61 and Q62 are strongly conserved among retroviral integrases, and the positioning of Q62 in the active site region suggests that it may be involved in this interaction. Crosslinking to the aromatic W61 residue might also explain the rather strong signal observed at this position. A recent more detailed crystal structure of the core domain (Goldgur *et al.*, 1998) provides a starting point for analyzing the likelihood of these predicted contacts. As shown in Figure 7, all of the residues potentially involved in the crosslinking are in the active site region. In addition, the spacing of these residues is consistent with the length of DNA involved; if the A(1) position binds in the vicinity of K159 and E152, then the C(–1) position interacting with Q62 and the A(–2) position with Q148 is well within the allowed spacing. The disordered loop has been traced in this structure, and Y143 is shown facing inward into the middle of the loop; however, the temperature factors for these residues are quite high, implying significant flexibility, so determining the exact positioning of Y143 and Q148 is difficult.

Further upstream, in the proximal LTR region, there are clear interactions with the C-terminal domain of integrase. Originally, it was believed that this region, which has a non-specific DNA-binding activity, was involved in target DNA binding (van Gent *et al.*, 1991; Engelman *et al.*, 1993). Recent results appear to rule this out; target specificity of chimeric integrases clearly aligns with the catalytic core region and not the C-terminus (Shibagaki and Chow, 1997; Katzman and Sudol, 1998). These data, coupled with data from Heuer and Brown (1997), suggest that the C-terminal domain is mainly involved in stabilizing the interaction of integrase with the viral DNA. This would explain the requirement for the C-terminal domain for 3' processing, but not for disintegration, which does not require any proximal LTR sequences (Balakrishnan and Jonsson, 1997). Stabilization of the proximal LTR sequence may be required in order to 'fit' DNA into the proper conformation to allow tighter binding of the terminal sequences; this could be accomplished by random interactions with DNA until the C-terminal domain 'catches' a proper sequence in this region, whereupon the rest of the viral DNA is more tightly bound at the active site. The apparent lack of metal ion effects on these proximal positions argues in favor of a C-terminal interaction; these regions are presumably far from the active site, and should be less influenced by the choice of metal cofactor.

Perhaps the most interesting finding is the region in the distal LTR which is both affected by mutation, and also crosslinks to integrase. This region exhibits a dramatic dependence on magnesium, which suggests that it is likely to be an interaction which involves the core domain.

Fig. 7. Structural summary of crosslinking data. The structure of the core domain of HIV-1 integrase near the active site is shown in white, with residues discussed in this study highlighted in color. The three active site residues, D64, D116 and E152 are shown in red, while K159 is shown in blue. Q62, a strongly conserved residue on the peptide (51–64) which is crosslinked to the C(–1) position is shown in yellow. Y143 is shown in green, and Q148 in cyan.

However, the levels of crosslinking are too low to have allowed identification of the exact region that is involved. It is unlikely that this region is contacting the same monomer of the core that is being contacted by the terminal LTR, which sits nearly 30 Å away. However, it is possible that this region is contacting the second monomer of the core found in the normal core dimer, or possibly even a completely different core dimer, since it is likely that the active complex includes at least two dimers of integrase. It is also important to note that this region, centered around position 9–10 is approximately one helical turn away from the region of terminal contacts, implying that both surfaces lie on the same face of the helix and may be contacted by adjacent integrase surfaces. These contacts are clearly vital in the presence of magnesium since mutation of positions 9–11 can nearly eliminate activity. It is not clear, however, whether these same contacts are required *in vivo* to the same extent. Some data seem to suggest that DNA past positions 6–7 are not necessary *in vivo*; however, there are several reports of mutations at other regions of the LTR which appear to affect viral replication. For instance, in Moloney murine leukemia virus, a mutation at positions 10 and 11 inhibits viral replication, and appears to suffer a defect in 3' processing (Murphy *et al.*, 1993). It is also possible that other retroviral integrases use slightly different patterns of DNA recognition. In the case of FIV integrase, the C-

terminal domain appears to not be required for 3' processing activity (Shibagaki *et al.*, 1997), suggesting that the stabilization by the HIV C-terminus may be carried out by a different portion of the protein, or may not be necessary at all.

Finally, there is the question of strand-transfer activity. Once cleavage occurs, strand transfer is carried out, presumably using the same active site that was used for the cleavage reaction. Strand transfer has previously been suggested to require much less sequence specificity than 3' processing (Balakrishnan and Jonsson, 1997). Our data confirm that strand transfer with precleaved substrates does not appear to be affected by sequences upstream of position 3, in the presence of either cation, while mutations at positions 1 and 2 show a significant effect on levels of strand transfer (data not shown). One explanation for this is that the DNA must be bound in a more specific confirmation for the 3' processing reaction than for strand transfer; thus binding of the upstream regions may no longer be required for the interactions involved in the integration step. This may suggest that once 3' processing occurs, there is a slight change in the orientation or positioning of the DNA in the active site which helps to position the DNA for strand transfer. However, the C-terminus of integrase is still required for strand transfer, although there is no evidence that it carries out the same role as it does in 3' processing.

Further determination of the exact contacts between the crosslinked DNA and integrase will help to locate precisely the position of the LTR on the protein. It is known that lysine 159, near the active site, is involved in interactions which may orient the DNA properly for cleavage (Jenkins *et al.*, 1997). It is possible that a similar orientation of viral DNA occurs upon binding of T(5) to the C-terminal domain, perhaps in the large cleft seen in the structure of that region. Isolation of the residues which contact the T(5) position may also help us to connect the structures of the C-terminal domain and the core to produce a clearer picture of the complete integrase–DNA complex.

Materials and methods

Preparation of HIV-1 integrase

Full-length HIV-1 integrase (F185K/C280S) was expressed in *E. coli* and purified from soluble extracts. Integrase contained a histidine tag at the N-terminus of the protein to facilitate purification. A thrombin cleavage site was engineered adjacent to the histidine tag to allow removal of the tag after purification. After sonication and preparation of a soluble extract, integrase was purified over a Ni-affinity column, incubated with thrombin to cleave the tag, and then purified over a benzamidine Sepharose column to remove thrombin. The protein was further purified over MonoS to chromatographic homogeneity as visualized by Coomassie Blue stain. The details of these purification procedures have been previously described (Jenkins *et al.*, 1996). Final integrase samples were dialyzed into buffer containing 1 M NaCl and 50 μ M zinc acetate. Concentrations were determined using calculated extinction coefficients based on amino acid composition. In order to reduce variability in specific activity, purified integrase was divided into separate aliquots and frozen in liquid nitrogen; aliquots were used only once, and any remaining protein was discarded. All experiments in this study were carried out using integrase from a single purification.

Preparation of oligonucleotides

Oligonucleotides containing mutations in the LTR sequence were synthesized on a Millipore Expedite DNA synthesizer at the 40 nmol scale. After deprotection at 55°C in ammonium hydroxide for 8 h, samples were dried down, resuspended in 10 mM Tris–HCl pH 8.0, and ethanol precipitated. The oligonucleotides were then gel purified on a 20% denaturing acrylamide gel, passively eluted from the gel, and ethanol precipitated. These oligonucleotides were completely pure as judged by chromatography of samples labeled with [γ -³²P]ATP. For oligonucleotides containing a 3' biotin tag, synthesis was carried out on a Biotin-TEG support matrix (Glen Research Inc.). Oligonucleotides containing 5-iododeoxyuracil or 5-iododeoxycytidine from Glen Research were synthesized using normal protocols, except that deprotection was carried out at room temperature for 24 h. Substrates for 3' processing and strand-transfer reactions were made by labeling the top strand oligonucleotide (containing the 3' biotin tag) at the 5' end with [γ -³²P]ATP (3000 Ci/mmol) using T4 polynucleotide kinase, and annealing the labeled primer with the bottom strand oligonucleotide as previously described (Jenkins *et al.*, 1997). Annealed substrates were purified over a Bio-Rad P6 column. Substrates for photocrosslinking were made in the same manner, except that the strand containing the photoactive residue was always labeled.

Integrase activity assays

Manganese (Mn)-based 3' processing reactions contained 25 mM 3-morpholinopropanesulfonic acid (MOPS) pH 7.2, 10 mM β -mercaptoethanol, 10% (v/v) glycerol, 100 μ g/ml bovine serum albumin, 10 mM MnCl₂, 5 nM DNA substrate, and 1 μ M integrase in a final volume of 100 μ L. NaCl was added to a final concentration of 50 mM including contributions from the protein. The magnesium (Mg)-based 3' processing reactions contained 25 mM 3-morpholinopropanesulfonic acid (MOPS) pH 7.2, 5% (v/v) PEG-8000, 5% (v/v) dimethylsulfoxide, 10 mM DTT, 10 mM MgCl₂, 15 mM potassium glutamate, 5 nM DNA substrate, and 1 μ M integrase in a final volume of 100 μ L. An additional 30 mM NaCl was introduced into these reactions with the protein. In both cases, reactions were incubated for varying amounts of time, and 10 μ L aliquots were removed, stopped with 5 μ L loading dye, and electrophoresed on NuPage 4–12% MES gels (Novex, Inc.). Quantitation of reaction products

was carried out using a PhosphorImager and ImageQuant 4.2 software (Molecular Dynamics).

Selection of LTR sequences

Selection substrates were constructed by mixing oligonucleotide DOM-155 (5'-CAGCTTCGTCGACGGTCAG) with one of the bottom strand degenerate oligonucleotides and annealing by heating to 95°C for 5 min followed by slow cooling to room temperature. Bottom strand primers had previously been phosphorylated at their 5' ends with T4 polynucleotide kinase and unlabeled ATP. The top strand of the annealed duplex was extended with the Klenow fragment of *E. coli* DNA polymerase, and the duplex substrate was purified over a Bio-Rad P-6 desalting column. Standard 3' processing assays were carried out under Mg or Mn conditions as described above, using 5 nM duplex substrate, for 2 h at 37°C. This sample was then ethanol precipitated, and resuspended in 10 mM Tris–HCl pH 8.0 (20 μ L). The LM-PCR linker was constructed by mixing oligonucleotides DOM-152 (5'-GTATGCACGATG) and DOM-153 (5'-CCCGTGACCGTGAGCATCGTGCAT), and annealing by heating to 95°C for 5 min followed by slow cooling to room temperature. A quarter of the reaction mixture was added to 0.7 μ M annealed linker in 1 $\times$ T4 DNA ligase buffer and 10% PEG-8000. Ten units of T4 DNA ligase were added to the total volume of 20 μ L, and ligation was carried out for 2 h at 25°C. The sample was diluted to 50 μ L with water, and desalted over a Bio-Rad P-6 column. A small sample (1 μ L) of the ligated product was used as template in a 50 μ L PCR containing 1 $\times$ PCR buffer, 2 mM MgCl₂, 100 μ M dNTPs, 100 ng flanking primers, and 2.5 U *Taq* polymerase. After 25 cycles of PCR, the samples were digested with *Eco*RI and *Hind*III, purified over Wizard columns (Promega), and ligated to pUC18 linearized with *Eco*RI and *Hind*III. The ligated DNA was transformed into *E. coli* DH5 α , and white colonies were picked and grown overnight in LB broth. DNA was prepared using the Qiagen Turbo method on a BioRobot 9600 (Qiagen, Inc.), and sequenced on an ABI 391 sequencer using the manufacturer's protocols (PE Biosystems). Two separate primers were used to sequence the inserts; one from the top strand on the left side of the pUC polylinker, and one on the opposite strand from the right side. Sequence analysis was carried out with EditView (PE Biosystems) and MacDNASIS (Hitachi Software).

Photocrosslinking and isolation of crosslinked peptides

Reaction mixtures contained the same components as the 3' processing reaction, with 5 nM DNA substrate and 1 μ M integrase in a final volume of 5 mL. Samples were incubated for 15 min on ice, and transferred to a 96-well microtiter plate and irradiated with 302 nm light from a handheld ultraviolet lamp with a mean flux of 2.0 mW/cm² for 35 min. Samples were then centrifuged at 16 000 *g* for 10 min, and the pellet resuspended in buffer containing 25 mM HEPES pH 7.5, 10 mM EDTA, 10 mM DTT, and 6 M urea, boiled for 5 min, and electrophoresed on a NuPage 4–12% MES gel (Novex, Inc.). The gel was autoradiographed, and portions corresponding to crosslinked product were excised, crushed in buffer containing 20 mM Tris–HCl pH 8.5 and 0.5% SDS, and allowed to elute passively overnight at 37°C. Recovery of labeled counts from this procedure was quantitative. Protein–DNA complexes were precipitated with 10% trichloroacetic acid (TCA), washed with acetone, dried, and resuspended in protease buffer. For cleavage with cyanogen bromide (CNBr), the samples were resuspended in 100 μ L 70% trifluoroacetic acid, and 100 mM CNBr was added, followed by incubation for 45 min at 70°C. These samples were dried down, and resuspended in NuPage gel loading buffer (Novex Inc.). For cleavage with endoproteinase Glu-C, samples were resuspended in 50 mM Tris–HCl pH 8.5, 5 mM EDTA, 5 mM DTT, 1 M NaCl and 0.1% SDS, and 200 μ g Glu-C (Boehringer Mannheim) were added per microgram of integrase and incubated for 12 h at 37°C. Samples were precipitated with TCA and acetone, dried, and resuspended in gel loading buffer. The peptides were then electrophoresed on NuPage 4–12% MES gels, and transferred to PVDF membranes (Trans-Blot, Bio-Rad Laboratories) using a Novex blotting apparatus. Radioactive peptides were identified by autoradiography, excised from the membrane, and analyzed on an HP 241 protein sequencer (Hewlett-Packard).

Acknowledgements

We thank Dr Tanya Paull for careful reading of the manuscript, and Dr Dale Ramsden for advice on the LM-PCR selection experiments. Molecular graphics images were produced using the MidasPlus program from the Computer Graphics Laboratory, University of California, San

Francisco. This work was supported by the National Institutes of Health Intramural AIDS Targeted Antiviral Program.

References

- Balakrishnan,M. and Jonsson,C.B. (1997) Functional identification of nucleotides conferring substrate specificity to retroviral integrase reactions. *J. Virol.*, **71**, 1025–1035.
- Bujacz,G., Alexandratos,J., Qing,Z.L., Clement-Mella,C. and Wlodawer,A. (1996) The catalytic domain of human immunodeficiency virus integrase: ordered active site in the F185H mutant. *FEBS Lett.*, **398**, 175–178.
- Burke,C.J., Sanyal,G., Bruner,M.W., Ryan,J.A., LaFemina,R.L., Robbins,H.L., Zeff,A.S., Middaugh,C.R. and Cordingley,M.G. (1992) Structural implications of spectroscopic characterization of a putative zinc finger peptide from HIV-1 integrase. *J. Biol. Chem.*, **267**, 9639–9644.
- 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.
- Cai,M., Zheng,R., Caffrey,M., Craigie,R., Clore,G.M. and Gronenborn,A.M. (1997) Solution structure of the N-terminal zinc binding domain of HIV-1 integrase. *Nature Struct. Biol.*, **4**, 567–577.
- Chow,S.A., Vincent,K.A., Ellison,V. and Brown,P.O. (1992) Reversal of integration and DNA splicing mediated by integrase of human immunodeficiency virus. *Science*, **255**, 723–726.
- Deitz,T.M., Trebra,R.J.v., Swanson,B.J. and Koch,T.H. (1987) Photochemical coupling of 5-bromouracil (BU) to a peptide linkage. A model for BU-DNA protein photocrosslinking. *J. Am. Chem. Soc.*, **109**, 1793–1797.
- Dyda,F., Hickman,A.B., Jenkins,T.M., Engelman,A., Craigie,R. and Davies,D.R. (1994) Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases. *Science*, **266**, 1981–1986.
- Eijkelenboom,A.P., Lutzke,R.A., Boelens,R., Plasterk,R.H., Kaptein,R. and Hard,K. (1995) The DNA-binding domain of HIV-1 integrase has an SH3-like fold. *Nature Struct. Biol.*, **2**, 807–810.
- Ellison,V. and Brown,P.O. (1994) A stable complex between integrase and viral DNA ends mediates human immunodeficiency virus integration *in vitro*. *Proc. Natl Acad. Sci. USA*, **91**, 7316–7320.
- 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.
- Engelman,A. and Craigie,R. (1995) Efficient magnesium-dependent human immunodeficiency virus type 1 integrase activity. *J. Virol.*, **69**, 5908–5911.
- Engelman,A., Bushman,F.D. and Craigie,R. (1993) Identification of discrete functional domains of HIV-1 integrase and their organization within an active multimeric complex. *EMBO J.*, **12**, 3269–3275.
- 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 non-specific DNA binding. *J. Virol.*, **68**, 5911–5917.
- Gerton,J.L., Ohgi,S., Olsen,M., DeRisi,J. and Brown,P.O. (1998) Effects of mutations in residues near the active site of human immunodeficiency virus type 1 integrase on specific enzyme-substrate interactions. *J. Virol.*, **72**, 5046–5055.
- Goldgur,Y., Dyda,F., Hickman,A.B., Jenkins,T.M., Craigie,R. and Davies,D.R. (1998) Three additional structures of the core domain of HIV-1 integrase: An active site that binds magnesium. *Proc. Natl Acad. Sci. USA*, **95**, 9150–9154.
- Haugan,I.R., Nilsen,B.M., Worland,S., Olsen,L. and Helland,D.E. (1995) Characterization of the DNA-binding activity of HIV-1 integrase using a filter binding assay. *Biochem. Biophys. Res. Commun.*, **217**, 802–810.
- Heuer,T.S. and Brown,P.O. (1997) Mapping features of HIV-1 integrase near selected sites on viral and target DNA molecules in an active enzyme-DNA complex by photo-cross-linking. *Biochemistry*, **36**, 10655–10665.
- Heuer,T.S. and Brown,P.O. (1998) Photo-cross-linking studies suggest a model for the architecture of an active human immunodeficiency virus type 1 integrase-DNA complex. *Biochemistry*, **37**, 6667–6678.
- Hiom,K. and Gellert,M. (1997) A stable RAG1-RAG2-DNA complex that is active in V(D)J cleavage. *Cell*, **88**, 65–72.
- Jenkins,T.M., Engelman,A., Ghirlando,R. and Craigie,R. (1996) A soluble active mutant of HIV-1 integrase: involvement of both the core and carboxyl-terminal domains in multimerization. *J. Biol. Chem.*, **271**, 7712–7718.
- Jenkins,T.M., Esposito,D., Engelman,A. and Craigie,R. (1997) Critical contacts between HIV-1 integrase and viral DNA identified by structure-based analysis and photo-crosslinking. *EMBO J.*, **16**, 6849–6859.
- Junop,M.S. and Hanford,D.B. (1997) Factors responsible for target site selection in Tn10 transposition: a role for the DDE motif in target DNA capture. *EMBO J.*, **16**, 2646–2655.
- Katz,R.A. and Skalka,A.M. (1994) The retroviral enzymes. *Annu. Rev. Biochem.*, **63**, 133–173.
- Katzman,M. and Sudol,M. (1996a) Influence of subterminal viral DNA nucleotides on differential susceptibility to cleavage by human immunodeficiency virus type 1 and visna virus integrases. *J. Virol.*, **70**, 9069–9073.
- Katzman,M. and Sudol,M. (1996b) Nonspecific alcoholysis, a novel endonuclease activity of human immunodeficiency virus type 1 and other retroviral integrases. *J. Virol.*, **70**, 2598–2604.
- Katzman,M. and Sudol,M. (1998) Mapping viral DNA specificity to the central region of integrase by using functional human immunodeficiency virus type 1/visna virus chimeric proteins. *J. Virol.*, **72**, 1744–1753.
- Katzman,M., Katz,R.A., Skalka,A.M. and Leis,J. (1989) The avian retroviral integration protein cleaves the terminal sequences of linear viral DNA at the *in vivo* sites of integration. *J. Virol.*, **63**, 5319–5327.
- Kulkosky,J., Jones,K.S., Katz,R.A., Mack,J.P. and Skalka,A.M. (1992) Residues critical for retroviral integrative recombination in a region that is highly conserved among retroviral/retrotransposon integrases and bacterial insertion sequence transposases. *Mol. Cell Biol.*, **12**, 2331–2338.
- LaFemina,R.L., Callahan,P.L. and Cordingley,M.G. (1991) Substrate specificity of recombinant human immunodeficiency virus integrase protein. *J. Virol.*, **65**, 5624–5630.
- 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.
- Murphy,J.E., De Los Santos,T. and Goff,S.P. (1993) Mutational analysis of the sequences at the termini of the Moloney murine leukemia virus DNA required for integration. *Virology*, **195**, 432–440.
- Polard,P. and Chandler,M. (1995) Bacterial transposases and retroviral integrases. *Mol. Microbiol.*, **15**, 13–23.
- Reicin,A.S., Kalpana,G., Paik,S., Marmon,S. and Goff,S. (1995) Sequences in the human immunodeficiency virus type 1 U3 region required for *in vivo* and *in vitro* integration. *J. Virol.*, **69**, 5904–5907.
- Rice,P., Craigie,R. and Davies,D.R. (1996) Retroviral integrases and their cousins. *Curr. Opin. Struct. Biol.*, **6**, 76–83.
- Rowland,S.J. and Dyke,K.G.H. (1990) Tn552, a novel transposable element from *Staphylococcus aureus*. *Mol. Microbiol.*, **4**, 961–975.
- Savilahti,H., Rice,P.A. and Mizuuchi,K. (1995) The phage Mu transpososome core: DNA requirements for assembly and function. *EMBO J.*, **14**, 4893–4903.
- Scottoline,B.P., Chow,S., Ellison,V. and Brown,P.O. (1997) Disruption of the terminal base pairs of retroviral DNA during integration. *Genes Dev.*, **11**, 371–382.
- 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.
- Shibagaki,Y. and Chow,S.A. (1997) Central core domain of retroviral integrase is responsible for target site selection. *J. Biol. Chem.*, **272**, 8361–8369.
- Shibagaki,Y., Holmes,M.L., Appa,R.S. and Chow,S.A. (1997) Characterization of feline immunodeficiency virus integrase and analysis of functional domains. *Virology*, **230**, 1–10.
- van den Ent,F.M., Vink,C. and Plasterk,R.H. (1994) DNA substrate requirements for different activities of the human immunodeficiency virus type 1 integrase protein. *J. Virol.*, **68**, 7825–7832.
- van den Ent,F.M., Vos,A. and Plasterk,R.H. (1998) Mutational scan of the human immunodeficiency virus type 2 integrase protein. *J. Virol.*, **72**, 3916–3924.
- van Gent,D.C., Elgersma,Y., Bolk,M.W., Vink,C. and Plasterk,R.H. (1991) DNA binding properties of the integrase proteins of human immunodeficiency viruses types 1 and 2. *Nucleic Acids Res.*, **19**, 3821–3827.
- 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.

- van Gent,D.C., Ramsden,D.A. and Gellert,M. (1996) The RAG1 and RAG2 proteins establish the 12/23 rule in V(D)J recombination. *Cell*, **85**, 107–113.
- Vink,C. and Plasterk,R.H. (1993) The human immunodeficiency virus integrase protein. *Trends Genet.*, **9**, 433–438.
- Vink,C., van Gent,D.C., Elgersma,Y. and Plasterk,R.H. (1991) Human immunodeficiency virus integrase protein requires a subterminal position of its viral DNA recognition sequence for efficient cleavage. *J. Virol.*, **65**, 4636–4644.
- Vink,C., Oude Groeneger,A.M. and Plasterk,R.H. (1993) Identification of the catalytic and DNA-binding region of the human immunodeficiency virus type I integrase protein. *Nucleic Acids Res.*, **21**, 1419–1425.
- Vink,C., Lutzke,R.A. and Plasterk,R.H. (1994) Formation of a stable complex between the human immunodeficiency virus integrase protein and viral DNA. *Nucleic Acids Res.*, **22**, 4103–4110.
- Willis,M.C., Hicke,B.J., Uhlenbeck,O.C., Cech,T.R. and Koch,T.H. (1993) Photocrosslinking of 5-iodouracil-substituted RNA and DNA to proteins. *Science*, **262**, 1255–1257.
- Willis,M.C., LeCuyer,K.A., Meisenheimer,K.M., Uhlenbeck,O.C. and Koch,T.H. (1994) An RNA–protein contact determined by 5-bromouridine substitution, photocrosslinking and sequencing. *Nucleic Acids Res.*, **22**, 4947–4952.
- 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.
- Yang,W. and Steitz,T.A. (1995) Recombining the structures of HIV integrase, RuvC and RNase H. *Structure*, **3**, 131–134.
- Yoshinaga,T. and Fujiwara,T. (1995) Different roles of bases within the integration signal sequence of human immunodeficiency virus type 1 *in vitro*. *J. Virol.*, **69**, 3233–3236.
- Zheng,R., Jenkins,T.M. and Craigie,R. (1996) Zinc folds the N-terminal domain of HIV-1 integrase, promotes multimerization and enhances catalytic activity. *Proc. Natl Acad. Sci. USA*, **93**, 13659–13664.

*Received July 6, 1998; revised July 29, 1998;
accepted August 5, 1998*