Supporting Information

Lim et al. 10.1073/pnas.1001402107

SI Text

Supporting Materials and Methods. ***Mapping of viral integration sites.*** A previously developed, high-throughput method for identifying human immune deficiency virus integration sites (1) was adapted to our system as follows. The genomic DNA of infected cells was isolated using Qiagen mini DNA kit (QIAGEN). The genomic DNA was then digested with various enzymes—BamH I, Hinc II, HpyCH4 III, HpyCH4 IV, Mse I, Hind III and Taq^{a}I [New England Biolabs (NEB)]—to minimize potential biases introduced by the choice of a few restriction enzymes. The fragmented genomic DNA was then used as the template for linear amplification with a biotinylated primer, bMLV, (5′- b-ATTTGTTAAAGACAGGATATCAGTGGTCCAG-3′) complementary to a sequence upstream of the 3′ long terminal repeat (LTR) of integrated viral genome. The resulting amplification product was selectively concentrated using Dynabeads M-280 Streptavidin (Invitrogen) after PCR cleanup and then restricted by Mme I (NEB). The restricted fragments containing viral-host genome junctions were ligated to preannealed linker DNA (linkA, 5′-CGGATCCCGCATCAT-3′, linkB, 5′- TGTCACACCTGGAGATATGATGCGGGATCCGNN -3′) and then used for PCR with primers annealing to the 3′ LTR and the linker (U5_5L, 5′- CTC CTGGATCCCCTCTTGCAGTTGCATCCGTCTT-3′, hi-linker, 5′-TCACACCTGGAGATATGATGCGG-3′). The amplified viral-host genome junctions were cloned into pBS SK SP plasmid after restriction with BamH I (NEB) (the BamH I sites within primers is underlined above) and then sequenced. As a complementary approach, the conventional method for cloning viral-host genome junctions was also applied here as described in detail elsewhere (2). The retroviral-host junction sequences were mapped to locations on the human genome (February 2009 assembly) using the BLASTN program on the Ensembl genome browser web site (www.ensembl.org). The chromosomal location mapping was considered for further statistical analysis only when the following conditions were satisfied: (*i*) no deletion was found before and at the CA sequence in the 3′ end of 3′ LTR (Fig. S1*B*) and (*ii*) a single chromosomal location can be assigned. Based on the chromosomal locations various genomic annotations for each retroviral integration site were made via genome browser web sites [Ensembl (www.ensembl.org) and UCSC genome browser (genome.ucsc.edu)].

Probability of having retroviral integrations at the same locations on the human genome. The probability that both 16.p12.zfd1 and 375.RT.zfd1 clones integrated into the same location on the human genome is calculated as follows:

$$\text{The size of human genome} = 3.1 \times 10^9.$$

For 16.p12.zfd1, the number of distinct virus-host junction sequences matching a single location on the human genome = 2

For 375.RT.zfd1, the number of distinct virus-host junction sequences matching a single location on the human genome = 4

Multiple virus-host junctions with the same sequence for each mutant case were counted one.

$$\begin{aligned} P &= {}_2C_1 \times (4/(3.1 \times 10^9)) \times ((3.1 \times 10^9 - 4)/(3.1 \times 10^9)) \\ &\sim {}_4C_1 \times (2/(3.1 \times 10^9)) \times ((3.1 \times 10^9 - 2))/(3.1 \times 10^9))^3 \\ &= 2.6 \times 10^{-9}. \end{aligned}$$

The probability that both 273.IN.zfd2 and 375.RT.zfd2 clones integrated into four common locations on the human genome is calculated as follows:

For 273.IN.zfd2, the number of distinct virus-host junction sequences matching a single location on the human genome = 22

For 375.RT.zfd2, the number of distinct virus-host junction sequences matching a single location on the human genome = 14

Multiple virus-host junctions with the same sequence for each mutant case were counted one.

$$\begin{aligned} P &\sim {}_{14}C_1 \times (22/(3.1 \times 10^9)) \times {}_{13}C_1 \\ &\quad \times (21/(3.1 \times 10^9)) \times {}_{12}C_1 (20/(3.1 \times 10^9)) \\ &\quad \times {}_{11}C_1 \times (19/(3.1 \times 10^9)) \times ((3.1 \times 10^9 - 22)/(3.1 \times 10^9))^{10} \\ &\sim {}_{22}C_1 \times (14/(3.1 \times 10^9)) \times {}_{21}C_1 \times (13/(3.1 \times 10^9)) \times {}_{20}C_1 \\ &\quad \times (12/(3.1 \times 10^9)) \times {}_{19}C_1 \times (11/(3.1 \times 10^9)) \\ &\quad \times ((3.1 \times 10^9 - 14)/(3.1 \times 10^9))^{18} = 4.6 \times 10^{-29}. \end{aligned}$$

1. Kim S, Kim Y, Liang T, Sinsheimer JS, Chow SA (2006) A high-throughput method for cloning and sequencing human immunodeficiency virus type 1 integration sites. *J Virol* 80:11313–11321.

2. Wu X, Li Y, Crise B, Burgess SM (2003) Transcription start regions in the human genome are favored targets for MLV integration. *Science* 300:1749–1751.

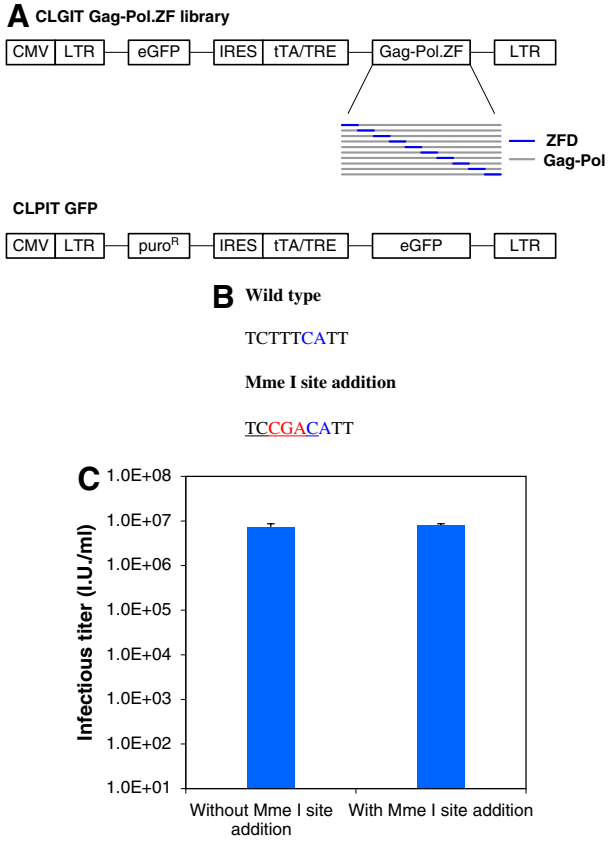

Fig. S1. (*A*) Retroviral vector constructs, including the CLGIT Gag-Pol.ZF library and CLPIT GFP. The library size was 4.3×10^5, estimated by colony counting after transformation of electrocompetent *Escherichia coli*. Based on the 5,214 nucleotide positions in murine leukemia virus (MLV) *gag-pol*, the library covered most or all of the possible insertion sites. LTR, long terminal repeat; IRES, internal ribosome entry site; tTA, tetracycline-controlled transactivator; TRE, tetracycline response element. (*B*) Sequence of the 3′ end of MLV 3′ LTR. For integration site analysis based on high-throughput linear amplification mediated PCR, a Mme I site (underlined) was introduced into the end of 3′ LTR by mutations of three nucleotides (in red color). The last two nucleotides (TT) are cleaved after reverse transcription, and the resulting CA sequence end (in blue color) with an active 3′-OH is involved in integration (1). This mutation did not impact infectious titer of MLV vector with wild-type Gag-Pol (*C*). (*C*) Infectious titers of MLV vector with wild-type Gag-Pol and with and without Mme I site addition into 3′ LTR. Viruses were used for infection of 293T cells after concentration via ultracentrifugation. Infectious titers of viruses were measured in duplicate. Standard deviations of infectious titer are shown by bars.

1. Coffin JM, Hughes SH, Varmus HE (1997) *Retroviruses* (Cold Spring Harbor Lab Press, Cold Spring Harbor, NY), 1st Ed.

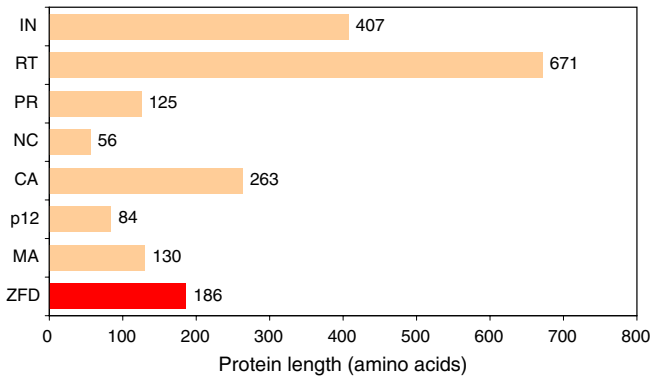

Fig. S2. Amino acid lengths of MLV Gag Pol proteins relative to the polydactyl zinc finger domains.

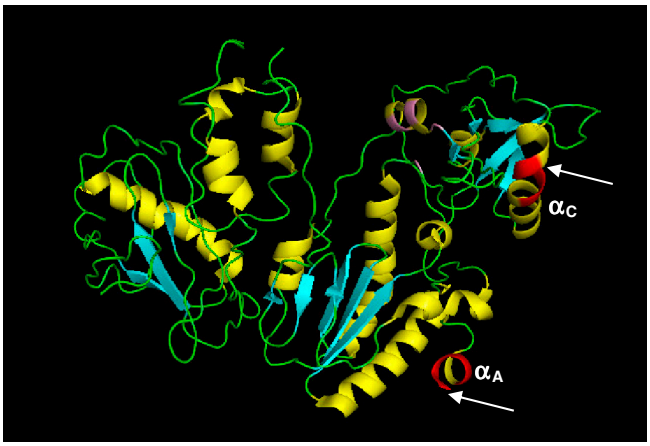

Fig. S3. Structure of wild-type MLV reverse transcriptase. The arrows indicate the zinc finger domain (ZFD) insertion sites, and regions following the inserted zinc fingers are shown in red. The four key residues (Asp114, Arg116, Asn119, and Gly191) involved in the contacts between reverse transcriptase and the primer-template duplex are shown in pink. The orientations of inserted ZFDs and the comprehensive effects of the insertions on the conformations of Gag-Pol proteins could not be reliably predicted by homology-based modeling, due to the lack of the crystal structures of the full length MLV proteins and suitable protein templates with the sequences similar to the fusion protein junctions. The protein structure (PDB ID code 1RW3) was rendered using PyMol, and note that the N-terminal 23 amino acids (prior to the αA chain) were truncated in the original structure (1).

1. Das D, Georgiadis MM (2004) The crystal structure of the monomeric reverse transcriptase from Moloney murine leukemia virus. *Structure* 12:819–829.

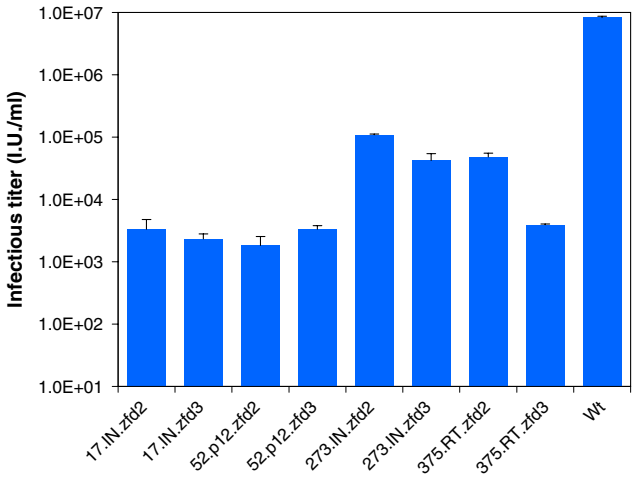

Fig. S4. Infectious titers of wild-type virus and ZFD insertion mutants. Viruses were used for infection of 293T cells after concentration via ultracentrifugation. Infectious titers of viruses were measured in duplicate. Standard deviations of infectious titer are shown by bars.

Table S1. The number of sequenced virus-host genome junctions

	Total number of sequenced junctions	The number of junctions matching a single location on the human genome
WT	294	107
17.IN.zfd1	76	10
16.p12.zfd1	6	2
375.RT.zfd1	34	4
273.IN.zfd1	8	2
273.IN.zfd2	197	22
273.IN.zfd3	101	9
375.RT.zfd2	93	14

Table S2. Virus-host genome junction sequences identified for 273.IN.zfd2 and 273.IN.zfd3

Clones	Virus-host junction sequence	The number of junctions with the shown sequence	Total number of sequenced junctions for each mutant	% out of the total number of sequenced junctions	Integration position on the human genome	
					Chromosome	Position
273.IN.zfd2	ACATTTGGTAGCTGGGATGTTAG	22	197	11.2	X	153357666
273.IN.zfd2	ACATGATGGGCAAAAGTCACCC	10	197	5.1	1	94142416
273.IN.zfd2	ACATTTGGGGGCTAAACAAATTT	10	197	5.1	X	113308851
273.IN.zfd2	ACATTACAATTACTAATAGTAT	8	197	4.1	2	13496337
273.IN.zfd2	ACATTTTCTTTCTAAAGTGCCTT	6	197	3.0	15	93845542
273.IN.zfd2	ACATTTCATTAAGCGTAGGGGC	6	197	3.0	13	74454561
273.IN.zfd2	ACATTTGGGCTTGGTAGGATTGG	5	197	2.5	7	86384864
273.IN.zfd2	ACATTTGGGGGCTTAATATTTTT	4	197	2.0	16	52422344
273.IN.zfd2	ACATCTAGACAGTTCAGAAAAA	3	197	1.5	2	146335293
273.IN.zfd2	ACAGTTGGAGGGCACAGGATAT	3	197	1.5	21	41419380
273.IN.zfd2	ACATCGGATCCCCCGGGCTGCA	1	197	0.5	11	17549051
273.IN.zfd2	ACATTTGGGATCTGCCCATATCA	1	197	0.5	11	116161720
273.IN.zfd2	ACATTTGGGGTCCTACATTCAG	1	197	0.5	16	58516469
273.IN.zfd2	ACATTTTCTTTCTAAGGTGCCTT	1	197	0.5	4	173871541
273.IN.zfd2	ACATTTACTTTCTAAAGTGCCT	1	197	0.5	8	12828475
273.IN.zfd2	ACATTTGGGCTCCAAGATATTGC	1	197	0.5	7	62333487
273.IN.zfd2	ACATTTGGTAGCTGGTATGTTG	1	197	0.5	3	140396792
273.IN.zfd2	ACATTTGAAGTCCACAGCCTGCTGG	1	197	0.5	11	11519998
273.IN.zfd2	ACATTGAGTCAAAACTAGAGCCT	1	197	0.5	15	88175538
273.IN.zfd2	ACATTTGGGGGCCTGGACCACT	1	197	0.5	18	43408004
273.IN.zfd2	ACATTGGTAGCTGGGATGTTAGG	1	197	0.5	X	153357665
273.IN.zfd2	ACATTTGGAGACTAAATAAAAAC	1	197	0.5	1	97789945
273.IN.zfd3	ACATATTAATAGCAGTGTGCAC	18	101	17.8	5	93023433
273.IN.zfd3	ACATTACCCTGGACCACTGATAT	15	101	14.9	14	52491298
273.IN.zfd3	ACATTGAGCCTGGACCACTGAT	7	101	6.9	16	75208997
273.IN.zfd3	ACATTGGAGGGGAGGAATGCCGT	5	101	5.0	18	46415241
273.IN.zfd3	ACATTTGGGGGCCAAGCTGCCT	5	101	5.0	16	56729532
273.IN.zfd3	ACATTACCCTGGACCACTGATGT	3	101	3.0	8	143413073
273.IN.zfd3	ACATTTGGCGCCCCGAGTGAGGG	1	101	1.0	1	180199185
273.IN.zfd3	ACATTACCCTGGACACTGATATC	1	101	1.0	17	8050843
273.IN.zfd3	ACATTTCCCTTACCTACATTGTT	1	101	1.0	19	24183658

Virus-host genome junction sequences are shown, along with the fraction of total junctions corresponding to that integration position for each mutant. The underlined sequences are viral, and the others are from the host genome.