Analysis of Wild-Type and Mutant SL3-3 Murine Leukemia Virus Insertions in the c-*myc* Promoter during Lymphomagenesis Reveals Target Site Hot Spots, Virus-Dependent Patterns, and Frequent Error-Prone Gap Repair

Anne Ahlmann Nielsen,[1] Annette Balle Sørensen,[1] Jörg Schmidt,[2] and Finn Skou Pedersen[1,3]*

Department of Molecular Biology[1] and Department of Medical Microbiology and Immunology,[3] University of Aarhus, Aarhus, Denmark, and Department of Comparative Medicine, GSF-National Research Center for Environment and Health, Neuherberg, Germany[2]

The murine leukemia retrovirus SL3-3 induces lymphomas in the T-cell compartment of the hematopoetic system when it is injected into newborn mice of susceptible strains. Previously, our laboratory reported on a deletion mutant of SL3-3 that induces T-cell tumors faster than the wild-type virus (S. Ethelberg, A. B. Sørensen, J. Schmidt, A. Luz, and F. S. Pedersen, J. Virol. 71:9796–9799, 1997). PCR analyses of proviral integrations in the promoter region of the c-*myc* proto-oncogene in lymphomas induced by wild-type SL3-3 [SL3-3(wt)] and the enhancer deletion mutant displayed a difference in targeting frequency into this locus. We here report on patterns of proviral insertions into the c-*myc* promoter region from SL3-3(wt), the faster variant, as well as other enhancer variants from a total of approximately 250 tumors. The analysis reveals (i) several integration site hot spots in the c-*myc* promoter region, (ii) differences in integration patterns between SL3-3(wt) and enhancer deletion mutant viruses, (iii) a correlation between tumor latency and the number of proviral insertions into the c-*myc* promoter, and (iv) a [5′-(A/C/G)TA(C/G/T)-3′] integration site consensus sequence. Unexpectedly, about 12% of the sequenced insertions were associated with point mutations in the direct repeat flanking the provirus. Based on these results, we propose a model for error-prone gap repair of host-provirus junctions.

The retroviral replication cycle includes an obligate integration step in which a reverse-transcribed double-stranded DNA copy of the RNA genome is inserted into the genome of the target cell. The integration step is initiated by a removal of typically 2 bases at either 3′ end and is followed by a DNA strand transfer reaction where the retrovirally encoded enzyme integrase (IN) catalyzes the joining of the two ends to the host target DNA a few bases apart. The repair of the two resulting gaps includes DNA synthesis, removal of the two viral 5′ dinucleotide overhangs generated during 3′ processing, and ligation of the resulting nicks (for a review, see reference 21). In general, a virus-specific stretch of 4 to 6 bp of the host DNA is duplicated at the integration site; however, atypical virus-host DNA junctions, including alterations in repeat length and point mutations, are occasionally generated (14, 30, 39, 52, 71, 72).

In contrast to the retrovirus-like Ty retrotransposons in yeast, which are very selective in the choice of integration sites (as reviewed in reference 9), retroviruses integrate throughout the chromosomes (10, 22, 33, 37, 42, 43, 46, 49, 50, 63, 65, 67, 70, 75, 76). Although studied in vivo and by use of simplified in vitro models during the last decades (for a review, see references 8 and 36), integration site selection still remains poorly understood. Based on these reports, factors such as nucleosomal structure, DNase I-hypersensitive sites, and methylation seem to affect integration (44, 55–58, 60, 73). Moreover, genes appear to be favored targets for both murine leukemia viruses (MLVs) and human immunodeficiency virus type 1 (HIV-1) as examined in cell cultures (63, 76). In contrast to MLV, which prefers integration near the start of transcriptional units, the entire transcriptional unit except upstream of the transcriptional start is favored by HIV-1 (76).

Simple non-acutely transforming retroviruses induce hematopoietic malignancies by a complex process including insertional mutagenesis of host genes (as reviewed in references 41 and 51). Extensive analyses of proviral integration sites in mice, cats, rats, and birds have revealed that c-*myc* is one of the most frequently targeted genes (for a review, see references 20, 41, and references therein). In chickens, 3′ promoter insertion is the predominant form of activation, while c-*myc* expression is deregulated primarily by enhancer activation in mammals. In mice, both MLVs of the gammaretrovirus genus (e.g., Moloney MLV [Mo-MLV], SL3-3, and MCF 69L1) and the thymotropic betaretroviral leukemia virus (TBLV), which is closely related to mouse mammary tumor virus (5), target this proto-oncogene, giving rise to hematopoietic malignancies such as T-cell lymphomas and erythroleukemias (3, 6, 19, 33, 41, 43, 49, 54, 59, 67, 70). In SL3-3-induced T-cell lymphomas, the proviral insertions are predominantly located upstream of the first exon and the majority of insertions are integrated in the transcriptionally opposite orientation relative to that of c-*myc* (3, 43, 67).

One of the major leukemogenic determinants of MLVs is the transcriptional enhancer containing densely packed binding sites for a variety of transcription factors. MLVs are able to

* Corresponding author. Mailing address: Department of Molecular Biology, University of Aarhus, C. F. Møllers Allé Bldg. 130, DK-8000 Aarhus C, Denmark. Phone: 45 8942 3188. Fax: 45 8619 6500. E-mail: fsp@mb.au.dk.

infect a broad spectrum of cell types; however, tumor development is favored in target cells expressing the array of transcription factors that matches the profile of the enhancer framework of the infecting retrovirus (41). As we have previously reported, the introduction of an extra wild-type repeat in combination with the deletion of two NF1 (nuclear factor 1) binding sites in the SL3-3 enhancer region generates a potent inducer of T-cell tumors, the SL3-3(turbo) virus (also known as SL3-3[2Δ18-3]) (26). In contrast to wild-type SL3-3 [SL3-3(wt)]-induced tumors, of which 20 to 25% display clonal rearrangements in the c-*myc* locus due to proviral insertions (35, 48, 53, 67), initial data from lymphomas induced by this enhancer variant demonstrated no such rearrangements (23). However, subsequent PCR analyses of proviral integration sites in 12 tumors induced by SL3-3(turbo) revealed that 92% of these harbored nonclonal insertions in c-*myc* (A. A. Nielsen, A. B. Sørensen, and F. S. Pedersen, unpublished data). In order to make a thorough examination of this observation, about 250 lymphomas induced by SL3-3(wt) and a variety of enhancer variants of SL3-3 were tested for insertions in the promoter region of this proto-oncogene. Results from this study point to (i) integration site hot spots in the c-*myc* promoter region, (ii) frequent cases of various atypical provirus-host junction structures, (iii) a virus-dependent distribution of integration sites, and (iv) a correlation between latency and the number of proviral insertions detected per tumor.

MATERIALS AND METHODS

Pathogenicity experiments. Newborn inbred or randomly bred NMRI mice were injected with SL3-3(wt) or enhancer variants of the SL3-3 MLV. Tumors isolated from female mice originated from experiments performed by Hallberg et al. (32), Ethelberg et al. (25, 26), and Lund et al. (48). In brief, in these experiments, the animals were injected with 10^5 to 10^7 infectious virus particles and, as controls, were mock injected with complete medium. In the new injection round with SL3-3(wt) and SL3-3(turbo) injected into newborn male inbred NMRI mice, 40 and 115 mice were injected with 10^3 to 10^6 infectious SL3-3(wt) and enhancer variant virus particles, respectively. The number of infectious virus particles was measured by infectious center assay as described previously (62). As a control, 10 mice were mock injected with complete medium. The mice were checked for tumor development on 5 days of the week and killed at the time of apparent illness. Sacrificed animals were autopsied and diagnosed according to criteria described previously (61).

DNA. Genomic DNA was extracted from frozen tumor tissues with a DNeasy tissue kit (QIAGEN) according to the manufacturer's instructions.

PCR amplification. c-*myc*-specific PCRs were performed with a 50-μl volume containing 5 μl of 10× *Taq* buffer (Invitrogen), a 0.2 mM concentration of each deoxynucleoside triphosphate (Invitrogen), 1.25 U of *Taq* DNA polymerase (5 U/μl; Invitrogen), 1.5 mM MgCl$_2$ (Invitrogen), and 10 pmol of each primer (see below). One hundred to 1,000 ng of genomic tumor DNA was used as the template in the PCRs. The primers were as follows: v1, 5′-XGAATTCGATATC GATCCCCGGTCATCTGGG-3′; v2, 5′-XTGCGGCCGCGATTCCCAGATG ACCGGGGGATC-3′ (the underlined sequences in v1 and v2 anneal to the viral sequence, and the remainder of the primer sequences consists of linker sequences added for other purposes, with X being biotin); myc1, 5′-TGTGTATG TATACGGTTTGGGGATTGTAC-3′; and myc2, 5′-CACTCCAGCACCTCCG GTTCCGGACT-3′. The proviral primers v1 and v2 correspond to positions 8204 to 8187 and 8187 to 8204, respectively, of GenBank and EMBL database accession number AF169256. The two gene-specific primers myc1 and myc2 correspond to positions 77 to 104 and 710 to 686, respectively, of GenBank and EMBL database accession number M12345. Oligonucleotides were synthesized at DNA Technology ApS, Aarhus, Denmark. The fragments were amplified in a Touch-Down thermal cycler (Hybaid) with the following program: 1 cycle of denaturation at 94°C for 3 min and then 40 cycles of denaturation at 94°C for 1 min, annealing at 62°C for 1 min, and extension at 72°C for 3 min, and finally 1 cycle of extension at 72°C for 10 min. Subsequently, the amplification products were visualized on ethidium bromide-stained 1.5% agarose (Invitrogen Life Technol-

ogies) gels in 0.5× Tris-borate-EDTA (Invitrogen Life Technologies) buffer. The GeneRuler 1-kb DNA ladder and the GeneRuler 100-bp DNA ladder were purchased from Fermentas.

Purification of PCR products. Amplified PCR products were purified by using streptavidin-coated magnetic beads (Dynabead M280-streptavidin; Dynal AS, Oslo, Norway), the Wizard DNA clean-up System (Promega), or the GFX PCR DNA and GelBand purification kit (Amersham Biosciences) according to the manufacturer's instructions.

Sequencing procedure and sequence comparison. The amplified PCR products were sequenced with the DYEnamic ET terminator cycle sequencing kit (Amersham Pharmacia Biotech) by following the manufacturer's recommendations, and reaction products were analyzed on an automated DNA sequencer (genetic analyzer model 3100; Applied Biosystems Inc.). Amplification products harboring the 5′ long terminal repeat (LTR) or 3′-LTR regions were sequenced with primer v3 (5′-CTCTGGTATTTTCCCATG-3′) and primer v4 (5′-TCCGA ATCGTGGTCTCGCTGATCCTTGG-3′), respectively. Sequencing primers v3 and v4 correspond to positions 7904 to 7886 and 69 to 96, respectively, of GenBank and EMBL database accession number AF169256. Both oligonucleotides were purchased from DNA Technology ApS, Aarhus, Denmark. Sequencing of enhancer regions was performed by using the viral primer v1. Sequences were edited by use of the program Sequencer, version 3.0 (Gene Codes Corporation), and Vector NTI (InforMax, Inc.). Edited sequences were compared with available sequences in databases by using the nucleotide-nucleotide BLAST (basic local alignment search tool) search tool with comparison to sequences in the nonredundant database (http://www.ncbi.nlm.nih.gov/BLAST/), the mouse genome BLAST search tool (http://www.ncbi.nlm.nih.gov/genome/seq/MmBlast .html), and the Ensembl (http://www.ensembl.org/Mus_musculus/) and University of California at Santa Cruz (assembly date, October 2003) (http://genome .UCSC.edu/) assemblies.

Statistical analysis. The difference in the distributions of target sites in c-*myc* selected for during tumor development induced by SL3-3(wt) and SL3-3(turbo) was tested by use of Fisher's exact test for count data (1).

The proviral insertions in c-*myc* per tumors were analyzed in groups with respect to the latency periods, assuming a Poisson distribution. This assumption was tested by calculating a conditional chi-square statistic. This test statistic was shown to work well even for small Poisson λ parameters (7). The Poisson parameter is simply estimated by the average number of proviral insertions (m_i) in each group. The confidence intervals for the Poisson parameters are based on the following equations: $\theta_1 = \gamma_{T,2:\alpha/2}/(2n)$ and $\theta_u = \gamma_{T + 1,2;1 - \alpha/2}/(2n)$, where $\gamma_{\alpha,\beta;\delta}$ is the deviate associated with the lower tail probability δ of the gamma distribution with a shape parameter α and scale parameter β (27).

To test whether the parameters of two independent Poisson random variables (λ_1 and λ_2) are equal, a binomial test is used (40). This test is based on the number of successes ($n_1 \cdot m_1$), the number of trials ($n_1 \cdot m_1 + n_2 \cdot m_2$), and the hypothesized probability of success [$n_1/(n_1 + n_2)$], where m_i and n_i are the average number of insertions and the sample size in group i ($i = 1$ or 2).

All the statistical calculations were done with R—a language and environment for statistical computing and graphics which is available as free software (www .r-project.org) (38).

RESULTS

In female inbred NMRI mice, the enhancer variant SL3-3(turbo) induces T-cell lymphomas with 100% incidence and a mean latency period 20% shorter than that of SL3-3(wt). As shown in Fig. 1, the SL3-3(turbo) enhancer harbors two identical 18-bp deletions encompassing the NF1 site in addition to an extra wild-type 72-bp repeat.

Previously, in order to study proviral insertion sites in 12 SL3-3(turbo)-induced tumors, 45 proviral flanking tags were amplified (Nielsen et al., unpublished) by use of a simple two-step PCR method (68). Comparison of tags within publicly available databases revealed that 22% of the sequences demonstrated similarity to the c-*myc* promoter region. Subsequently, the insertions except for one positioned about 4.2 kb upstream of exon 1 were verified by using the gene-specific PCR approach shown in Fig. 2. Furthermore, additional proviral insertions in the c-*myc* gene were detected, and all together, 20 different proviral insertions in this proto-oncogene
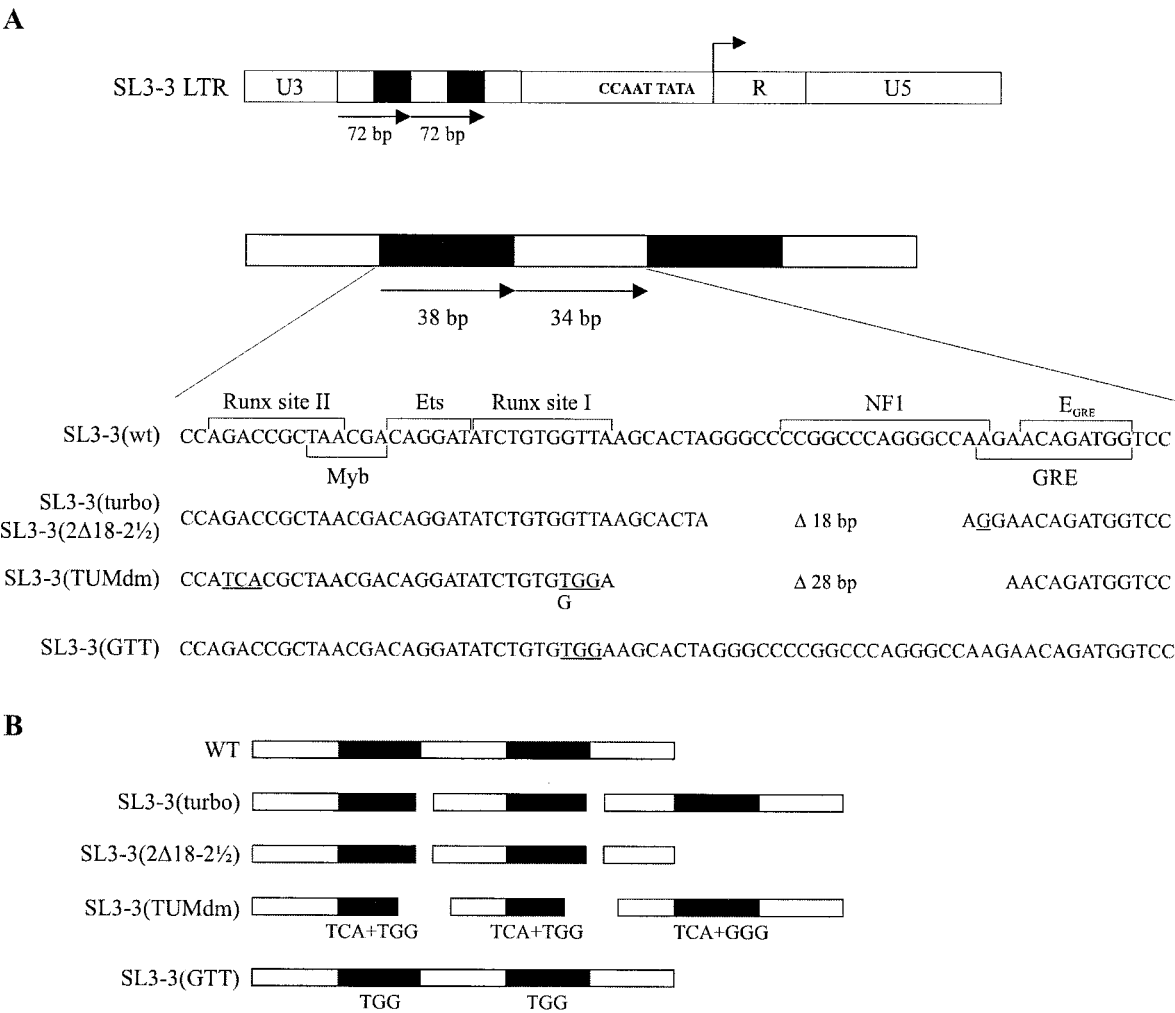
A

B

FIG. 1. Enhancer structures of wild-type and mutant SL3-3 MLVs used in the study (24, 25, 26). (A) The SL3-3(wt) enhancer located in U3 consists of a 72-bp direct repeat followed by a third repetition of 34 bp containing binding sites for a variety of transcription factors. Both SL3-3(turbo) and SL3-3(2Δ18–2½) harbor two identical 18-bp deletions encompassing the NF1 site. In addition, the SL3-3(turbo) enhancer contains an extra 72-bp wt repeat. SL3-3(GTT) and SL3-3(TUMdm) are both Runx binding site mutants. The SL3-3(GTT) enhancer contains the GTT-to-TGG mutation in both Runx site I binding sites. The SL3-3(TUMdm) enhancer harbors several alterations compared to the SL3-3(wt) enhancer structure. In addition to an extra 72-bp repeat, the enhancer harbors two identical 28-bp deletions encompassing the NF1 site. Furthermore, GTT-to-TGG and GAC-to-TCA mutations are present in Runx site I and site II, respectively. A T-to-G base substitution is present in the last Runx site I. (B) The enhancer structures are illustrated schematically with boxes; deletions are marked by gaps, and Runx binding site mutations are indicated by letters. WT, wild type.

originating from 11 tumors were identified. Similarly, 18 SL3-3(wt)-induced tumors from the same injection round (25, 26) were analyzed by the same PCR approach. In contrast to the enhancer variant, only 7 of the 18 tumors displayed insertions in the c-*myc* promoter region. These results suggest a correlation between tumor latency and proviral insertions in c-*myc*. To analyze this finding further, 110 and 40 inbred male NMRI mice were injected with SL3-3(turbo) and SL3-3(wt) viruses, respectively. The enhancer variant induced tumors with a mean latency period about 13% shorter than that of SL3-3(wt) ($P < 0.001$, as calculated by Student's two-sample t test grouped by GROUP), a result resembling the previously observed difference in female mice (26). However, we note that the results suggest a lower c-*myc* targeting frequency in SL3-3(wt)-induced tumors in females (39 to 60%) than in males (95%) of

the inbred NMRI mouse strain. At present, the reason for this difference remains uncertain.

Integration site hot spots in the c-*myc* promoter region. Tumor DNAs isolated from the new injection round were analyzed with respect to provirus insertions in the c-*myc* promoter region by use of the gene-specific PCR approach, and the results are summarized in Table 1. In addition, tumors induced by SL3-3(wt) and a variety of other enhancer mutants of SL3-3 (enhancer structures are illustrated in Fig. 1) from previous experiments (25, 26, 32, 48) were analyzed. The c-*myc*-specific PCRs were done in duplicate or triplicate except with tumors originating from the study by Lund et al. (48). To visualize the amplification products, 20% of the PCR volume was loaded onto ethidium bromide-stained agarose gels (Fig. 2B).
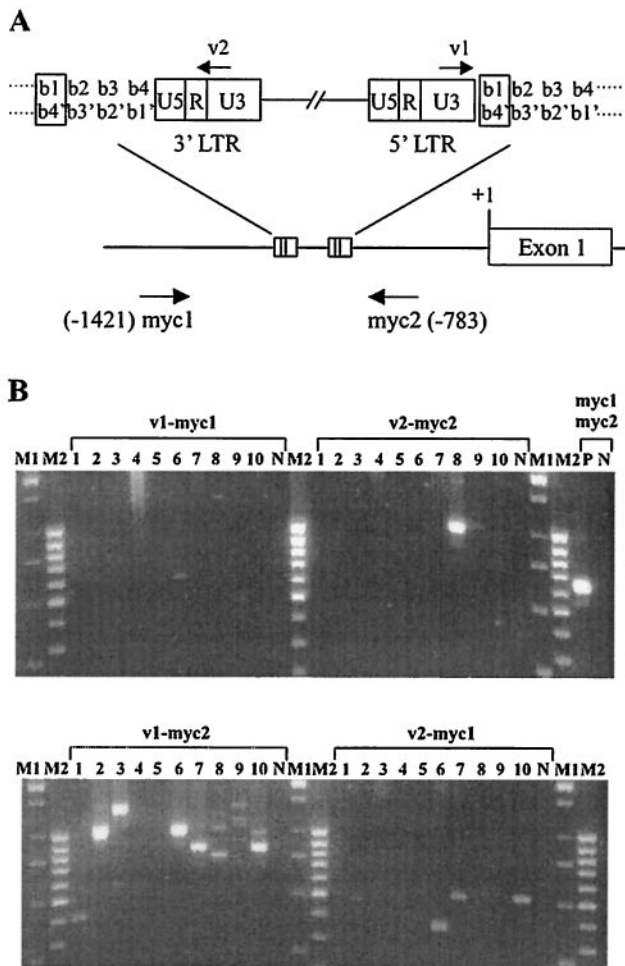
FIG. 2. Detection of proviral integrations in the c-*myc* promoter region. (A) Provirus integrations in the c-*myc* promoter region in the SL3-3(wt) and variant-induced tumors were detected by using a PCR method employing two virus-specific (v1 and v2) and two gene-specific (myc1 and myc2) primers. PCRs with the four primer combinations v1-myc1, v2-myc1, v1-myc2, and v2-myc2 were performed in order to detect viruses integrated in either orientation. The four base pairs b1-b4′ to b4-b1′ duplicated during proviral integrations of MLVs are shown. We have defined the first base pair (in the c-*myc* sense orientation) in this duplicated stretch of bases (b1-b4′) to be the site of integration (boxed). (B) The amplified PCR products were visualized on ethidium bromide-stained agarose gels. As examples, amplification products obtained from 10 SL3-3(turbo)-induced tumors in male mice are shown. M1, 1-kb ladder (2, 1.5, 1, 0.75, 0.5, and 0.25 kb); M2, 100-bp ladder (1,031, 900, 800, 700, 600, 500, 400, 300, and 200 bp, respectively); N, negative control; P, positive control. Tumor numbers and primer combinations are listed above the lanes.

To obtain an overview of the dispersion of proviral insertions along the c-*myc* promoter region, 212 PCR amplification products were sequenced and validated. Approximately half of these represent 5′-LTR- and 3′-LTR-flanking tags from the same proviral insertions, and in total, the obtained sequences represent 164 different integrations in the c-*myc* promoter region. During MLV retroviral integration, a stretch of 4 bp is duplicated at the integration site (as reviewed in reference 74) (Fig. 2A), and we have defined the base pairs located at positions b1 to b4′ to be the site of integration. According to this definition, positions and orientations of the 164 insertions are

illustrated in Fig. 3. The indicated integration sites specify the upstream positions relative to transcription start site (+1) at exon 1 (c-*myc* promoter P1). All insertions except for one (at position −4219) are located within a stretch of 2,100 bp, and the majority of the proviruses are in the transcriptionally antisense orientation relative to that of c-*myc*. Interestingly, several integration site hot spots where up to nine proviruses of different tumors are integrated at the exactly same base pair (at some sites in both orientations) were identified. To eliminate the possibility that these hot spots were due to contamination artifacts from the PCRs, the majority of the 5′-LTR enhancer regions were sequenced. Due to the experimental setup, it was not possible to directly obtain enhancer sequences from the 3′-LTR-flanking tags. In all cases except for a few recombined mink cell focus-forming (MCF) viruses, the enhancer structures represented the injected virus type (occasionally with gains or losses of repeat elements). Furthermore, observed alterations in the target sequences (see below) were used to eliminate contamination products from the data set. These analyses verified that the most frequently targeted sites indeed harbor different proviruses; however, the possibility that a few of the common insertions are results of contamination cannot be formally excluded. As an example, nine virus insertions were observed at position −1306, and of these, one was an MCF virus, two were SL3-3(wt) viruses, and the remainder were SL3-3(turbo) viruses. Seven of the proviruses were inserted in the antisense orientation. Furthermore, atypical junction structures were observed in seven cases (see below).

Alterations in duplicated host target sequences. Atypical virus-chromosomal junction structures were identified for 23 of the 164 insertions (Table 2). These proviral integrations were distributed at 15 different target sites. Among these, 11 were located at three insertion site hot spots: 7 at position −1306 upstream of exon 1, three at position −1199, and one at position −1112. Mutations in the duplicated sequence were detected for 20 different integrations, where either a nucleotide immediately adjacent to the virus, the second nearby nucleotide, or two nucleotides next to the virus were mutated. Alterations of the third and fourth positions from the virus were never observed. Moreover, in four cases the length of the duplicated target sequence was changed to 3 or 5 bp.

We had amplified both the 5′-LTR- and 3′-LTR-flanking region from 12 insertions harboring mutations in the direct repeats, and in all cases, the alteration was observed in one of the junctions only. From the eight remaining mutated repeat structures, either the 5′-LTR or the 3′-LTR-flanking region was amplified. A few single-nucleotide changes in the c-*myc* promoter sequence can be observed between different mouse strains (e.g., by comparison of the published c-*myc* promoter sequence M12345 [BALB/c mice] and the data set [NMRI mice]); therefore, provirus flanking sequences from the present data set overlapping with the insertion sites were analyzed to verify the target sequence.

Generally, the single point mutations introduced a V-to-T (V denotes A, C, or G) nucleotide alteration at the position juxtaposed to the virus or at the second nearby position; i.e., for viruses integrated in the antisense orientation, the nucleotide of the target sequence at position b1 or b2 (the nomenclature is shown in Fig. 2A) is altered from V to T when the mutation is detected at the 5′-LTR-flanking junction, and the

TABLE 1. Frequency of tumors harboring integrations into c-*myc* in different experimental series

Virus	NMRI mouse strain[a]	Mean latency (days) ± SD	Tumor type[b]	Gender[c]	Frequency of tumors harboring c-*myc* integrations[d]	No. of insertions amplified from each tumor series
SL3-3(wt)[e]	R	88 ± 25	T	F	5/20 (25%)	7
SL3-3(wt)[f]	I	54 ± 8	T	F	9/15 (60%)	15
SL3-3(wt)[g]	I	64 ± 8	T	F	7/18 (39%)	9
SL3-3(wt)[i]	I	56 ± 8	ND	M	37/39 (95%)	80
SL3-3(turbo)[g]	I	51 ± 5	T	F	11/12 (92%)	20
SL3-3(turbo)[i]	I	49 ± 6	ND	M	100/110 (91%)	259
SL3-3(2Δ18-2½)[g]	I	68 ± 10	T	F	4/12 (33%)	5
SL3-3(TUMdm)[h]	I	121 ± 32	T	F	5/17 (29%)	6
SL3-3(GTT)[e]	R	266 ± 146	T	F	3/8 (38%)	4

[a] R, randomly bred; I, inbred.
[b] T, T-cell lymphomas; ND, not determined.
[c] F, female mice; M, male mice.
[d] As determined by PCR.
[e] Hallberg et al. (32).
[f] Lund et al. (48).
[g] Ethelberg et al. (26).
[h] Ethelberg et al. (25).
[i] New round of injection.

nucleotide at position b1′ or b2′ is altered from V to T at the 3′-LTR-flanking junction. The opposite is seen for viruses integrated in the sense orientation. At all three junctions harboring a dinucleotide alteration, two identical nucleotides, CC or GG, were mutated. The mutations were observed equally in 5′-LTR-flanking and 3′-LTR-flanking repeat structures (45 and 55%, respectively). All together, the observed nucleotide changes suggest that defects in one of the steps in the integration process, perhaps due to an imperfect removal of the viral 5′ dinucleotide left over, introduce mutations in the stretch of 4 bp duplicated at the MLV integration.

A [5′-(A/C/G)TA(C/G/T)-3′] consensus sequence for SL3-3 MLV target sites. In total, the 164 insertions were distributed on 83 different integration sites. To clarify whether the selection of target sites correlates with a consensus sequence, the repeats at the integration sites displaying a typical MLV-generated duplication of 4 bp (Table 3) were compared. At all four positions, the percentage of each nucleotide was calculated and the results are presented in Tables 4 to 6. The calculations were based on nucleotides located at the first to fourth positions from the 5′ LTR of the virus and into the flanking DNA; i.e., for viruses in the sense orientation, the four nucleotides are 5′-b1′b2′b3′b4′-3′, and for viruses in the antisense orientation, the nucleotides are 5′-b1b2b3b4-3′ (the nomenclature is shown in Fig. 2A). As seen, a T nucleotide at the second position and an A or T nucleotide at the third position is favored for viruses in either orientation (Tables 4 and 5). Several integration sites are repeatedly targeted, and when we include this fact in the calculations (Table 6), the data suggest a weak consensus sequence [5′-(A/C/G)TA(C/G/T)-3′] for SL3-3 MLV target sequences. The apparent palindromic nature of the consensus sequence indicates that IN does not distinguish between 5′-LTR or 3′-LTR *att* elements at the strand transfer step.

Virus-dependent integration pattern in the c-*myc* promoter. The proviral integrations shown in Fig. 3 include 50 and 98 SL3-3(wt) and SL3-3(turbo) insertions, respectively. Both viruses are predominantly integrated into the c-*myc* promoter region in the antisense orientation, suggesting activation of this proto-oncogene by enhancer insertion. To analyze whether different insertion sites are selected for in end-stage tumors induced by these two viruses, the percentages of insertions obtained from each data set within windows of 30 bp were investigated (Fig. 4). The SL3-3(wt) insertions are located within a region of 1,400 bp, with 32% of the insertions occurring in a narrow region of 30 bp (positions −1397 to −1426) (Fig. 4B). In contrast, no such dense clusters of SL3-3(turbo) insertions are observed. When we look in detail at the distribution of virus insertions located in the 30-bp region spanning positions −1397 to −1426 upstream of exon 1, the difference in target site selection for SL3-3(wt) and SL3-3(turbo) appears at position −1408 (Fig. 4C). Based on the observation that 16 out of 50 SL3-3(wt) and 5 out of 98 SL3-3(turbo) insertions independently appear in a region of 30 bp, Fisher's exact test for count data results in a significant ($P = 0.00002$) difference in the binomial parameters for insertions with respect to the groups [SL3-3(wt) and SL3-3(turbo)]. All together, these results demonstrate a virus-specific pattern for integration sites in the c-*myc* promoter region.

Correlation between tumor latency and number of provirus integrations in the c-*myc* promoter region. To investigate the putative correlation between tumor latency and the number of provirus insertions in the c-*myc* promoter, the number of different fragments generated in the PCR analysis was counted. All in all, a total of 495 amplification products representing 405 different insertions were achieved from the approximately 250 tumors analyzed. Subsequently, every tumor analyzed was pictured according to latency period and number of insertions (Fig. 5A). Furthermore, the average number of proviral integrations in the c-*myc* promoter as a function of latency period to disease development was calculated. The total observation period of 300 days was divided into subgroups of 10 days each; however, due to the small amount of data samples for latency periods above 80 days, these have been grouped. Within each 10-day time period, the counted number of insertions in c-*myc* was assumed to follow a Poisson distribution. To test this assumption, the conditional chi-square statistic was calculated. The results showed that for all five subgroups, the
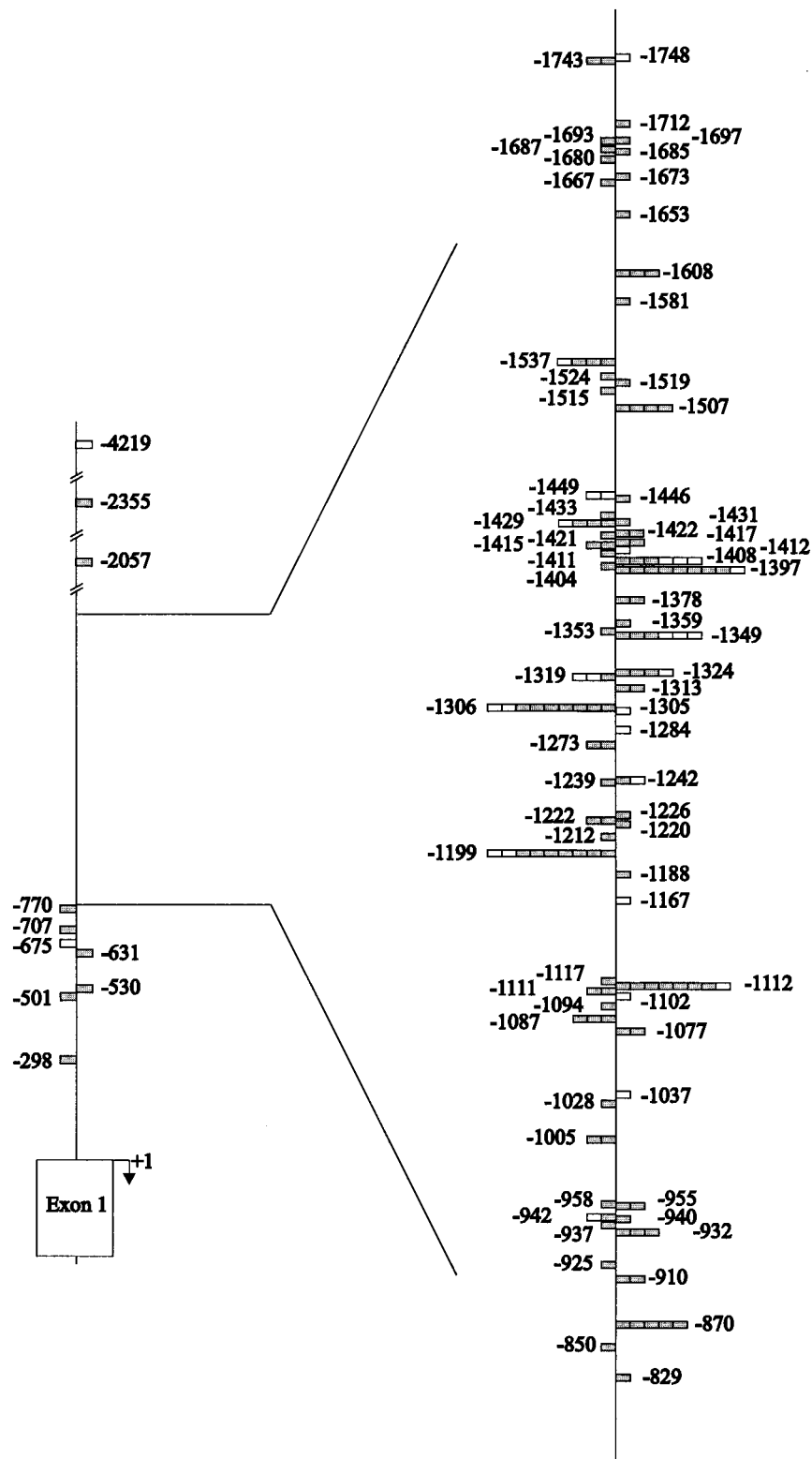
FIG. 3. Integration sites in the c-*myc* promoter region. Integration sites in the c-*myc* promoter obtained by gene-specific PCRs are shown relative to the promoter sequence with the GenBank and EMBL accession number NT_078782. The region from approximately −800 to −1,800 bp upstream of exon 1 is enlarged in order to illustrate the densely packed integration sites. Each proviral integration is indicated by a box, and the orientations are contrasted, with white indicating the sense direction and grey indicating the antisense direction.

TABLE 2. A typical virus-chromosome junction structures

Virus	Mouse strain (gender)[b]	Insertion site[c]	Virus orientation[d]	Target sequence[e]	5′-LTR-flanking junction[f]	3′-LTR-flanking junction[f]
SL3-3(turbo)	I (M)	−1537	Sense	GGAT	ND	**TT**AT
SL3-3(turbo)	I (F)	−1431	Antisense	ATAT	**TT**AT	ND
SL3-3(2Δ18-2½)	I (F)	−1422	Antisense	GTGT	**TT**GT	ND
SL3-3(wt)	I (F)	−1408	Antisense	GTTT	**TTT**T	ND
SL3-3(turbo)[a]	I (M)	−1359	Antisense	CTGG	CTGG	CT**AA**
SL3-3(turbo)	I (M)	−1306	Antisense	ATAC	ATAC	ATA**A**
SL3-3(turbo)	I (M)	−1306	Antisense	ATAC	ND	AT**GC**
SL3-3(turbo)	I (M)	−1306	Antisense	ATAC	**TT**AC	ATAC
SL3-3(turbo)	I (M)	−1306	Antisense	ATAC	**TT**AC	ND
SL3-3(wt)	R (F)	−1306	Antisense	ATAC	ATAC	ATA**A**
SL3-3(turbo)	I (M)	−1306	Sense	ATAC	ND	**TT**AC
SL3-3(turbo)	I (M)	−1305	Sense	TAC[g]	TAC	**TT**C
SL3-3(turbo)	I (M)	−1242	Sense	TAAT	ND	**TT**AT
SL3-3(turbo)	I (M)	−1199	Antisense	CTGG	**TT**GG	CTGG
SL3-3(wt)	I (M)	−1199	Antisense	CTGG	CTGG	CT**AA**
SL3-3(wt)	R (F)	−1199	Sense	CTGG	CTGG	**TT**GG
SL3-3(turbo)	I (M)	−1188	Antisense	GGGTT[h]	GGGTT	GGGTT
SL3-3(turbo)	I (M)	−1167	Sense	CTAG	CT**AA**	CTAG
SL3-3(wt)[a]	I (M)	−1112	Antisense	CTGG	**TT**GG	CTGG
SL3-3(2Δ18-2½)	I (F)	−1111	Antisense	TGGG	**TT**GG	TGGG
SL3-3(turbo)	I (F)	−1087	Antisense	CCCCC[h]	CCCCC	CCCCC
SL3-3(turbo)	I (F)	−958	Antisense	GCCCT[h]	GCCCT	GCCCT

[a] The virus harbors a recombined U5 region.

[b] Randomly bred and inbred NMRI mice are marked with R and I, respectively. The gender of the mouse is given in parentheses, with F indicating female and M indicating male.

[c] The number indicates the upstream position of the integration site relative to the transcription start site of the c-*myc* promoter, P1.

[d] The orientation is relative to the transcriptional orientation of c-*myc*.

[e] The wild-type target sequence is shown as 5′-b1b2b3b4-3′ (Fig. 2A).

[f] To help the reader, the duplicated stretch of host DNA is presented as 5′-b1b2b3b4-3′ for both the 5′-LTR- and 3′-LTR-flanking junctions. Mutated nucleotides are marked in bold. ND, not determined.

[g] A stretch of 3 bp was duplicated at this proviral integration site.

[h] Five base pairs were duplicated at this target site.

Poisson assumption was retained (P values were 0.695, 0.227, 0.578, 0.143, and 0.419).

The average number of insertions in c-*myc* per tumor within a given subgroup of latency periods is an estimator for the parameter λ_i of the Poisson distribution in subgroup *i*. Figure 5B shows these estimators in common with 95% confidence intervals based on the gamma distribution (see Materials and Methods). There is a strong correlation between the average number of insertions and the grouped latency periods. Comparing two adjacent latency periods showed significant differences (binomial test P values were 0.031, 0.006, 0.246, and 0.043). In view of these stringent conditions, there was no significant difference in Poisson parameters between the latency period from 60 to 69 days and the latency period from 70 to 79 days. However, this may be due to the relatively small number of samples in these groups.

DISCUSSION

In mice, both MLVs and the betaretrovirus TBLV target c-*myc* during T-cell tumor induction with a genus-specific difference in the distributions of insertion sites. Integration sites of MLVs such as Mo-MLV and MCF 69L1 in both nontransgenic and transgenic mice are predominantly observed in clusters within a 3-kb region upstream of c-*myc*, and few insertions are located within the first exon and intron (2, 3, 15, 34, 43, 54, 64, 67). The major parts of the viruses are integrated in the transcriptionally opposite orientation relative to that of the proto-oncogene. In contrast, TBLV insertions are detected upstream, downstream, and within the c-*myc* gene, and insertions in either orientation are observed with apparently similar frequencies (6). Like other MLVs, the presented proviral insertion sites of SL3-3(wt) and various enhancer variants of SL3-3 were primarily inserted in the antisense orientation within a region spanning approximately 2.4 kb upstream of c-*myc*. Several integration site hot spots with up to nine different proviral integrations were identified. By sequencing the enhancer regions of the integrated viruses, it was possible to eliminate contamination products from the PCRs, thus verifying that these hot spots truly represent highly frequently targeted positions. The first indication that specific nucleotide positions may be strongly preferred at retroviral integration was reported in 1988 by Shih et al. (66), but these results were later questioned in a report from the same laboratory (75). Also, in vitro analyses of MLV integration into minichromosomes have shown that specific nucleotide positions may in some instances be preferred (57). More-recent large-scale screenings of proviral integration sites (33, 37, 43, 46, 49, 50, 65, 67, 70) have detected insertions sited at the exactly same base as listed in the RTCG database (http://RTCGD.ncifcrf.gov) (2).

The integration sites are not uniformly distributed along the promoter sequence (Fig. 3), and the insertion pattern both reflects target site selection by IN and represents sites selected for during cancer development. Several of the targeted positions are separated by a few base pairs only, and we believe that this is due to intrinsic features of IN regarding target site

TABLE 3. Sequences of the base pair repeats identified at the 83 different integration sites

Insertion site[a]	4-bp b1b2b3b4[b]	No. of insertions[c]	Insertion site	4-bp b1b2b3b4	No. of insertions	Insertion site	4-bp b1b2b3b4	No. of insertions
−4219	TAGT	1	−1417	TATG	2	−1112	CTGG	8
−2355	CTTT	1	−1415	TGTA	2	−1111	TGGG	2
−2057	GTCT	1	−1412	ATAC	1	−1102	CGGG	1
−1748	CTAT	1	−1411	TACG	1	−1094	CCCC	1
−1743	ATTC	2	−1408	GTTT	6	−1087	CCCC[e]	3
−1712	AGTG	1	−1404	GGGG	1	−1077	ATGG	2
−1697	CTTT	1	−1397	GTAC	9	−1037	CCAG	1
−1693	GAAC	1	−1378	GTAT	2	−1028	AAAC	1
−1687	TGAG	1	−1359	CTGG	1	−1005	GCTC	2
−1685	AGCG	1	−1353	AATT	1	−958	GCCCT[d]	1
−1680	CTTA	1	−1349	AATG	6	−955	CTGG	2
−1673	ACGC	1	−1324	AATT	4	−942	GAGG	2
−1667	CTGT	1	−1319	GCTT	3	−940	GGTG	1
−1653	CTTT	1	−1313	CACA	2	−937	GTAT	1
−1608	AGAG	3	−1306	ATAC	9	−932	GGGT	3
−1581	CGAG	1	−1305	TAC[d]	1	−925	GACC	1
−1537	GGAT	4	−1284	TTAC	1	−910	CTGG	2
−1524	CCGA	1	−1273	TGAT	2	−870	ACTG	5
−1519	GGAA	1	−1242	TAAT	2	−850	CTAC	1
−1515	TATG	1	−1239	TTAC	1	−829	CCAC	1
−1507	ATAT	4	−1226	AATA	1	−770	GGGG	1
−1449	TAAG	2	−1222	AAAG	2	−707	AGTC	1
−1446	GCTA	1	−1220	AGGG	1	−675	ACAC	1
−1433	GTAT	1	−1212	GCTT	1	−631	GTGG	1
−1431	ATAT	1	−1199	CTGG	9	−530	CTAT	1
−1429	ATAT	4	−1188	GGGGT[d]	1	−501	CTAG	1
−1422	GTGT	2	−1167	CTAG	1	−298	GTAC	1
−1421	TGTG	1	−1117	CCTT	1			

[a] The insertion site indicates the upstream position in the c-*myc* promoter region relative to the transcription start site of promoter P1 according to the sequence of GenBank and EMBL accession number NT_078782. Flanking sequences are available upon request.

[b] The four base pairs listed indicate positions b1, b2, b3, and b4 relative to the integration site (Fig. 2A).

[c] The total numbers of proviral integrations identified at these insertion sites are indicated.

[d] Atypical duplication of either 3 or 5 bp was observed at this integration site.

[e] One of the integrations sited at this position is correlated with the 5-bp repeat 5′-CCCCC-3′, and the two other insertions are correlated with a 5′-CCCC-3′ 4-bp repeat.

selection rather than divergences in the potential of the integrated proviruses to deregulate the expression of c-*myc*.

Single point mutations in the direct repeat element were observed at either the 5′-LTR- or 3′-LTR-flanking junction for 17 of the integration sites (Table 2). In thirteen of the mutated direct repeats, the nucleotide immediately adjacent to the virus was altered, and in four integration events, the second nearby nucleotide was changed. In all cases except for one, a V-to-T mutation was introduced. Based on this observation, we propose the model for error-prone gap repair illustrated in Fig. 6. By this model, mutations are introduced when the repeated target sequence harbors a T nucleotide at either the first or second position from the virus that is able to base pair with one of the two A nucleotides in the viral 5′ overhang, thereby generating a region of microhomology. The DNA polymerase involved in gap repair fills in the dinucleotide gap, and subsequent editing of the mismatching base introduces the mutation.

For three proviral integrations, two identical bases at the 3′-LTR-flanking virus-chromosome junction were changed relative to the target sequence. In neither case did the target sequence provide a T nucleotide that could base pair with either position 1 or position 2 of the viral 5′ dinucleotide left over; hence, these mutations cannot be explained by the model proposed above. Likewise, analysis of the virus-host DNA junctions demonstrated no insertion of additional nucleotides; thus, the alterations are not due to imperfect 3′ processing prior to strand transfer. At present, it is unknown how these mutations are generated.

TABLE 4. Distribution of each nucleotide at the four positions in 21 target sequences from viruses in the sense orientation[a]

Base	No. (%) of sequences with indicated base at position:			
	b1′	b2′	b3′	b4′
A	8 (38)	4 (19)	10 (48)	4 (19)
C	8 (38)	5 (24)	2 (10)	5 (24)
G	5 (24)	0	3 (14)	6 (29)
T	0	12 (57)	6 (29)	6 (29)

[a] The single target site that correlated with a 3-bp duplication only is not included. Calculations are based on wild-type target sequences only. The consensus sequences for b1′, b2′, b3′, and b4′ were A/C/G, T, A, and N, respectively.

TABLE 5. Distribution of each nucleotide at the four positions in 72 target sequences from viruses in the antisense orientation[a]

Base	No. (%) of sequences with indicated base at position:			
	b1	b2	b3	b4
A	19 (27)	14 (20)	24 (34)	8 (11)
C	20 (28)	11 (15)	8 (11)	16 (23)
G	22 (31)	17 (24)	16 (23)	24 (34)
T	10 (14)	29 (41)	23 (32)	23 (32)

[a] Target sites that correlated with a 5-bp duplication only are not included. That is, target site −1087 displays two insertions with 4-bp duplications; therefore, this insertion site is included. Calculations are based on wild-type target sequences only. The consensus sequences for b1, b2, b3, and b4 were A/C/G, T, A/T, and C/G/T, respectively.

TABLE 6. Distribution of each nucleotide as calculated for all of the 160 proviral insertions with 4-bp duplications[a]

Base	No. (%) of sequences with indicated base at position:			
	b1/b1′	b2/b2′	b3/b3′	b4/b4′
A	54 (34)	30 (19)	63 (39)	15 (9)
C	48 (30)	23 (14)	13 (8)	41 (26)
G	44 (28)	26 (16)	38 (24)	56 (35)
T	14 (9)	81 (51)	46 (29)	48 (30)

[a] Insertion sites that correlated with atypical lengths of duplicated host DNA are not included; however, the two insertions at −1087 with 4-bp duplications are included. The calculations are based on wild-type target sequences only. The consensus sequences for b1/b1′, b2/b2′, b3/b3′, and b4/b4′ were A/C/G, T, A, and C/G/T, respectively.

The gene-specific PCRs were performed with a regular *Taq* polymerase without proofreading capacity (see Materials and Methods). However, as the mutations are position specific (never observed at the third or fourth position from the virus) and nucleotide specific (V to T) and only a very few differences were detected during comparison with sequences in publicly available databases, we can rule out the possibility that the observed base changes resulted from errors generated during the elongation step.

The gap repair system involved at proviral integration still remains to be determined. Based on studies with cell culture, both DNA-dependent protein kinase participating in the repair of double-stranded DNA breaks and the poly(ADP-ribose) polymerase-1 (PARP-1) implicated in the repair of single-stranded as well as double-stranded breaks have been proposed to be involved (16–18, 29, 31). However, in both cases, investigators have obtained deviating results (4, 13, 21, 28, 47). In addition, both viral and host polymerases have been shown to repair gapped DNA substrates in concert with flap endonuclease and host ligases in vitro (77). The point mutations introduced at error-prone gap repair may be generated due to the mismatch repair system responsible for removal of base mismatches introduced during replication or by other processes (reviewed in the work of Christmann et al. [12]). The mechanism of strand discrimination in eukaryotes is still not clear, and mismatch repair may correct the base present at either strand of the mismatching base pair, leading to introduction of point mutations in some cases only. Alternatively, DNA replication may have taken place prior to repair of the mismatching base pair, thereby generating daughter cells harboring different junction structures.

The 17 single-nucleotide mutations generated by error-prone gap repair were detected at 10 different integration sites dispersed regularly along the *c-myc* promoter region, and 9 of the cases were located at three target site hot spots (at positions −1306, −1199, and −1112). At position −1306, error-prone gap repair was detected at seven out of nine insertion events. The target sequence at this hot spot is 5′-ATAC-3′ (Table 3), which provides the opportunity of base pairing between target sequence and the viral 5′ overhang at either junction (Fig. 6), a fact that may partly explain this high frequency. In contrast, none of the nine proviral integrations located at the integration site hot spot at position −1397 also harboring an optimal target sequence for error-prone gap repair (5′-GTAC-3′) (Table 3) were correlated with point mu-

tations in the junction structures. This finding raises the possibility that the location of the insertion site may play a role in the frequency of error-prone gap repair.

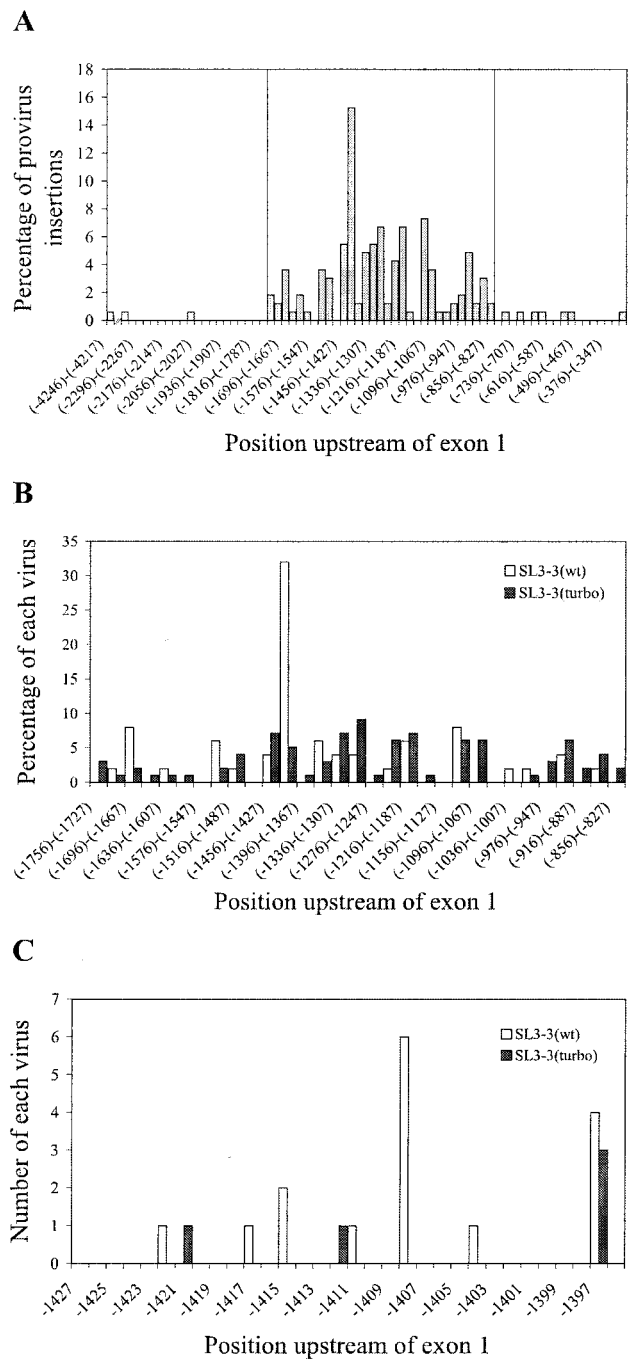Atypical lengths of the duplicated repeats at virus-host DNA

FIG. 4. Virus-dependent integration pattern in the *c-myc* promoter region. (A) The promoter region is divided into sections of 30 bp, and within each section the percentages of all virus insertions are calculated. (B) Similarly, the percentages of SL3-3(wt) and SL3-3(turbo) insertions within 30-bp windows from positions −827 to −1756 are shown (this region is indicated by lines in panel A; i.e., a total of 10 insertions outside this region are not included). (C) The numbers of SL3-3(wt) and SL3-3(turbo) insertions at individual positions from positions −1396 to −1427 are illustrated.
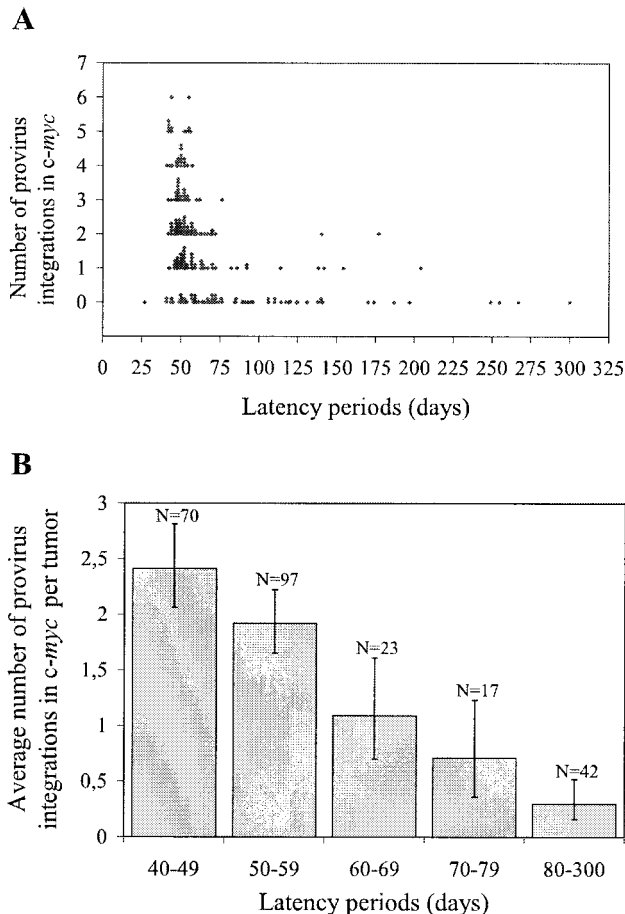
A

B

FIG. 5. Correlation between tumor latency and number of proviral integrations in the c-*myc* promoter region. (A) The number of provirus insertions in the individual tumors is shown as a function of the latency period. Identical data points (latency period and number of integrations) are shown above each other. The numbers of different integrations in c-*myc* in the individual tumors are obtained by counting the numbers of amplification products achieved in the repeated rounds of PCRs. (B) The average number of proviral integrations in the c-*myc* gene as a function of latency period to disease development is shown along with 95% confidence intervals. The total observation period of 300 days has been divided into subgroups of 10 days each; however, due to the small amount of data samples for each latency period above 80 days, these latency periods have been grouped. N, number of mice within a given subgroup.

junctions have been reported previously (14, 30, 39, 52, 71, 72). Likewise, one case of point mutations similar to those described in this report was identified at proviral integration sites by a Mo-MLV-based vector in fibroblasts by Jin et al. (39). Of the 38 integration sites analyzed by Jin et al. (39), 6 were correlated with a duplication of a 5-bp repeat, 1 displayed single-nucleotide point mutations, and 1 was correlated with another type of aberrant junction structure. Thus, an atypical length of repeats is the foremost alteration in the study by Jin et al. (39), while point mutations predominate in our data set. Differences in the cell types, chromosomal regions affected, and integrating mutagens, i.e., virus versus vector, may explain this divergence. Differences in the IN enzyme among MLVs might also play a role. Our data stem from T cells, in contrast to the fibroblasts used with the Mo-MLV-based vector. Non-

homologous end joining plays a central role in T cells due to the involvement in rearrangements of the T-cell receptor, a fact that may relate to the higher frequency of error-prone gap repair in our data set. Moreover, these cells are tumor cells in which gap repair systems may be impaired. In addition, the importance of the organization of chromatin in different cell types cannot be excluded.

The different target sites conformed to a weak consensus sequence [5'-(A/C/G)TA(C/G/T)-3'], in which a T and an A nucleotide at the second and third positions, respectively, are favored (Table 6). Previously, the weak consensus sequence [5'-GT(A/T)AC-3'] was identified for HIV-1 integration (10, 69). As with our observation, a bias for a T nucleotide and an A nucleotide at the second and fourth positions, respectively, was observed (10). Likewise, biases for A's and T's at the central positions of the target sequence have been reported for
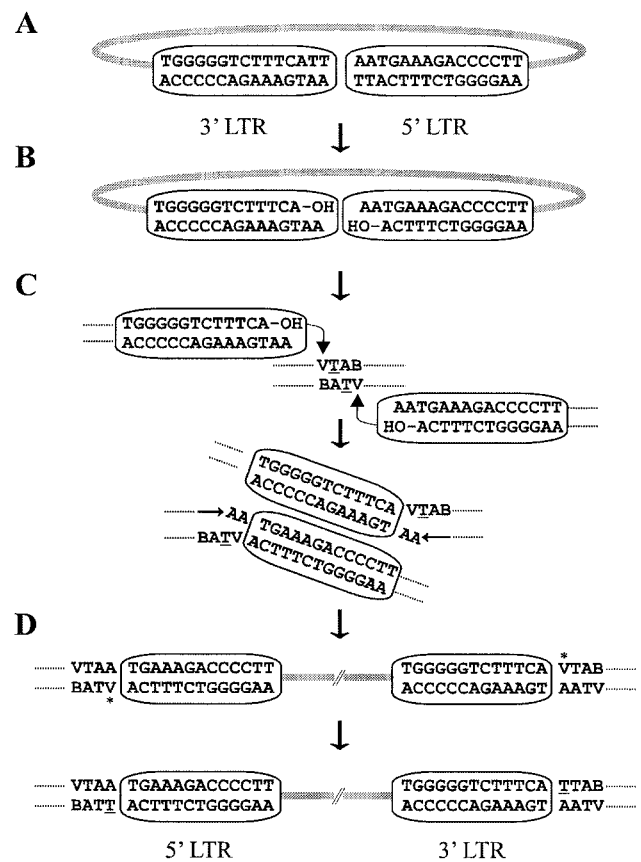

FIG. 6. Model for error-prone gap repair at proviral integration. The figure illustrates the introduction of mutations at either virus-chromosome junction. (A) Schematic representation of the SL3-3 proviral DNA copy of the genome. The 3'-LTR and 5'-LTR *att* sites are shown. (B) The integration step is initiated by 3' processing of two T nucleotides. (C) During strand transfer, the two DNA ends are joined in concert to the chromosomal target DNA 4 bp apart. The target sequence VTAB (V denotes A, C, or G; B denotes C, G, or T) is shown, and the T nucleotide located at the second position is underlined. Due to base pairing, the 5' dinucleotide overhangs are not removed during the gap repair process. (D) Subsequent editing of the mismatching base pairs (marked with an asterisk) generates the V-to-T mutation at either junction. The figure illustrates the introduction of mutations at both junctions for the same proviral integration; however, it is noted that this is never observed in our data set.

Mo-MLV integration in vitro and in cell culture (39, 57) and for human T-lymphotropic virus type 1 insertions isolated from patient samples (11, 45). Based on these observations, both SL3-3 and HIV-1 INs target rotationally symmetric sequences, and both enzymes prefer a combination of T and A at the central positions of the targeted stretch of host DNA.

As seen in Fig. 5, a clear correlation between tumor latency and the number of insertions in the c-*myc* promoter was observed. In several tumors, more than two insertions in c-*myc* were identified (Fig. 5A). The clonality status of the insertions remains to be analyzed by Southern blotting analysis. Yet, the vast majority of the insertions may be polyclonally integrated, as demonstrated by Southern blotting analysis for the 20 insertions detected in 12 SL3-3(turbo)-induced lymphomas in female mice (Nielsen et al., unpublished). Similarly, tumors displaying single or multiple polyclonal TBLV insertions in c-*myc* were observed in the study by Broussard et al. (6). The lymphomas induced by wild-type and enhancer variants of SL3-3 may be generated from numerous clones in which the c-*myc* gene is targeted with similar frequencies, or the tumors may be made up of a smaller number of clones, in which the frequency of insertions in c-*myc* varies. To address this question, gene-specific PCRs have to be performed on DNA isolated from single cells.

In conclusion, the presented study demonstrates (i) a [5′-(A/C/G)TA(C/G/T)-3′] integration target site consensus sequence; (ii) nonuniform dispersion of proviral integration sites in the c-*myc* promoter, including repeated targeting of specific nucleotide positions; (iii) dependence of tumorigenic target site selection upon the integrating virus; (iv) correlation of faster latency periods with higher average numbers of c-*myc* integrations per tumor; and (v) frequent mutation of the provirus-flanking repeat structures upon error-prone gap repair.

ACKNOWLEDGMENTS

REFERENCES

1. **Agresti, A.** 1990. Categorical data analysis. Wiley, New York, N.Y.
2. **Akagi, K., T. Suzuki, R. M. Stephens, N. A. Jenkins, and N. G. Copeland.** 2004. RTCGD: retroviral tagged cancer gene database. Nucleic Acids Res. **32:**D523–D527.
3. **Amtoft, H. W., A. B. Sørensen, C. Bareil, J. Schmidt, A. Luz, and F. S. Pedersen.** 1997. Stability of AML1 (core) site enhancer mutations in T lymphomas induced by attenuated SL3-3 murine leukemia virus mutants. J. Virol. **71:**5080–5087.
4. **Baekelandt, V., A. Claeys, P. Cherepanov, E. De Clercq, B. De Strooper, B. Nuttin, and Z. Debyser.** 2000. DNA-dependent protein kinase is not required for efficient lentivirus integration. J. Virol. **74:**11278–11285.
5. **Ball, J. K., H. Diggelmann, G. A. Dekaban, G. F. Grossi, R. Semmler, P. A. Waight, and R. F. Fletcher.** 1988. Alterations in the U₃ region of the long terminal repeat of an infectious thymotropic type B retrovirus. J. Virol. **62:**2985–2993.
6. **Broussard, D. R., J. A. Mertz, M. Lozano, and J. P. Dudley.** 2002. Selection for c-*myc* integration sites in polyclonal T-cell lymphomas. J. Virol. **76:**2087–2099.
7. **Brown, L. D., and L. H. Zhao.** 2002. A test for the Poisson distribution. Sankhya Ser. A **64:**611–625.
8. **Bushman, F. D.** 2002. Integration site selection by lentiviruses: biology and possible control. Curr. Top. Microbiol. Immunol. **261:**165–177.
9. **Bushman, F. D.** 2003. Targeting survival: integration site selection by retroviruses and LTR-retrotransposons. Cell **115:**135–138.
10. **Carteau, S., C. Hoffmann, and F. Bushman.** 1998. Chromosome structure and human immunodeficiency virus type 1 cDNA integration: centromeric alphoid repeats are a disfavored target. J. Virol. **72:**4005–4014.
11. **Chou, K. S., A. Okayama, I. J. Su, T. H. Lee, and M. Essex.** 1996. Preferred nucleotide sequence at the integration target site of human T-cell leukemia virus type I from patients with adult T-cell leukemia. Int. J. Cancer **65:**20–24.
12. **Christmann, M., M. T. Tomicic, W. P. Roos, and B. Kaina.** 2003. Mechanisms of human DNA repair: an update. Toxicology **193:**3–34.
13. **Coffin, J. M., and N. Rosenberg.** 1999. Retroviruses. Closing the joint. Nature **399:**413
14. **Colicelli, J., and S. P. Goff.** 1985. Mutants and pseudorevertants of Moloney murine leukemia virus with alterations at the integration site. Cell **42:**573–580.
15. **Corcoran, L. M., J. M. Adams, A. R. Dunn, and S. Cory.** 1984. Murine T lymphomas in which the cellular myc oncogene has been activated by retroviral insertion. Cell **37:**113–122.
16. **Daniel, R., R. A. Katz, G. Merkel, J. C. Hittle, T. J. Yen, and A. M. Skalka.** 2001. Wortmannin potentiates integrase-mediated killing of lymphocytes and reduces the efficiency of stable transduction by retroviruses. Mol. Cell. Biol. **21:**1164–1172.
17. **Daniel, R., R. A. Katz, and A. M. Skalka.** 1999. A role for DNA-PK in retroviral DNA integration. Science **284:**644–647.
18. **Daniel, R., S. Litwin, R. A. Katz, and A. M. Skalka.** 2001. Computational analysis of retrovirus-induced scid cell death. J. Virol. **75:**3121–3128.
19. **Denicourt, C., E. Edouard, and E. Rassart.** 1999. Oncogene activation in myeloid leukemias by Graffi murine leukemia virus proviral integration. J. Virol. **73:**4439–4442.
20. **Dudley, J. P., J. A. Mertz, L. Rajan, M. Lozano, and D. R. Broussard.** 2002. What retroviruses teach us about the involvement of c-Myc in leukemias and lymphomas. Leukemia **16:**1086–1098.
21. **Engelman, A.** 2003. The roles of cellular factors in retroviral integration. Curr. Top. Microbiol. Immunol. **281:**209–238.
22. **Erkeland, S. J., M. Valkhof, C. Heijmans-Antonissen, A. Van Hoven-Beijen, R. Delwel, M. H. Hermans, and I. P. Touw.** 2004. Large-scale identification of disease genes involved in acute myeloid leukemia. J. Virol. **78:**1971–1980.
23. **Ethelberg, S.** 1997. Transcriptional control and pathogenic properties of the lymphomagenic murine retrovirus SL3-3 investigated via enhancer variants selected for during tumorigenesis. Ph.D. thesis. University of Aarhus, Aarhus, Denmark.
24. **Ethelberg, S., B. Hallberg, J. Lovmand, J. Schmidt, A. Luz, T. Grundström, and F. S. Pedersen.** 1997. Second-site proviral enhancer alterations in lymphomas induced by enhancer mutants of SL3-3 murine leukemia virus: negative effect of nuclear factor 1 binding site. J. Virol. **71:**1196–1206.
25. **Ethelberg, S., J. Lovmand, J. Schmidt, A. Luz, and F. S. Pedersen.** 1997. Increased lymphomagenicity and restored disease specificity of AML1 site (core) mutant SL3-3 murine leukemia virus by a second-site enhancer variant evolved in vivo. J. Virol. **71:**7273–7280.
26. **Ethelberg, S., A. B. Sørensen, J. Schmidt, A. Luz, and F. S. Pedersen.** 1997. An SL3-3 murine leukemia virus enhancer variant more pathogenic than the wild type obtained by assisted molecular evolution in vivo. J. Virol. **71:**9796–9799.
27. **Evans, M., N. A. J. Hastings, and B. Peacock.** 1993. Statistical distributions, 2nd ed. Wiley, New York, N.Y.
28. **Fulop, G. M., G. C. Bosma, M. J. Bosma, and R. A. Phillips.** 1988. Early B-cell precursors in *scid* mice: normal numbers of cells transformable with Abelson murine leukemia virus (A-MuLV). Cell. Immunol. **113:**192–201.
29. **Gäken, J. A., M. Tavassoli, S. U. Gan, S. Vallian, I. Giddings, D. C. Darling, J. Galea-Lauri, M. G. Thomas, H. Abedi, V. Schreiber, J. Menissier-de Murcia, M. K. Collins, S. Shall, and F. Farzaneh.** 1996. Efficient retroviral infection of mammalian cells is blocked by inhibition of poly(ADP-ribose) polymerase activity. J. Virol. **70:**3992–4000.
30. **Gaur, M., and A. D. Leavitt.** 1998. Mutations in the human immunodeficiency virus type 1 integrase D,D(35)E motif do not eliminate provirus formation. J. Virol. **72:**4678–4685.
31. **Ha, H. C., K. Juluri, Y. Zhou, S. Leung, M. Hermankova, and S. H. Snyder.** 2001. Poly(ADP-ribose) polymerase-1 is required for efficient HIV-1 integration. Proc. Natl. Acad. Sci USA **98:**3364–3368.
32. **Hallberg, B., J. Schmidt, A. Luz, F. S. Pedersen, and T. Grundström.** 1991. SL3-3 enhancer factor 1 transcriptional activators are required for tumor formation by SL3-3 murine leukemia virus. J. Virol. **65:**4177–4181.
33. **Hansen, G. M., D. Skapura, and M. J. Justice.** 2000. Genetic profile of insertion mutations in mouse leukemias and lymphomas. Genome Res. **10:**237–243.
34. **Haupt, Y., A. W. Harris, and J. M. Adams.** 1992. Retroviral infection accelerates T lymphomagenesis in Eμ-N-*ras* transgenic mice by activating c-*myc* or N-*myc*. Oncogene **7:**981–986.
35. **Hays, E. F., G. Bristol, and S. McDougall.** 1990. Mechanisms of thymic lymphomagenesis by the retrovirus SL3-3. Cancer Res. **50:**5631S–5635S.
36. **Holmes-Son, M. L., R. S. Appa, and S. A. Chow.** 2001. Molecular genetics and target site specificity of retroviral integration. Adv. Genet. **43:**33–69.
37. **Hwang, H. C., C. P. Martins, Y. Bronkhorst, E. Randel, A. Berns, M. Fero, and B. E. Clurman.** 2002. Identification of oncogenes collaborating with

p27[Kip1] loss by insertional mutagenesis and high-throughput insertion site analysis. Proc. Natl. Acad. Sci. USA **99:**11293–11298.

38. **Ihaka, R., and R. Gentleman.** 1996. R-a language for data analysis and graphics. J. Comput. Graph. Statist. **5:**299–314.

39. **Jin, Y. F., T. Ishibashi, A. Nomoto, and M. Masuda.** 2002. Isolation and analysis of retroviral integration targets by solo long terminal repeat inverse PCR. J. Virol. **76:**5540–5547.

40. **Johnson, N. L., S. Kotz, and A. W. Kemp.** 1993. Univariate discrete distributions, 2nd ed. Wiley, New York, N.Y.

41. **Jolicoeur, N. R. A. P.** 1997. Retroviral pathogenesis, p. 475–585. *In* J. M. Coffin, S. H. Hughes, and H. E. Varmus (ed.), Retroviruses. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

42. **Joosten, M., Y. Vankan-Berkhoudt, M. Tas, M. Lunghi, Y. Jenniskens, E. Parganas, P. J. Valk, B. Löwenberg, E. van den Akker, and R. Delwel.** 2002. Large-scale identification of novel potential disease loci in mouse leukemia applying an improved strategy for cloning common virus integration sites. Oncogene **21:**7247–7255.

43. **Kim, R., A. Trubetskoy, T. Suzuki, N. A. Jenkins, N. G. Copeland, and J. Lenz.** 2003. Genome-based identification of cancer genes by proviral tagging in mouse retrovirus-induced T-cell lymphomas. J. Virol. **77:**2056–2062.

44. **Kitamura, Y., Y. M. Lee, and J. M. Coffin.** 1992. Nonrandom integration of retroviral DNA *in vitro*: effect of CpG methylation. Proc. Natl. Acad. Sci. USA **89:**5532–5536.

45. **Leclercq, I., F. Mortreux, M. Cavrois, A. Leroy, A. Gessain, S. Wain-Hobson, and E. Wattel.** 2000. Host sequences flanking the human T-cell leukemia virus type 1 provirus in vivo. J. Virol. **74:**2305–2312.

46. **Li, J., H. Shen, K. L. Himmel, A. J. Dupuy, D. A. Largaespada, T. Nakamura, J. D. Shaughnessy, Jr., N. A. Jenkins, and N. G. Copeland.** 1999. Leukaemia disease genes: large-scale cloning and pathway predictions. Nat. Genet. **23:**348–353.

47. **Li, L., J. M. Olvera, K. E. Yoder, R. S. Mitchell, S. L. Butler, M. Lieber, S. L. Martin, and F. D. Bushman.** 2001. Role of the non-homologous DNA end joining pathway in the early steps of retroviral infection. EMBO J. **20:**3272–3281.

48. **Lund, A. H., J. Schmidt, A. Luz, A. B. Sørensen, M. Duch, and F. S. Pedersen.** 1999. Replication and pathogenicity of primer binding site mutants of SL3-3 murine leukemia viruses. J. Virol. **73:**6117–6122.

49. **Lund, A. H., G. Turner, A. Trubetskoy, E. Verhoeven, E. Wientjens, D. Hulsman, R. Russell, R. A. DePinho, J. Lenz, and M. van Lohuizen.** 2002. Genome-wide retroviral insertional tagging of genes involved in cancer in Cdkn2a-deficient mice. Nat. Genet. **32:**160–165.

50. **Mikkers, H., J. Allen, P. Knipscheer, L. Romeijn, A. Hart, E. Vink, A. Berns, and L. Romeyn.** 2002. High-throughput retroviral tagging to identify components of specific signaling pathways in cancer. Nat. Genet. **32:**153–159.

51. **Mikkers, H., and A. Berns.** 2003. Retroviral insertional mutagenesis: tagging cancer pathways. Adv. Cancer Res. **88:**53–99.

52. **Moreau, K., C. Torne-Celer, C. Faure, G. Verdier, and C. Ronfort.** 2000. *In vivo* retroviral integration: fidelity to size of the host DNA duplication might be reduced when integration occurs near sequences homologous to LTR ends. Virology **278:**133–136.

53. **Morrison, H. L., B. Soni, and J. Lenz.** 1995. Long terminal repeat enhancer core sequences in proviruses adjacent to c-*myc* in T-cell lymphomas induced by a murine retrovirus. J. Virol. **69:**446–455.

54. **O'Donnell, P. V., E. Fleissner, H. Lonial, C. F. Koehne, and A. Reicin.** 1985. Early clonality and high-frequency proviral integration into the c-*myc* locus in AKR leukemias. J. Virol. **55:**500–503.

55. **Pruss, D., F. D. Bushman, and A. P. Wolffe.** 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.

56. **Pruss, D., R. Reeves, F. D. Bushman, and A. P. Wolffe.** 1994. The influence of DNA and nucleosome structure on integration events directed by HIV integrase. J. Biol. Chem. **269:**25031–25041.

57. **Pryciak, P. M., A. Sil, and H. E. Varmus.** 1992. Retroviral integration into minichromosomes *in vitro*. EMBO J. **11:**291–303.

58. **Pryciak, P. M., and H. E. Varmus.** 1992. Nucleosomes, DNA-binding proteins, and DNA sequence modulate retroviral integration target site selection. Cell **69:**769–780.

59. **Rajan, L., D. Broussard, M. Lozano, C. G. Lee, C. A. Kozak, and J. P. Dudley.** 2000. The c-*myc* locus is a common integration site in type B retrovirus-induced T-cell lymphomas. J. Virol. **74:**2466–2471.

60. **Rohdewohld, H., H. Weiher, W. Reik, R. Jaenisch, and M. Breindl.** 1987. Retrovirus integration and chromatin structure: Moloney murine leukemia proviral integration sites map near DNase I-hypersensitive sites. J. Virol. **61:**336–343.

61. **Schmidt, J., V. Erfle, F. S. Pedersen, H. Rohmer, H. Schetters, K. H. Marquart, and A. Luz.** 1984. Oncogenic retrovirus from spontaneous murine osteomas. I. Isolation and biological characterization. J. Gen. Virol. **65:**2237–2248.

62. **Schmidt, J., A. Luz, and V. Erfle.** 1988. Endogenous murine leukemia viruses: frequency of radiation-activation and novel pathogenic effects of viral isolates. Leukoc. Res. **12:**393–403.

63. **Schröder, A. R., P. Shinn, H. Chen, C. Berry, J. R. Ecker, and F. Bushman.** 2002. HIV-1 integration in the human genome favors active genes and local hotspots. Cell **110:**521–529.

64. **Selten, G., H. T. Cuypers, M. Zijlstra, C. Melief, and A. Berns.** 1984. Involvement of c-myc in MuLV-induced T cell lymphomas in mice: frequency and mechanisms of activation. EMBO J. **3:**3215–3222.

65. **Shen, H., T. Suzuki, D. J. Munroe, C. Stewart, L. Rasmussen, D. J. Gilbert, N. A. Jenkins, and N. G. Copeland.** 2003. Common sites of retroviral integration in mouse hematopoietic tumors identified by high-throughput, single nucleotide polymorphism-based mapping and bacterial artificial chromosome hybridization. J. Virol. **77:**1584–1588.

66. **Shih, C. C., J. P. Stoye, and J. M. Coffin.** 1988. Highly preferred targets for retrovirus integration. Cell **53:**531–537.

67. **Sørensen, A. B., M. Duch, H. W. Amtoft, P. Jørgensen, and F. S. Pedersen.** 1996. Sequence tags of provirus integration sites in DNAs of tumors induced by the murine retrovirus SL3-3. J. Virol. **70:**4063–4070.

68. **Sørensen, A. B., M. Duch, P. Jørgensen, and F. S. Pedersen.** 1993. Amplification and sequence analysis of DNA flanking integrated proviruses by a simple two-step polymerase chain reaction method. J. Virol. **67:**7118–7124.

69. **Stevens, S. W., and J. D. Griffith.** 1996. Sequence analysis of the human DNA flanking sites of human immunodeficiency virus type 1 integration. J. Virol. **70:**6459–6462.

70. **Suzuki, T., H. Shen, K. Akagi, H. C. Morse, J. D. Malley, D. Q. Naiman, N. A. Jenkins, and N. G. Copeland.** 2002. New genes involved in cancer identified by retroviral tagging. Nat. Genet. **32:**166–174.

71. **Taganov, K., R. Daniel, R. A. Katz, O. Favorova, and A. M. Skalka.** 2001. Characterization of retrovirus-host DNA junctions in cells deficient in non-homologous-end joining. J. Virol. **75:**9549–9552.

72. **Van Beveren, C., E. Rands, S. K. Chattopadhyay, D. R. Lowy, and I. M. Verma.** 1982. Long terminal repeat of murine retroviral DNAs: sequence analysis, host-proviral junctions, and preintegration site. J. Virol. **41:**542–556.

73. **Vijaya, S., D. L. Steffen, and H. L. Robinson.** 1986. Acceptor sites for retroviral integrations map near DNase I-hypersensitive sites in chromatin. J. Virol. **60:**683–692.

74. **Whitcomb, J. M., and S. H. Hughes.** 1992. Retroviral reverse transcription and integration: progress and problems. Annu. Rev. Cell Biol. **8:**275–306.

75. **Withers-Ward, E. S., Y. Kitamura, J. P. Barnes, and J. M. Coffin.** 1994. Distribution of targets for avian retrovirus DNA integration in vivo. Genes Dev. **8:**1473–1487.

76. **Wu, X., Y. Li, B. Crise, and S. M. Burgess.** 2003. Transcription start regions in the human genome are favored targets for MLV integration. Science **300:**1749–1751.

77. **Yoder, K. E., and F. D. Bushman.** 2000. Repair of gaps in retroviral DNA integration intermediates. J. Virol. **74:**11191–11200.