Foamy virus vector integration sites in normal human cells

Grant D. Trobridge*,†, Daniel G. Miller‡, Michael A. Jacobs§, James M. Allen*, Hans-Peter Kiem†, Rajinder Kaul§, and David W. Russell*¶

*Department of Medicine, Division of Hematology, †Department of Medicine, Division of Medical Genetics, ‡Department of Pediatrics, Division of Genetics and Developmental Medicine, and §Department of Biochemistry, University of Washington, Seattle, WA 98195; and ¶Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109

Edited by Inder M. Verma, The Salk Institute for Biological Studies, La Jolla, CA, and approved December 5, 2005 (received for review November 18, 2005)

Foamy viruses (FVs) or spumaviruses are retroviruses that have been developed as vectors, but their integration patterns have not been described. We have performed a large-scale analysis of FV integration sites in unselected human fibroblasts (n = 1,008) and human CD34+ hematopoietic cells (n = 1,821) by using a bacterial shuttle vector and a comparable analysis of lentiviral vector integration sites in CD34+ cells (n = 1,331). FV vectors had a distinct integration profile relative to other types of retroviruses. They did not integrate preferentially within genes, despite a modest preference for integration near transcription start sites and a significant preference for CpG islands. The genomewide distribution of FV vector proviruses was nonrandom, with both clusters and gaps. Transcriptional profiling showed that gene expression had little influence on integration site selection. Our findings suggest that FV vectors may have desirable integration properties for gene therapy applications.

Integrating viral vectors can provide the long-term transgene expression required in many gene therapy applications, especially when proliferating cells such as hematopoietic stem cells are transduced. However, integration also results in insertional mutagenesis that may activate oncogenes, as seen in X-linked severe combined immune deficiency patients that developed leukemia after retroviral vectors integrated near the LMO2 protooncogene (1). Large-scale integration site analyses can help define the spectrum of insertional mutagenesis expected of a viral vector, and studies of murine leukemia virus (MLV), HIV, and avian sarcoma virus (ASV) have demonstrated distinct integration patterns (2–6). The relationship of integration to transcription is especially important, as it may determine the likelihood of oncogene activation. The two most commonly used retroviral vectors differ in this respect, with HIV vectors integrating preferentially throughout transcription units, and MLV vectors integrating preferentially near transcription start sites (2, 3).

Foamy viruses (FVs), or spumaviruses, comprise a class of retroviruses distinct from oncoviruses such as MLV and lentiviruses such as HIV. FVs are found in several mammalian species (7) but are not endemic in human populations (8), and the prototype “human FV” isolate (9) is now thought to be a chimpanzee virus [SFVcpz(hu)] (10, 11). FVs have several distinguishing properties, including a transcriptional transactivator expressed from an internal promoter (12–14), a DNA-based genome that forms by reverse transcription in virion-producing cells (15, 16), and an essential cis-acting region overlapping the pol and env genes (17–19).

FV vectors have been developed (19–22) and shown to have a broad host range (21, 23, 24), large packaging capacity (25), stable virions that can be concentrated by ultracentrifugation (26), and prolonged survival of the preintegration complex in quiescent cells (27). They are especially effective at transducing hematopoietic stem cells of mice and humans (26, 28–31), making them promising vectors for the treatment of hematologic diseases. Although FV vectors integrate into the host genome of transduced cells, little is known about where integration occurs. Previous studies mapped 12 FV vector integrations to both introns and intergenic regions in human hematopoietic cells (29) and 7 FV integrations in satellite DNA and ribosomal DNA in a murine fibroblast cell line (32).

Importantly, wild-type FVs are not pathogenic in their natural hosts (33) or in humans exposed to primate FVs (34). Despite their prevalence in well studied captive primate populations used for research, FVs have not been associated with the development of malignancies. Although this lack of pathogenicity is reassuring and contrasts with the biology of MLV, it could be caused by cytotoxic effects of the wild-type virus rather than differences in integration site selection. Depending on where proviruses integrate, replication-incompetent FV vectors that contain strong internal promoters and enhancers may activate protooncogenes. To determine where FV proviruses integrate, we have performed a large-scale analysis of FV vector integration sites in normal human cells and compared our results with those obtained with other integrating vectors.

Results

We used a shuttle vector system to rescue and sequence provirus integration sites (Fig. 1). The FV vector ΔΦFmcsNO contains a GFP gene under the control of a phosphoglycerate kinase (PGK) promoter that allowed us to measure transduction frequencies, and plasmid sequences consisting of a Tn5 promoter, neomycin resistance gene (neo), and p15A replication origin that function in bacteria. Normal human fibroblasts and human CD34+ peripheral blood cells were infected with ΔΦFmcsNO and cultured without selection to allow time for integration to occur. Genomic DNA was then isolated from infected cells, digested with specific restriction enzymes to release the bacterial plasmid and flanking genomic sequences from each provirus, circularized, and rescued in Escherichia coli. We sequenced a total of 4,963 rescued plasmids. Different restriction enzymes were used to avoid potential biases in recognition site distribution, and several criteria were used to exclude sequence reads that might lead to inaccuracies, including those with deleted LTRs, poor alignments, and low BLAST scores (see Table 2, which is published as supporting information on the PNAS web site). A total of 638 plasmids contained junction sequences that were identical to other plasmid junctions, and these were considered to be duplicate recoveries of the same provirus. With these criteria, 1,008 and 1,821 unique integrants were localized to build 35 of the human genome sequence in fibroblasts and CD34+ cells, respectively. Five integrants (0.18%) mapped to the 43-kb ribosomal DNA repeats, which are not localized in the genome sequence but are present at an estimated 400 copies in a chromosome 14.

Conflict of interest statement: No conflicts declared.

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: FV, foamy virus; MLV, murine leukemia virus; ASV, avian sarcoma virus; AAV, adeno-associated virus; PGK, phosphoglycerate kinase; RefSeq, Reference Sequence.

To whom correspondence should be addressed. E-mail: drussell@u.washington.edu.

© 2006 by The National Academy of Sciences of the USA
To provide an appropriate data set for comparison, we also rescued and sequenced proviruses from CD34+ cells infected with HIV-1-based vector HIV-PYmcsNO, which expresses yellow fluorescent protein from the PGK promoter and contains the same shuttle vector sequences as ΔΔPFmcsNO. A total of 1,331 HIV vector integrants were localized to the human genome sequence by using the same criteria as for FV vector sites (see Table 2, Table 1). This finding supports the conclusion that differences seen between FV and other vectors were not caused by the method used to isolate junctions and suggests that neither method imposes a large bias to the data. This comparison also demonstrates that integration preferences are similar in CD34+ cells and transformed cell lines, at least for HIV vectors.

Compared with the random control set, FV vectors preferentially integrated in or near CpG islands, and there were fewer integrants than expected in genes and long interspersed nuclear element (LINE) repeats. In contrast, the three other types of retroviruses all integrated preferentially in genes. MLV vectors were similar to FV vectors regarding their preference for CpG islands and reduced integration in LINE repeats. FV vector proviruses were found in all human chromosomes (Fig. 2A). Relative to the set of random sites, the three largest chromosomes had fewer proviruses than expected, whereas several of the smallest chromosomes (17–21) had more proviruses than expected. The cytogenetic distributions of both FV vector and random sites are shown in Fig. 2B. The FV sites exhibited both clustering and gaps, which were quantified by determining the distances between neighboring integration sites (Fig. 3A). Compared with random sites, FV vectors were more likely to integrate near other integration sites, with almost 1% of neighboring integrants lying within 1 kb of each other, as opposed to 0.11% of

![Fig. 1. FV shuttle vector rescue strategy. The shuttle vector ΔΔPFmcsNO is based on the ΔΔp (deleted foamy) backbone (25) and is shown with its PGK promoter, GFP transgene, multiple cloning site, neo gene driven by a TN5-derived bacterial promoter, p15A origin, 3' LTR, and flanking chromosomal DNA were circularized by ligation and transferred to bacteria, where they replicate and confer kanamycin resistance. Unwanted fragments containing 5' vector sequences or bacterial plasmid DNA were destroyed by digestion with the rare-cutting endonuclease I-SceI and the methylation-sensitive enzyme DpnI, respectively. Bacterial colonies were picked and their DNA.](image-url)

### Table 1. Genomic features of integration sites

<table>
<thead>
<tr>
<th>Virus or vector</th>
<th>Unique sites</th>
<th>RefSeq genes*</th>
<th>Cpg islands†</th>
<th>LINE</th>
<th>SINE</th>
<th>DNA</th>
<th>LTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>FV&lt;sub&gt;fibro&lt;/sub&gt;§</td>
<td>1,008</td>
<td>24.0</td>
<td>5.85</td>
<td>15.7</td>
<td>12.3</td>
<td>3.57</td>
<td>9.42</td>
</tr>
<tr>
<td>FV&lt;sub&gt;CD34&lt;/sub&gt;§</td>
<td>1,821</td>
<td>31.1</td>
<td>12.1</td>
<td>14.6</td>
<td>16.6</td>
<td>2.91</td>
<td>6.48</td>
</tr>
<tr>
<td>FV&lt;sub&gt;ALL&lt;/sub&gt;</td>
<td>2,829</td>
<td>28.6</td>
<td>9.93</td>
<td>15.0</td>
<td>15.1</td>
<td>3.14</td>
<td>7.52</td>
</tr>
<tr>
<td>Random</td>
<td>10,000</td>
<td>34.1</td>
<td>2.56</td>
<td>20.0</td>
<td>13.5</td>
<td>3.20</td>
<td>8.78</td>
</tr>
<tr>
<td>HIV&lt;sub&gt;shuttle&lt;/sub&gt;§</td>
<td>1,331</td>
<td>69.1</td>
<td>1.50</td>
<td>17.1</td>
<td>18.2</td>
<td>3.83</td>
<td>3.83</td>
</tr>
<tr>
<td>HIV&lt;sub&gt;neo&lt;/sub&gt;§</td>
<td>1,757</td>
<td>65.2</td>
<td>1.46</td>
<td>18.1</td>
<td>15.4</td>
<td>3.36</td>
<td>3.87</td>
</tr>
<tr>
<td>MLV&lt;sub&gt;TVA&lt;/sub&gt;</td>
<td>644</td>
<td>43.3</td>
<td>17.7</td>
<td>10.4</td>
<td>20.7</td>
<td>2.64</td>
<td>6.06</td>
</tr>
<tr>
<td>ASV&lt;sup‡‡&lt;/sup&gt;</td>
<td>480</td>
<td>42.9</td>
<td>2.71</td>
<td>20.6</td>
<td>12.7</td>
<td>2.92</td>
<td>7.29</td>
</tr>
<tr>
<td>AAV&lt;sup‡‡&lt;/sup&gt;</td>
<td>670</td>
<td>38.8</td>
<td>4.03</td>
<td>18.5</td>
<td>13.5</td>
<td>3.13</td>
<td>5.67</td>
</tr>
</tbody>
</table>

*Sites in transcribed regions of RefSeq genes.
†Sites inside of or within 1 kb of Cpg islands.
‡Repetitive elements as defined at the University of California, Santa Cruz genome browser.
§Genbank accession numbers DU798511–DU799518.
**Genbank accession numbers DU796690–DU798510.
††Genbank accession numbers DU799519–DU800849.
‡‡Genbank accession numbers DU799519–DU800849.
§§Genbank accession numbers DU798511–DU799518.
As shown in Table 1, FV, MLV, and AAV vectors frequently found near promoter regions and may regulate transcription units, in agreement with prior analyses (2, 4, 5). In contrast, HIV MLV in that it was greatest within 2.5 kb upstream or downstream of start sites (3), but of a lesser magnitude (10.6% of all FV integrants as compared with 18.7% for MLV). In contrast, HIV MLV integrated preferentially in sequences flanking CpG islands, but less so within islands, and AAV vectors integrated preferentially within islands but not in flanking regions. FV vectors also integrated more frequently near CpG islands in CD34+ cells than in fibroblasts (see Table 1; P < 0.01).

Additional analyses were performed to assess the relationship of restriction enzyme site distributions influenced these analyses, we plotted the locations of recognition sites for each restriction enzyme or enzyme combination used relative to Reference Sequence (RefSeq) genes (National Center for Biotechnology Information, www.ncbi.nlm.nih.gov/RefSeq) and CpG islands (Figs. 6 and 7, which are published as supporting information on the PNAS web site). Although enzyme-specific distributions were observed, they did not account for the distribution of FV vector integrants relative to these genomic features. However, in the case of published MLV data (3), the low number of integrants present within CpG islands may have been caused by a lack of MseI sites (Fig. 8, which is published as supporting information on the PNAS web site).

Fig. 2. Chromosomal distribution of integration sites. (A) FV vector integration sites from both CD34+ cells and fibroblasts (n = 2,829) are plotted as a percentage of all integrants in each chromosome and compared with a set of random sites (n = 10,000). Asterisks mark chromosomes with significantly different integration frequencies (P < 0.01). (B) The same FV vector integration sites are represented as individual dots positioned above each human chromosome and compared with 2,829 random sites placed below the chromosomes. Dots are 20% opaque to display multiple overlapping integrants.

Fig. 3. Clusters and gaps of FV vector integration sites. (A) The distances between adjacent unique FV vector integration sites were determined and binned by size, with the percent of proviruses within each bin plotted. (B) The number of FV vector integration hotspots (defined as three integrations) was determined for each sequence window size from 2.5 to 100 kb. Three size-matched sets of 2,829 random sites were also plotted with standard deviations as controls. Asterisks mark significant differences (P < 0.01).

CpG islands are sequences rich in CpG dinucleotides that are frequently found near promoter regions and may regulate transcription (36, 37). As shown in Table 1, FV, MLV, and AAV vectors all integrated preferentially within 1 kb of CpG islands. However, on further examination there were distinct differences for all three vector types (Fig. 4B). FV integration preferences were greatest within CpG islands and steadily decreased in sequences extending outward. MLV integrated preferentially in sequences flanking CpG islands, but less so within islands, and AAV vectors integrated preferentially within islands but not in flanking regions. FV vectors also integrated more frequently near CpG islands in CD34+ cells than in fibroblasts (see Table 1; P < 0.01).

Gene expression arrays were used to determine whether integration occurred preferentially in expressed genes. We ranked 13,069 RefSeq genes by expression level in uninfected and infected CD34+ cells and identified those where FV vectors integrated (Fig. 5). The average expression rank of genes where random sites were located in both infected and uninfected cells. The average expression rank of genes where
integrations occurred within 10 kb of transcription starts was 1.061- to 1.066-fold above that of random sites. These results support the hypothesis that gene expression does not have a major impact on FV vector integration. This finding was also confirmed by comparing the expression levels of genes where FV vectors integrated in fibroblasts and CD34+ cells, which showed little effect of cell-type-specific expression on integration (Fig. 9, which is published as supporting information on the PNAS web site).

Discussion

In this study, we used a shuttle vector system to rescue integrated FV vector proviruses as bacterial plasmids, and robotic processes were used to purify plasmid DNA and sequence junctions. Bacterial rescue of retroviral shuttle vectors is a well-established technique (38, 39), and here we have used this efficient and scalable method to recover thousands of proviruses from normal human cells without selection. There are certain advantages to this approach as compared with more commonly used PCR-based methods. The length of junction sequences recovered is not limited by the amplification steps of PCR, and the sequence obtained from plasmids is of high quality, so long sequence reads are generated. The average match length in our FV vector experiments was 337 bp, whereas the PCR methods used to sequence MLV and HIV integration junctions (2, 3) generated average match lengths of 107 and 236 bp, respectively when aligned by our criteria. This increase in sequence length allows a more accurate localization of junction sequences, especially those containing repeats, and additional sequence can be obtained from recovered plasmids if necessary. The major disadvantage of our approach is that the vector must include a bacterial replication origin and selectable marker. Although these would not typically be included in a gene therapy vector, they could be built into vectors when size permits. The similar results obtained for HIV integration sites generated by plasmid rescue or PCR validates both methods and strongly supports the conclusion that FV vectors have distinct integration preferences as compared with other integrating vectors.

FV vectors were found to integrate in a nonrandom distribution, with preferences for CpG islands and transcription start sites, and fewer integrations than expected in long interspersed nuclear elements. Although some of these patterns were similar to other integrating viruses that have been studied, in particular MLV and AAV, the combination of properties was unique to FV vectors. Thus, there may be both common and unique features in the life cycles of viruses before or during integration. Clustering of FV integrants was observed, and because of the large number of integrants localized, we could identify several chromosomal regions where FV vectors did not integrate. The basis for these gaps is not clear, as they contain genes expressed in both fibroblasts and CD34+ cells (data not shown) and are not simply inactive chromatin. The same regions also contained integrants of other retroviruses studied in different cell lines and HIV vector integrants isolated by shuttle vector rescue in CD34+ cells (data not shown).

A key issue addressed by our study is the potential for insertional mutagenesis by FV vectors and how this compares with other integrating vectors that might be used in gene therapy. Unlike MLV or HIV, FV vectors did not show an overall preference for integrating within genes (Table 1), and their preferences for integrating near transcription start sites and CpG islands were lower than those of MLV (Fig. 4), which may decrease the risk of mutating and/or activating cellular genes. We examined more directly the possibility of protooncogene activation and identified 126 FV vector integrants (4.4%) within protooncogene transcripts or 50 kb upstream from their start site (Table 3, which is published as supporting information on the PNAS web site). This percentage...
was slightly higher (P < 0.01) than the average obtained from three sets of 2,829 randomly generated sites (3.2% ± standard deviation of 0.5%), but lower (P < 0.01) than the HIV (shuttle 7.0%, PCR 6.7%) or MLV (7.8%) data sets.

One of four hotspots containing four or more FV vector integrants within 50 kb (Fig. 10, which is published as supporting information on the PNAS web site), was 470 kb from the Evi1 locus on chromosome 3, originally identified as a common MLV integration site in myeloid tumors of mice (40). It is not clear whether these integrants are close enough to influence Evi1 expression, although in mice an MLV provirus located >500 kb away from the start site increased Evi1 transcription (41). Interestingly, in nonhuman primates several MLV vector integrations were also observed near Evi1, but in the adjacent Mds1 gene that partially overlaps some Evi1 transcripts in humans (6). The closest FV vector integrant to the Lmo2 protooncogene activated by MLV vector insertions in leukemias that developed in the X-linked severe combined immune deficiency trait (1) was ~700 kb away.

Although our findings suggest that FV vectors may have a safer spectrum of insertional mutagenesis compared with MLV or HIV vectors, it is important to note that the integration preferences described generally varied by 2-fold or less, which would not provide a significant improvement in safety. Other factors are likely to play a more important role in determining the oncogenic risk of vectors, including differences in enhancer activity, potential insulating effects of proviral sequences, the number of integrations that occur in a clinical trial, and the cell types in which they occur. The lack of pathogenicity of wild-type FV is encouraging in this regard, but does not take into account potential effects of internal promoters and transgenes included in vectors. Future experiments will be required to address these factors and improve the safety of integrating viral vectors.

Materials and Methods

Vector Production. The FV vector plasmid pΔPPFmcsNO was constructed by using standard molecular biology techniques by replacing the murine stem cell virus promoter of pΔPPmcsFv (25) with a murine PGK promoter and adding bacterial plasmid sequences including a Tn5-driven neo gene and p15A replication origin (42), and a multiple cloning site containing I-SceI, BamHI, AvrII, PciI, NsiI, NdeI, and Bsp120I recognition sites. The HIV-1-based shuttle vector HIV-PYmcsNO also contains the same Tn5-driven neo gene, p15A replication origin, and multiple cloning site and was made from pRRLsinCPPT.Pgk.EYFP.wpre (kindly provided by Luigi Naldini, Fondazione San Raffaele del Monte Tabor, Milan), which contains a yellow fluorescent protein in place of the GFP in pRRLsin.hPGK.EGFP.wpre vector (43). FV vector preparations were produced by calcium phosphate transfection as described (25, 44) except that 12 μg of pΔPPFmcsNO, 12 μg of pGGSΔΨ (28), 1.6 μg of pGIPS, and 0.75 μg of pCIES in a total volume of 800 μl were used for each 10-cm tissue culture dish. HIV vector preparations were produced as described (45). Stocks were titered by determining the number of GFP-transducing units on normal human fibroblasts or human HT-1080 fibrosarcoma cells.

Cell Culture and Infections. Normal human fibroblasts obtained from the Correll Institute for Medical Research (Camden, NJ; repository GM05387) and IT-1080 cells (46) were cultured in DMEM with 10% heat-inactivated (56°C for 30 min) FBS (HyClone, Logan, UT), 1.25 μg/ml amphotericin, 100 units/ml of penicillin, and 100 μg/ml streptomycin at 37°C in 5% CO2 atmosphere. For fibroblast infections, two 10-cm dishes were each seeded with 5 × 10⁶ cells on day 1, vector ΔPPFmcsNO was added at a multiplicity of infection of 0.5 on day 2, the cells were expanded to four 10-cm dishes on day 5, and DNA was isolated on day 8.

Human CD34+ cells were collected from three male volunteers and isolated by magnetic beads (Miltenyi Biotec, Auburn, CA) according to the manufacturer’s instructions and stored in liquid nitrogen until use. Cells were thawed on day 1 and cultured in transduction medium (DMEM with 20% FBS, 100 ng/ml Fill-3 ligand, 100 ng/ml stem cell factor, and 100 ng/ml thrombopoietin). For infections, 5 × 10⁶ CD34+ cells thawed 1 day previously were added to each well of a 6-well plate pretreated with 50 μg/ml CH-296 fibronectin fragment (Takara, New York), and exposed to vector for 10 h at multiplicity of infection (MOI) 1.5 (ΔPPFmcsNO) or 18 h at MOI 5 (HIV-PYmcsNO). Cells were then washed, resuspended in transduction medium supplemented with 20 ng/ml IL-3 and 20 ng/ml IL-6, and plated in a 6-well plate. RNA was isolated for gene expression analysis 48 h later, and DNA was isolated 4 days later. All cytokines were obtained from PeproTech (Rocky Hill, NJ) and were of (recombinant) human origin.

Provirus Rescue and Sequencing of Junctions. DNA from infected fibroblasts or CD34+ cells was isolated by using a Puregene isolation kit (Gentra Systems) following the manufacturer’s instructions, followed by extraction with phenol and chloroform and ethanol precipitation. Twenty micrograms of DNA was digested for 3 h with 40 units of either NdeI, NsiI, AvrII, and SpeI, or BspHII and FclI, and the DNA fragments were extracted with phenol and chloroform, precipitated with ethanol, and resuspended in 355 μl of H2O. For ligations, 5 μl of T4 DNA ligase (400 units/μl; New England Biolabs) and 40 μl of 10× ligation buffer were added, incubated overnight at 15°C, heat-inactivated at 65°C for 10 min, and cooled to room temperature. Twenty units of I-SceI was added and incubated at 37°C for 1 h to eliminate ligated fragments containing sequences upstream of the multiple cloning site. The reaction was then adjusted to 50 mM NaCl and digested with 40 units of methylation-sensitive DpnI for 30 min at 37°C to eliminate bacterial plasmids that may have been present in the vector stock. The DNA was then extracted with phenol and chloroform, precipitated with ethanol, and resuspended in 10 μl of 1 mM Tris (pH 8.0) buffer with 0.1 mM EDTA. DH10B cells (47) were transformed by electroporation with 1 μl (~1–2 μg of DNA) at a time, and transformed bacteria were grown on agar containing 50 μg/ml kanamycin. Kanamycin-resistant colonies were picked by using a Qpix2 robot (Genetex, Boston), and plasmid DNA containing rescued proviruses was purified and sequenced as described (48) by using FV sequencing primer 5’-AACGACTTGAT-CCAGAAC-5’ and HIV sequencing primer 5’-TCTGCTTGCTT-GAGCATC-5’. Sequencing reactions were electrophoresed on ABI prism 3700 (Applied Biosystems) capillary sequencers and analyzed by using PHRED/PHRAP/CONSED software tools (49–51).

Microarray Analysis of Gene Expression. RNA from fibroblasts and CD34+ cells was isolated with the RNeasy kit and Qiachredder column (Qiagen, Valencia, CA) following the manufacturer’s instructions. Five micrograms of total RNA was labeled and 15 μg of cRNA was used per HG-U133 plus 2.0 array (Affymetrix, Santa Clara, CA) per the manufacturer’s instructions. The subset of probes that identify specific RefSeq transcripts (13,069) were used in our analysis. Probe sets that hybridized with more than one gene were excluded. Expression levels were determined by using Affymetrix GCOS 1.1.1 software, and when multiple probe sets reflected the expression level of a single RefSeq gene, the average expression level was used in the rankings.

Analysis of Integration Sites. DNA sequences were processed with PERL programs, and sequences containing an LTR junction were truncated at 500 bp, or at the first occurrence of a restriction site used during rescue. The resulting junction sequences were aligned to build 35 of the human genome, the vector plasmid sequence, and human ribosomal DNA by using a stand-alone version of BLAT (52) that generates a BLAST alignment score. The input script was as follows: blat chromosome.file query.file -out=blast.out -ooc=11.ooc output.file. Alignments were sorted by BLAST score, and those with

Trobridge et al.
the five highest scores were saved for further processing. To ensure that sequences represented true junctions and were accurately localized, additional screening criteria were used as described in Table 2. The average match length was 337 bp for all FV sequences and 301 bp for all HIV sequences. PERL programs were also used to compare localized integration sites to various chromosomal features by using tables available from the University of California, Santa Cruz database (53) and to determine the positions of restriction enzyme sites in the human genome. Additional published retroviral integration sites (2–5) and AAV vector integration sites (54) were processed as above for comparison (GenBank accession numbers in Table 1).

We also produced a randomly localized set of genomic positions by generating random numbers between 1 and 5,941,037,819 (the size of the build 35 diploid male genome with chromosomes laid end to end) with the PERL “rand” function and converting to chromosomal positions as size-matched controls.

To identify FV vector integrants in protooncogenes or near their promoters we searched the Sanger Institute Cancer Gene Census Table (www.sanger.ac.uk/genetics/CGP/Census), which contains 331 unique potential oncogenes with corresponding RefSeq transcripts (as of April 8, 2005), and the Mouse Retrovirus Tagged Gene Database (http://rtg.UNCfcrf.gov), which contains 363 potential oncogenes with corresponding RefSeq transcripts (as of November 5, 2005). PERL programs were used to identify RefSeq genes that contained FV vector integration sites or random sites within their transcribed regions or within 50 kb upstream of the transcript start site.

Statistical Analysis. Statistical significance was determined by using the χ2 test. P values < 0.01 were considered significant.

We thank Maynard Olson and the University of Washington Genome Center for helpful advice and sequencing. Erik Olson and Tulin Okbinoğlu for help with stock production, the Center for Expression Arrays (University of Washington) for microarray analysis, Shelly Heinfeld and the Fred Hutchinson Cell Processing Core for isolation of human CD34+ cells, and Keiko Akagi and Neal Copeland for providing a table of human genes homologous to potential mouse oncogenes from the Retrovirus Tagged Cancer Gene Database. This work was supported by grants from the National Institutes of Health, the Child Health Research Center at Children’s Hospital, Seattle, and the University of Washington Center for AIDS Research.

Supplemental Figure 1. Integration and restriction site analysis relative to RefSeq transcript start sites. (A-D) FV vector integration sites cloned with specific restriction enzymes (Avr II and Spe I; n = 292, BspH I and Pci I; n = 518, Nde I; n = 776, or Nsi I; n = 1068), a computer-generated set of random sites (n = 10,000), and the enzyme recognition sites (Avr II and Spe I; n = 981,550, BspH I and Pci I; n = 9,649,146, Nde I; n = 894,673, or Nsi I; n = 910,268) are plotted as a percent of sites/kb within different sequence windows extending from RefSeq transcript start sites. (E-G) The data for each restriction enzyme or combination are plotted as in A-D but at a different scale.
Supplemental Figure 2. Integration and restriction site analysis relative to CpG islands. (A-D) FV vector integration sites, random sites, and enzyme recognition sites (see Supplemental Figure 1 Legend for details) are plotted as a percent of sites within CpG islands and fourteen 0.75 kb windows flanking each island in the human genome. (E-G) The data for each restriction enzyme or combination are plotted as in A-D but at a different scale.
Supplemental Figure 3. MLV vector integration and restriction site analysis relative to CpG islands. A) Published MLV vector integration sites cloned with Mse I (n = 644) (3), a computer generated set of random sites (n = 10,000) and Mse I recognition sites (n = 18,992,371) are plotted as a percent of sites within CpG islands and fourteen 0.75 kb windows flanking each island in the human genome. B) The data for Mse I sites are plotted as in (A) but at a different scale.
Supplemental Figure 4. Microarray analysis in human CD34+ cells and fibroblasts. Expression levels of RefSeq genes containing FV vector integrants were divided into 8 equal bins based on the ranked expression levels of all 13,069 RefSeq genes assayed on the Affymetrix chip. CD34+Int-CD34+Rank are 423 genes with integrants in CD34+ cells binned according to expression ranks in CD34+ cells. FibroInt-CD34+Rank are 165 genes with integrants in fibroblasts binned according to expression rank in CD34+ cells. CD34+Int-FibroRank are 423 genes with integrants in CD34+ cells binned according to expression rank in fibroblasts. FibroInt-FibroRank are 165 genes with integrants in fibroblasts binned according to expression rank in fibroblasts.
Supplemental Figure 5. FV vector integration hotspot near EVI1 locus. The location of a FV vector integration hotspot (defined as at least 4 integrations within 50 kb) on chromosome 3 is shown, approximately 470 kb away from the EVI1 locus (image derived from the UCSC genome browser).