ARTICLE

doi:10.1016/j.ymthe.2006.06.006

The Integration Profile of EIAV-Based Vectors

Caroline V. Hacker,[1,*] Conrad A. Vink,[1,†] Theresa W. Wardell,[1] Sheena Lee,[1,‡] Peter Treasure,[2] Susan M. Kingsman,[1] Kyriacos A. Mitrophanous,[1] and James E. Miskin[1]

[1]*Oxford BioMedica UK Ltd., Medawar Centre, The Oxford Science Park, Oxford OX4 4GA, UK*
[2]*Peter Treasure Statistical Services Ltd., Hill Farm Houses, Stow Bridge, King's Lynn PE34 3NR, UK*

**To whom correspondence and reprint requests should be addressed. Fax: +44 (0) 1865 783001. E-mail: c.hacker@oxfordbiomedica.co.uk.*

[†]*Present address: UCL Institute of Child Health, 30 Guilford Street, London WC1N 1EH, UK.*

[‡]*Present address: Department of Human Anatomy & Genetics, University of Oxford, South Parks Road, Oxford OX1 3QX, UK.*

Lentiviral vectors based on equine infectious anemia virus (EIAV) stably integrate into dividing and nondividing cells such as neurons, conferring long-term expression of their transgene. The integration profile of an EIAV vector was analyzed in dividing HEK293T cells, alongside an HIV-1 vector as a control, and compared to a random dataset generated *in silico*. A multivariate regression model was generated and the influence of the following parameters on integration site selection determined: (a) within/not within a gene, (b) GC content within 20 kb, (c) within 10 kb of a CpG island, (d) gene density within a 2-Mb window, and (e) chromosome number. The majority of the EIAV integration sites (68%; *n* = 458) and HIV-1 integration sites (72%; *n* = 162) were within a gene, and both vectors favored AT-rich regions. Sites within genes were examined using a second model to determine the influence of the gene-specific parameters, gene region, and transcriptional activity. Both EIAV and HIV-1 vectors preferentially integrated within active genes. Unlike the gammaretrovirus MLV, EIAV and HIV-1 vectors do not integrate preferentially into the promoter region or the 5′ end of the transcription unit.

Key Words: lentiviral vectors, EIAV, gene therapy, integration, safety

INTRODUCTION

Upon host cell entry, viral-encoded reverse transcriptase catalyzes the production of a double-stranded DNA (dsDNA) copy of the lentiviral RNA genome. The resulting dsDNA integrates into the host genome, a process that is mediated by viral integrase. The stable integration of a DNA copy of the viral genome into the host cell enables long-term expression of viral proteins and the subsequent production of progeny virus [1] in the case of the parental replication-competent viruses. Replication-defective lentiviral vectors have been engineered to exploit these biological characteristics to introduce therapeutic genes into target cell chromosomes, enabling long-term expression of the transgene. Vectors based on lentiviruses such as human immunodeficiency virus type-1 (HIV-1) [2,3] and equine infectious anemia virus (EIAV) [4,5] have the advantage over retroviral vectors of being able to integrate into the genome of quiescent cells. We have previously shown effective long-term gene transfer and expression into nondividing cells using EIAV vectors; examples are expression for up to 5 months of genes in the dopamine biosynthetic pathway in rat striatal neurons [6] and expression of a functional arginine vasopressin gene in magnocellular neurons of the Brattleboro rat model of diabetes insipidus for up to 1 year [7].

The publication of the human genome sequence and the development of highly sensitive PCR techniques such as ligation-mediated PCR (LM-PCR) [8] and linear-amplification-mediated PCR [9] have enabled large-scale integration site profiling of integrating vectors such as HIV-1 [10,11] and simian immunodeficiency virus (SIV) [12], as well as the gammaretroviruses avian sarcoma leukosis virus (ASLV) [11,13] and murine leukemia virus (MLV) [14]. Profiling has shown that site selection by retroviruses is not a random event. HIV-1 [10] and SIV [12] show a preference for transcription units and, more specifically, actively transcribed genes. In contrast, MLV has a preference for the start of a transcription unit and also the promoter region [14]. Specific integration into transcription units has also been described for ASLV [11,13]. The basis for the variation in retroviral integration profiles is not understood. Localized genomic features such as the primary sequence at the integration site are thought to

1525-0016/$30.00

have only a weak influence on site selection [15]. Other factors suggested to influence site selection include: (a) the openness of the host chromatin, (b) the interaction between the preintegration complex and the host cellular DNA-binding proteins, and (c) the stage of the host cell cycle at the time of transduction/infection [16].

We investigated the integration profile of an EIAV vector in dividing HEK293T cells, using LM-PCR [8], and compared it to that of an HIV-1 vector. We used statistical methods to determine whether there were any significant relationships between a set of defined parameters that might be expected to influence integration. We used a logistic regression method to determine the probability of integration at a site using either a single parameter (univariate analysis) or multiple parameters (multivariate analysis). In some cases, as expected, correlations obtained in the univariate model were not significant or had an altered magnitude in the multivariate model. Our analyses indicate that EIAV and HIV-1 vectors share a weak preference for integration at a similar palindromic sequence, a preference for AT-rich regions, and a highly significant preference for integration within transcription units, particularly those that are actively transcribed. A preference for actively transcribed regions has been described previously for HIV-1 and SIV vectors [10,12]. We have therefore added to an increasing body of data pointing to a conserved integration profile for all lentiviral vectors studied to date (HIV-1, SIV, and now EIAV). Furthermore the integration profiles of all of these lentiviral vectors are significantly different from that observed with the gammaretrovirus MLV.

RESULTS

Generation of *in Vitro* and Random Control Datasets

We mapped 620 independent lentiviral integration sites (n = 458 for EIAV, n = 162 for HIV-1) in HEK293T cells using LM-PCR [8] (GenBank Accession Nos. DQ498202–DQ498763). In addition, we generated 10,000 random sites subject to the same criteria as the *in vivo* dataset, resulting in a final control dataset of 7860 sites.

We carried out multivariate and univariate logistic regressions (based on the analysis carried out by Mitchell *et al.* [11]) to predict the viral integration sites for HIV-1 and EIAV vectors. The analyses used the following variables as predictors: (1) whether or not the site was within a gene, (2) the gene density in a 2-Mb window centered on the integration site, (3) if the distance to the nearest CpG island was greater that 10 kb, (4) the GC content in a 20-kb window centered on the integration site, (5) whether or not the site was within the 2-kb promoter region, and (6) the chromosome number. A gene-only analysis was also carried out and used the following variables as predictors: (1) gene region and (2) gene expression level.

Analysis of All Sites

We determined the transcriptional activity of HEK293T cells using the Affymetrix HG-focus chip, which assays over 8500 human genes (data not shown). We used this information, together with gene density information that was obtained from the public domain [17], to generate an image of the human chromosomes on which gene density and transcriptional activity were superimposed. We measured each parameter within a 2-Mb region, assembled the information for each chromosome, and represented it as shaded regions on a map. We then superimposed the mapped position of each integration site for EIAV and HIV-1 onto this diagram of the chromosomes, which resulted in a visual representation of integration and localized genomic features within the human genome (Fig. 1). Two striking observations are apparent from Fig. 1 for both vectors. First, there is no integration within centromeres; although the sequences of mammalian centromeres are highly repetitive and therefore not fully sequenced, the lack of integration within centromeres was not due to lack of sequence data in the database, as a similar figure generated using the first 460 sites from the random dataset resulted in multiple integrations within centromeres (data not shown). Second, there is a visible association between integration sites and regions of high gene density and transcription. The following sections describe the influence each parameter has on integration site selection.

Primary Sequence at Point of Integration

Primary sequence is thought to play only a minor role on integration site selection (reviewed in [18]). However, a weak palindromic sequence is often found at the point of integration for many retroviruses [19]. We determined the base composition 10 bp upstream and downstream of each integration site for the *in vitro* datasets and compared it to the *in silico* dataset (Figs. 2A and 2B). Any base that was substantially overrepresented (more than 10% greater than was predicted from the random dataset) is highlighted in green, while any base that was underrepresented (less than 10% than was predicted from the random dataset) is highlighted in orange. A weakly palindromic sequence was found centered around +3 from the integration site for HIV-1 and is broadly in agreement with recently published data for HIV-1 *in vivo* and *in vitro,* although in the previous publication G was not disfavored at +3 [19]. We now show that EIAV also exhibits a weak palindromic sequence centered on +3. The palindromic sequences showing the most favored nucleotides between the −1 and the +6 position are shown in Fig. 2C.

We characterized the 5′ -LTR-genomic junctions and 3′ -LTR-genomic junctions for six individual integration events to determine the size of the genomic repeat sequence generated by EIAV integration. As was the case for HIV-1, EIAV integration generated a 5-bp repeat

doi:10.1016/j.ymthe.2006.06.006

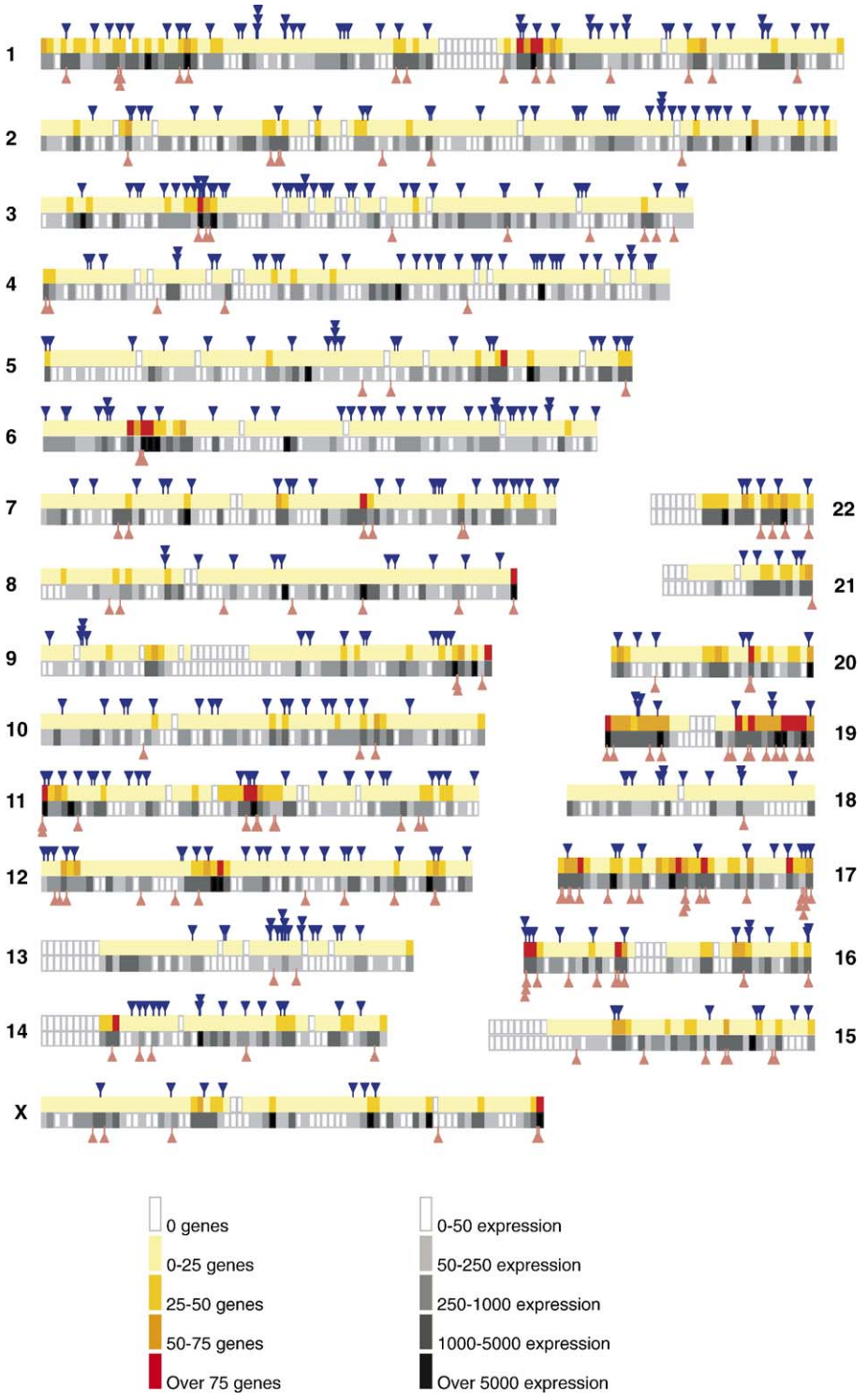


FIG. 1. Locations of integration sites in relation to gene density and transcriptional intensity in the human genome. Each shaded block represents a 2-Mb region of the human genome. The gene density and expression levels of all genes in the 2-Mb region were determined and the block was shaded appropriately. The HIV-1 integration sites are shown as pink triangles and EIAV integration sites are shown as blue triangles.

A

EIAV base composition at integration site (-10 to +10)																			
Base	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5	+6	+7	+8	+9	+10
A	13	1	-2	0	8	6	4	23	5	7	0	1	12	16	-16	-8	7	3	3	-1
C	-6	-4	0	-4	-4	-3	-6	-15	-6	-4	9	-9	-10	-10	13	11	1	-12	-1	0
G	-9	-3	-5	0	-2	-4	-4	-11	-4	3	9	-11	-11	-11	5	-7	-7	-17	-7	-10
T	2	6	8	3	-2	1	6	4	4	-6	-18	20	9	4	-2	5	-1	26	4	12

B

HIV base composition at integration site (-10 to +10)																				
Base	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5	+6	+7	+8	+9	+10
A	6	-2	-3	-2	1	3	-7	-10	0	-5	-2	-9	5	7	-20	-6	6	16	6	-7
C	-4	0	3	1	-5	-5	6	-10	-7	-2	1	-3	2	-1	23	17	6	1	1	4
G	-3	-1	-7	3	5	9	2	-2	3	7	22	0	-10	-5	-1	-6	-7	-8	6	5
T	1	4	7	-2	-1	-7	0	22	4	0	-20	12	2	-2	-2	-5	-5	-9	-13	-2

C

Vector	Most frequent base at integration site						
	-1	+1	+2	+3	+4	+5	+6
HIV-1 (19)	G	G	T	t	A	C	C
SIV (19)	G	G	T	T	A	C	C
MLV (19)	g	g	T	A	g	g/c	A
ASLV (19)	a	G	t	c	c	g	C
HIV-1 (this study)	g	G	T	a	a	C	C
EIAV (this study)	a	c/g	T	A	A	C	C

FIG. 2. Base composition surrounding EIAV and HIV-1 vector integration sites. (A) The base composition of the sequence 10 bp upstream (−10) and 10 bp downstream (+10) of each insertion site was determined for the EIAV dataset and compared to the random dataset. The difference between the two datasets was determined and recorded in the table. (B) The base composition of the sequence 10 bp upstream (−10) and 10 bp downstream (+10) of each insertion site was determined for the HIV-1 dataset and compared to the random dataset. Any base that appeared in a position 10% more often than predicted from the random dataset was highlighted in green, while any base that appeared in a position 10% less often than predicted from the random dataset was highlighted in orange. (C) A weak palindromic sequence centered on base +3 was identified for EIAV. This palindrome was compared to those previously described for other retroviruses and retroviral vectors.

sequence and the invariant CA sequence located 2 bp inside the termini of retroviruses was maintained [19,20].

EIAV Integration Targets Transcription Units and Gene-Dense Regions

We determined the number of integration sites within a transcription unit as defined by RefSeq [21] for each dataset and compared it to other retroviral integration profiles described in the literature (Table 1). In this study, the HIV-1 dataset (n = 162) showed a significant preference for integration within genes (72%), which is in line with data previously reported by Schroder *et al.* [10] and Wu *et al.* [14] and for the lentivirus SIV [12] (Table 1). In this study, EIAV (n = 458) was also shown to have a significant preference for integration within genes (68%), as described for other lentiviruses (Table 1). In addition, there was a positive correlation between gene density and HIV-1 and EIAV integration. An increase in the number of genes by 1 within a 2-Mb window resulted in an increase in the relative probability of integration of 3.9% for HIV-1 (95% CI (3.4%, 4.5%), $P \ll 0.01\%$) and 1.1% for EIAV (95% CI (0.7%, 1.6%), $P \ll 0.01\%$) (Fig. 3D).

GC Content of Integrated Region

GC content was considered over a 20-kb window around each integration site. Analysis using the univariate model showed that a 1% absolute increase in GC content in this region resulted in an increase in the probability of HIV-1 integration by 8.7% (95% CI (6.2%, 11.3%), $P \ll 0.01\%$) (Fig. 3A). However, a recent report investigating the influence of GC content on integration site selection by HIV-1 showed a weak negative correlation [22]. Other reports have shown that HIV-1 integration positively correlates with GC content [11,18]. The multivariate analysis will further resolve the interaction between specific parameters and is considered at the end of this section. For EIAV, an absolute increase in GC content of 1% resulted in a decrease in the relative probability of integration by 6.2% (95% CI (4.3%, 8.1%), $P \ll 0.01\%$) (Fig. 3A).

Integration Frequency within 2 kb Upstream of the Transcription Start Site

To assess the frequency of integration within the presumed promoter regions of genes, we assessed the number of integrations within a window located between 2 kb upstream of a gene and the transcription

doi:10.1016/j.ymthe.2006.06.006

TABLE 1: Integration profile of retroviral and lentiviral vectors within genes and nongenes, as defined by RefSeq [21]

Virus/vector (cell type, number of sites) and reference	Integration (%) Within a gene	Not within a gene
EIAV vector (HEK293T, n = 458)[a]	68	32
HIV-1 vector (HEK293T, n = 162)[a]	71	29
Random (n = 7861)[a]	36	64
HIV-1 and HIV vector (SupT1, n = 524) [10]	68	32
HIV-1 (H9/HeLa, n = 379) [14]	58	42
MLV (HeLa, n = 903) [14]	34	66
MLV vector (peripheral blood (PB) $CD34^+$ cells, n = 432) [26]	49	51
SIV (CEMx174, n = 148) [12]	74	26
SIV vector (PB $CD34^+$ cells, n = 328) [26]	73	27
ASLV vector (HEK293-TVA, n = 469) [11]	57	43
ASV vector (HeLa, n = 226) [13]	42	58

[a] Data determined in this study are compared to data from previously published studies (see references, where appropriate).

start site using the univariate model (Fig. 3B). There was no significant increase in the probability of insertion within the 2-kb region for HIV-1 or EIAV. This is in agreement with data previously published for HIV-1 [14]. Both HIV-1 and EIAV are in contrast to the integration preference of MLV, which shows a highly significant preference to the region 5 kb upstream of genes, with 11.2% of sites falling within this region compared to 2.1% of random sites [14]. The strong bias of MLV for integration within promoter regions in primary hematopoietic cells was also demonstrated using a promoter trapping method dependent on integration within the proximity of a cellular promoter. The promoter trapping efficiency (determined as the ratio between gene expression and integration) was significantly higher by between four- and fivefold for the MLV trap compared to the HIV-1 trap [23].

Integration Frequency within 10 kb of a CpG Island

CpG islands are enriched in the rare dinucleotide CG and are often associated with gene regulatory regions such as promoters [24]. The univariate model showed that being within 10 kb of a CpG island increased the relative probability of HIV-1 integration by a factor of 3.16 (95% CI (2.27%, 4.40%), $P \ll 0.01\%$). This was not expected, as previous data have shown that HIV-1 does not favor integration within ±1 kb of a CpG island [14]. In addition, our analysis revealed no significant favoring of HIV-1 insertion within the 2-kb promoter region, which is positively associated with CpG islands. The reason for the disparity between our results and those of Wu *et al.* may be due to the larger window considered in our analysis compared to the previous study (10 kb versus 1 kb). As CpG islands are also associated with gene-rich regions, the favoring of HIV-1 integration within 10 kb of a CpG island may be a result of other parameters such as gene density. This is considered in the multivariate analysis at the end of this section. The presence of a CpG island within 10 kb of an integration site does not increase the relative probability of EIAV integration (Fig. 3C). This is in contrast to MLV, which significantly favors CpG islands, with 16.8% of integrations found within ±1 kb of a CpG island compared to 2.1% of the randomly generated sites [14].

Integration within Individual Chromosomes

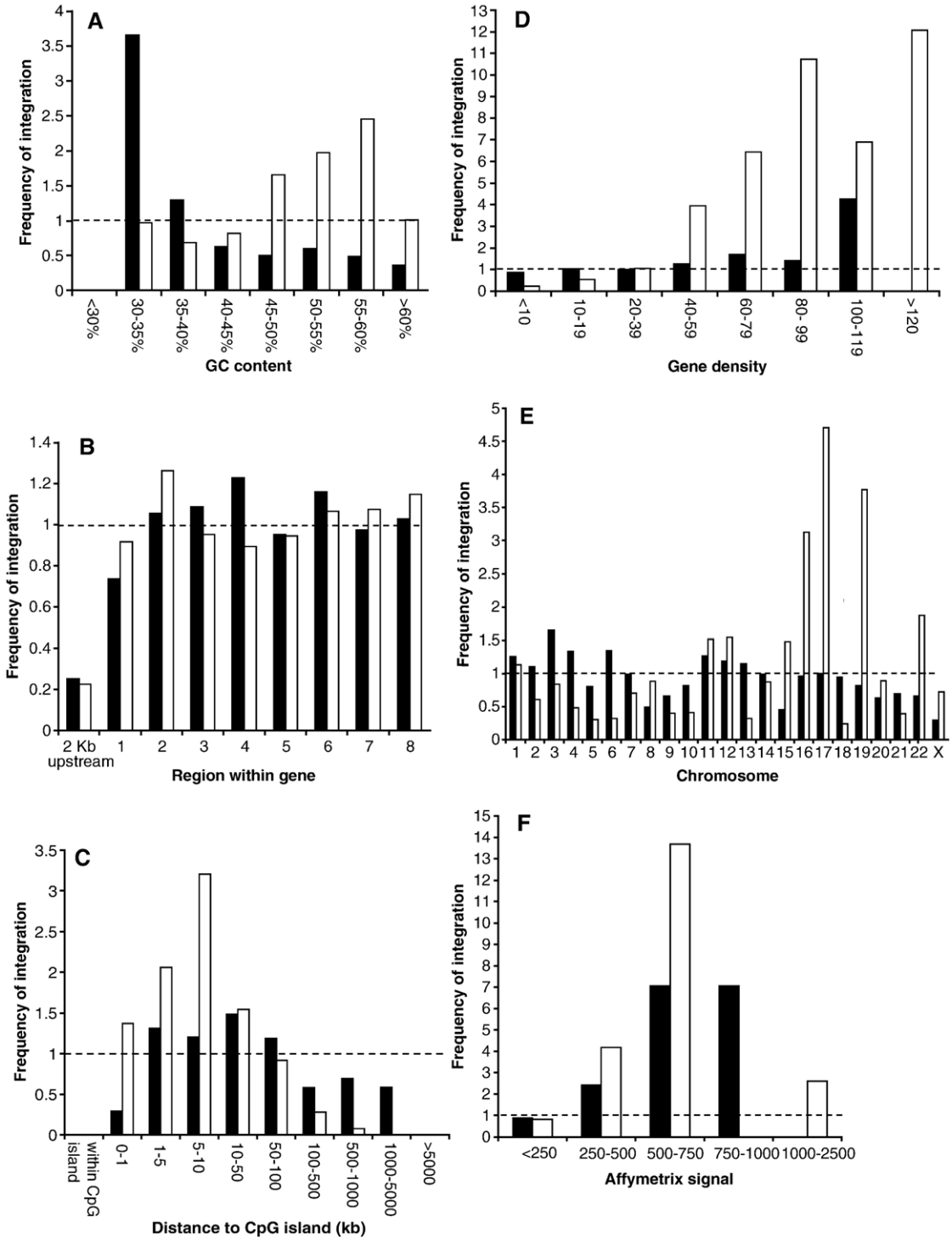
We examined the integration profile of EIAV and HIV-1 to determine whether specific chromosomes were favored or if any hot spots of integration were observed. The frequency of integration within each chromosome was roughly proportional to chromosome length for both vectors, but with some notable exceptions (Fig. 3E) that will be considered further at the end of this section using the multivariate model.

We examined the HIV-1 and EIAV datasets and 458 sites from the random dataset to determine if any hot spots of integration occurred. Hot spots have previously been defined as three or more integrations within a 100-kb region and have been identified for HIV-1 [10]. We identified one hot spot containing three HIV-1 integrations, or 1.8% of all sites mapped (n = 162), on chromosome 16 at position 16p13.3; each of the three integrations observed was within a separate gene. The hot spot may therefore be a consequence of the preference of HIV to integrate within genes and the high gene density of this particular chromosomal region. We found no hot spots for the EIAV dataset (n = 458) or within an equal number of sites analyzed from the random dataset.

FIG. 3. Integration site distribution of EIAV and HIV-1 vectors in dividing HEK293T cells. EIAV bars are in black and HIV-1 in white. A frequency of integration of 1 (represented by the dotted line) corresponds to the same level of integration as the random data, i.e., this category is neither favored nor disfavored for integration. (A) Percentage GC content within a 20-kb window centered on the integration site. (B) Gene region targeted for integration. Genes were divided into eight equally sized regions (proportional to gene length) and the number of integrations per region was determined for each dataset. In addition, the number of integrations in a region 2 kb upstream of the transcription start site was determined for each dataset. (C) Distance of the integration site from the nearest CpG island. (D) Number of genes within a 2-Mb window centered on the integration site. (E) Number of integrations per chromosome. (F) Expression level of gene targeted for integration. RNA from HEK293T cells was harvested and assayed on Affymetrix HG-Focus chips, which contain over 8500 human genes. The expression level of genes targeted for integration was determined and classified into five groups (<250, 250–500, 500–750, 750–1000, 1000–2500) and the number of integrations within each group determined.

Multivariate Analysis of All Sites

As many of the above parameters are influenced by each other, and therefore a relationship with an individual parameter may be heavily influenced by other parameters, we carried out a multivariate analysis to determine the relative contribution of each factor to integration frequency. Comparing the multivariate model to the univariate model for HIV-1, the influence of each



doi:10.1016/j.ymthe.2006.06.006

parameter changed quite substantially. The influence of the parameter within 10 kb of a CpG island decreased from 3.16 to 1.81% (95% CI (1.17%, 2.79%), $P \ll 0.1\%$), while the influence of gene density increased from 3.9 to 4.1% (95% CI (3.3%, 4.8%), $P \ll 0.1\%$). This suggests that the positive influence of CpG islands in the univariate model was most likely due to the influence of another factor or factors such as gene density, as an increase in gene density in a specific region results in an increase in promoter regions and thus CpG islands.

In addition, being within a 2-kb promoter region did not significantly influence integration when considered in the univariate model, but in the multivariate model, being within the 2-kb promoter region decreased the probability of integration, although this was not statistically significant. As 2-kb promoter regions are, by definition, associated with genes and given that gene-dense regions will possess a high proportion of the genomic promoter regions, it may be that the disfavoring of integration in the 2-kb promoter region was masked in the univariate model.

An absolute increase in GC content of 1% increased the probability of HIV-1 integration by 8.7% in the univariate model. However, in the multivariate model, a 1% absolute increase in GC content decreased the relative probability of integration by 5.9% (95% CI (2.5%, 9.2%), $P < 0.1\%$). The initial observations from the univariate model indicated a favoring of high GC content, but this may have been a function of gene density rather than GC content as previously described. Once all the parameters were taken into consideration in the multivariate model for HIV-1, the influence of other parameters revealed a preference for AT-rich regions. We hypothesize that this outcome was due to the HIV-1 vector-targeted genes having a lower GC content than the average gene (as represented by the random control dataset).

The univariate model for HIV-1 indicated a higher probability of integration within chromosomes 16, 17, and 19 and a lower probability of integration which was not significant in chromosomes 5 and 6. However, when considering the multivariate model, only integration within chromosome 17 was favored, and chromosomes 5 and 6 had a slightly lower probability of integration which was not significant ($P < 2\%$). This suggests that factors other than gene density are influencing integration in these chromosomes. The favoring of integration into chromosomes 17 and 19 and also to specific regions of chromosomes 6, 13, and 16 has previously been described for an MLV-based retroviral vector. However, regression analysis was not carried out in that particular study so the preference observed may have been due to the influence of gene density rather than chromosome number per se [25].

The multivariate model generated for EIAV shows that an increase in gene density increased the relative probability of integration from 1.1 to 2.3% (95% CI (1.7%, 2.9%), $P < 0.01\%$). This was not surprising as preferential integration of retroviruses and lentiviruses in genes and gene-rich regions has been well documented in the literature [10,11,18]. As was the case in the univariate model for EIAV, an increase in the local GC content led to a decrease in the relative probability of integration; an absolute increase in 1% GC content resulted in a decrease in the relative probability of EIAV integration by 13.5% (95% CI (11.1%, 15.9%), $P \ll 0.01\%$). All these factors taken together suggest that although EIAV preferentially integrates within gene-dense regions that are themselves GC rich, at the sequence level these regions are AT rich relative to the ''average'' gene and, as was the case for HIV-1, EIAV may prefer to target AT-rich regions. Within the more complete multivariate models examining the effects of GC content, HIV-1 and EIAV behaved similarly in that both vectors showed decreased integration within regions of high GC content.

Being within 10 kb of a CpG island did not affect integration of EIAV according to the univariate model. However, the multivariate model indicated that the presence of a CpG island within 10 kb increased the relative probability of integration by a factor of 1.78 (95% CI (1.31%, 2.42%), $P < 0.02\%$). This observation was not expected and suggests that integration within 10 kb of a CpG island is favored by EIAV, even after allowing for the influence of other parameters such as gene density. The apparent influence of CpG islands may be due to other factors that were not considered in this study, but crucially the influence of CpG islands on HIV-1 and EIAV vector integration was similar in the multivariate model.

The univariate model for EIAV suggests that there is a lower probability of integration in chromosomes 5, 8, 15, and X. However, when considering the multivariate model there was only weak evidence for a lower probability of integration on chromosomes 8 and X ($P < 10\%$) and no evidence for decreased integration on chromosomes 5 and 15. This suggests that factors other than low gene density are influencing integration on chromosome 8. The lower probability of integration on chromosome X may be because HEK293T cells appear to be female, as no insertions of HIV-1 or EIAV occurred on the Y chromosome, and in female cells one copy of the X chromosome is normally inactivated. In addition, we note that the karyotype of cultured cells is notoriously difficult to define. Cells are often aneuploid, with individual chromosomes over- and underrepresented, and this may have influenced the frequency of integration within individual chromosomes.

Analysis of Sites within Genes Only

As integration of EIAV and HIV-1 was favored within genes, we carried out further analysis to determine the influence of parameters specific to genes including (a) gene expression and (b) gene region.

Effect of gene expression on integration. We measured the level of expression of genes possessing an integration site from all the datasets and divided it into five categories. We normalized the number of genes in each category from each *in vitro* dataset to the number of genes from the random dataset in each category (Fig. 3F).

The univariate model showed that the probability of HIV-1 integration increased as the expression level increased (95% CI (0.437%, 0.798%), $P \ll 0.01\%$). EIAV also showed a preference for integration within active genes (95% CI (0.330%, 0.591%), $P \ll 0.01\%$), although we found no EIAV integration sites at the very highest level of expression, but this may be due to the limited dataset (expression between 1000 and 2500 units; see Fig. 3F). The favoring of lentivirus integration within active genes over inactive genes was previously described for HIV-1 and SIV [10,12].

Effect of gene region on integration. We further analyzed integration sites within genes to determine whether a particular region of the transcription unit was favored. The gammaretrovirus MLV significantly favors integration around the transcription start site (TSS), with 20.2% of all sites falling within ±5 kb of the TSS in HeLa cells ($n = 903$, $P < 0.0001$) [14] and 11% of integration sites falling within ±2kb of a TSS in peripheral blood $CD34^+$ cells ($n = 432$, $P < 0.0001$) [26]. Previous studies indicated that HIV-1 did not exhibit a preference for integration with close proximity to the TSS [11,14]. The lentivirus SIV also did not exhibit a preference for a specific gene region or for the region 5 kb upstream of the TSS in dividing CEMx174 cells *in vitro* [12] or in primate peripheral blood $CD34^+$ cells *in vivo* [26]. The results of this study are in agreement with the published observations and indicate a lack of preference for gene region or the 2kb region upstream of the TSS by HIV-1 vectors (Fig. 3B). In the univariate model, EIAV integration showed no significant preference for a gene region and thus behaves in a manner similar to that of the other lentiviruses for which integration site preference has been characterized.

Multivariate Analysis of Sites within Genes

The level of gene expression may govern the site of integration within a gene and thus the gene-specific parameters of gene region and gene expression may be counterdependent. We carried out a multivariate analysis to determine the contribution of each of these parameters to integration. The multivariate model for both vectors showed that the parameters gene region and gene expression were independent of each other (data not shown).

Discussion

EIAV vectors integrated preferentially into active genes as previously described for HIV-1 and SIV [10–12]. Integration correlated with gene-dense regions and regions of high expression. This integration bias differs from the gammaretroviruses (MLV and ASLV), which have only a weak preference for active genes [11]. HIV-1 and EIAV vectors showed no preference for a specific region of the transcription unit or for the 2-kb promoter region, while both preferentially integrated into AT-rich regions. This characteristic differs from that of the gammaretrovirus MLV, which favors integration within the promoter region of the transcription unit [14]. Furthermore the site of integration of retroviral vectors appears to be weakly palindromic, and we have shown some sequence similarity within this region between EIAV and HIV-1.

HIV-1 and EIAV are evolutionarily divergent within the lentiviruses as they share limited nucleotide sequence similarity, yet the vectors show comparable integration site preferences. This may indicate that the mechanism of integration and the various cellular proteins with which the preintegration complex interacts are potentially conserved within all lentiviruses. The gammaretroviruses are in a separate genus of the *Retroviridae* and it is therefore possible that they have evolved different mechanisms of interaction with the host cell genome and cellular factors. Our data from EIAV-based vectors showing conserved features therefore consolidate the conclusions about conservation among lentiviruses and divergence from gammaretroviruses drawn in other studies [10,12,26]. Several cellular DNA-binding proteins have been shown to interact with HIV-1, including barrier-to-autointegration factor (BAF), HMGa1, Ini-1, Ku, and LEDGF/p75 [16]. BAF binds directly to linker histone H1.1 and core histone H3 both *in vitro* and *in vivo.* Histone H1.1 associates preferentially with open/active chromatin. The interactions between BAF and HIV-1 and BAF and histones H1.1 and H3 may explain the positioning of HIV-1 within active chromatin and thus active genes [27]. In addition, knockdown of LEDGF/p75 in three cell lines was shown to reduce partially integration of HIV-1 within transcription units. Knockdown of LEDGF/p75 resulted in an increase in HIV-1 integration within GC-rich regions, suggesting that LEDGF targets HIV-1 integration either directly or indirectly to AT-rich regions [22]. Based on these observations, it can be postulated that the EIAV preintegration complex interacts with a cellular binding factor, analogous to the characterized HIV-1 integrase–LEDGF interaction. Further comparative analyses between HIV-1, EIAV, and MLV and their interactions with cellular factors should provide insight into the reasons for the shared integration profiles among lentiviral vectors and the differences from MLV.

Given the propensity of lentiviral vectors to integrate into actively transcribed genes, the gene expression profile of the target cell could influence integration site selection. This site selection could vary according to the cell type and/or the proliferation status. One study indicated a greater preference for HIV-1 vector integration within genes rather than intergenic regions of a growth-arrested cell, compared to cells undergoing proliferation, and it was

doi:10.1016/j.ymthe.2006.06.006

postulated that this might reflect the compaction of chromatin in the intergenic regions [28]. Further analysis of the integration profiles in a range of cell types and cell states is required to confirm whether there are any statistically significant differences between the profiles. We note that our study analyzed the integration profile within cultured cells that may be aneuploid, and therefore further profiling using primary cells is warranted, although this is beyond the scope of this study.

We have now added to an increasing body of data that indicates that there is a highly conserved set of integration features between all lentiviral vectors studied to date (HIV-1, SIV, and now EIAV). It is therefore possible that a conserved mechanism of integration exists for all the lentiviruses. Furthermore the integration profiles of all of these lentiviral vectors are significantly different from that observed with the gammaretrovirus MLV because of their increased bias to integrate within active genes and lack of preference for promoter regions. Therefore, it can be reasoned that lentiviral vector integration may be less likely to lead to activation of normally quiescent genes compared to MLV vector integration.

MATERIALS AND METHODS

Cells. HEK293T cells were maintained in Dulbecco's modified Eagle's medium (Sigma Chemical Co., UK) supplemented with 10% (v/v) fetal bovine serum (Sigma), 1× nonessential amino acids (Sigma), and 2 mM glutamine at 37°C.

Vector production. Vectors were produced via a three-plasmid cotransfection with plasmids encoding gag/pol, VSV-G envelope, and genome as previously described ([4], F. J. Wilkes, manuscript in preparation). EIAV vectors were produced by transfecting HEK293T cells with the codon-optimized gag/pol-encoding plasmid pESYNGP [29], the VSV-G envelope-encoding plasmid pRV67, and either the transfer vector plasmid pONYKZ or the ProSavin vector genome plasmid pONYK1-ORT [30]. The HIV-1 vector pHF2G was produced by transfecting HEK293T cells with the gag/pol-expressing plasmid pSYNGP [31], the VSV-G envelope-expressing plasmid pHCMVG, and the vector genome pHF2G [32]. The supernatant was harvested and the vector titer determined by adding serial dilutions of vector onto HEK293T cells and after several passages assessing vector copy number using Q-PCR.

Cellular transduction and template DNA preparation. HEK293T cells were seeded at a density of 3×10^5 cells/well in a 6-well plate. Vector was then added at an m.o.i. of between 1.7 and 6.6 in a total of 600 μl fully supplemented medium including Polybrene (8 ng/ml). Cells were incubated at 37°C for 3 days. At this stage, a small proportion of the cell population was seeded at between 5 and 20 cells per well in a 96-well plate for partial clonal enrichment. These and the remaining pool of cells were maintained in culture until they reached confluence. The clonally enriched cells were transferred to a 24-well plate and amplified; cellular DNA was extracted using the MagNA Pure LC DNA Isolation Kit I (Roche).

LM-PCR. LM-PCR has been described previously [8] and was adapted using vector-specific primers to work with either the HIV-1 or the EIAV vector system, each using the same restriction endonuclease (*Nla*III). LM-PCR was carried out on 1 μg extracted DNA. DNA was digested with 4 U of *Nla*III (New England Biolabs) for 1 h at 37°C and then purified by ethanol precipitation. The restriction-digested DNA was then resuspended in the primer extension reaction mixture (2.5 U native *Pfu* DNA polymerase (Stratagene), 1× final concentration reaction buffer (Stratagene), 200 μM dNTPs (Invitrogen), 0.25 pmol biotin extension primer (E-LTR-1, 5′-biotin-GGGCACTCAGATTCTGCGGT-3′, for vector pONYKZ or pONYK1-ORT; H-LTR-1, 5′-biotin-GAGCTCTCTGGCTAACTAGG-3′ for vector pHF2G)) and the extension was carried out under the following conditions: 95°C for 5 min, 64°C for 30 min, 72°C for 15 min. The extension product was then purified by passing through a QIAquick PCR Purification Kit and resuspended in 40 μl nuclease-free water (Gibco). Dynabeads M-280 streptavidin (Dynal Biotech Ltd.; 200 μg) were added to a clean 1.5-ml microcentrifuge tube and washed twice in 40 μl of 2× B&W buffer (10 mM Tris–HCl, pH 7.5 (Sigma), 1 mM EDTA (Sigma), 2.0 M NaCl (Sigma)) and then resuspended in 40 μl 2× B&W buffer. The purified cDNA was then added to the bead solution and mixed by rotating on a shaking platform for between 1.5 and 5 h at room temperature. The beads and associated cDNA were collected using a magnet and the supernatant was discarded. The beads were then washed twice with 100 μl of nuclease-free water (Gibco). The cDNA was then ligated to the double-stranded blunt-ended cassette as previously described [8]. The cassette was prepared by the addition of 100 pmol oligonucleotide LM-OK1 (5′-GACCCGGGA-GATCTGAATTCAGTGGCACAGCAGTTAGG-3′) and 100 pmol oligonucleotide LM-OK2 (5′-CCTAACTGCTGTGCCACTGAATTCAGATCTCCCG-3′) to a solution of 250 mM Tris (pH 7.5) and 100 mM $MgCl_2$ in a total volume of 200 μl. The solution was heated to 95°C for 5 min and then left to cool. The cassette was concentrated by passing through a Microcon YM-30 30 kDa column (Millipore) and resuspended in 80 μl nuclease-free water. To the cDNA, 1 μl of concentrated cassette was added with 80 U T4 DNA ligase (NEB) and 1× ligation buffer (NEB) in a total volume of 10 μl; the reaction was incubated overnight at 16°C. The biotin-labeled cDNA was then selected and washed twice, as before. The beads were then resuspended in 10 μl nuclease-free water. For the first exponential PCR, 1 μl of the ligation product was added to 1× final concentration of Extensor Hi-Fidelity PCR Master Mix (AB gene), 10 pmol vector-specific primer (E-LTR-2, 5′-CTGAGTCCCTTCTCTGCTGG-3′ for vector template pONYKZ and pONYK1-ORT; H-LTR-2, 5′-CCCACTGCTTAAGCCTCAAT-3′, for vector template pHF2G) and 10 pmol cassette-specific primer OK1 (5′-GACCCGGGAGATCTGAATTC-3′) in a total of 25 μl. The cycling conditions were as follows: 94°C for 2 min; 94°C for 15 s, 60°C for 30 s, 68°C for 2 min for 30 cycles; 68°C for 10 min. A nested PCR was then carried out using 1 μl of a 1:100 dilution of the first PCR product as template. The conditions for the nested PCR were the same as the first PCR; however, the nested vector-specific primer (E-LTR-3, 5′-GGGCTGAAAAGGCCTTTGTA-3′, for vector template pONYKZ and pONYK1-ORT or H-LTR-3, 5′-AGCTTGCCTTGAGTGCTTCA-3′ for vector template pHF2G) and nested cassette-specific primer OK2 (5′-AGTGGCA-CAGCAGTTAGG-3′) were used. The amplified PCR products were visualized on a 2% (w/v) agarose gel. As an *Nla*III site is present internally in each vector, a common internal vector control band was amplified in each reaction (EIAV vector 665 bp; HIV-1 2161 bp) and thus acted as an internal positive control. The remaining 20 μl of the second PCR was purified using a PCR cleanup kit (Qiagen). The purified products were ligated into TOPO-TA vector (Invitrogen), before being transformed into TOP10 cells (Invitrogen). Cells were plated out onto ampicillin plates with 40 μl X-gal (2%) and incubated overnight at 37°C. White colonies were selectively isolated and used to inoculate 1 well of a 96-well plate containing approximately 100 μl of ampicillin agar. The 96-well plate was sent to GATC (GATC Biotech AG) for PCR product insert sequencing.

Sequence analysis. The sequence of the PCR product of the clones was further analyzed to determine whether they possessed the entire predicted LTR region as well as the ligated linker cassette sequence. If the clones met these criteria, the sequence of the genomic DNA between the LTR and the cassette was submitted into the UCSC human genome database (May 2004 assembly) at http://www.genome.ucsc.edu/cgi-bin/hgBlat. The chromosomal location of the DNA sequence was therefore mapped. A profile of integration was thus determined with respect to gene locality and chromosome number. This profile was compared to a random profile of 7861 integrations generated from the *in silico* dataset as described in the following section.

Generation of random dataset. To determine if there was a statistically significant bias for integration site selection *in vitro,* a comparative random set

of 10,000 genomic positions was generated *in silico*. Using the Excel "RAND()" function, a set of 10,000 random numbers was generated between 1 and 3,019,080,196, which is the size of the haploid female genome (see http://www.genome.ucsc.edu/goldenPath/stats.html#hg17). These random numbers were then converted into chromosomal positions as described below. Chromosomes were placed end to end in numerical order so that chromosome 1 was the first chromosome in the sequence and chromosome X the last chromosome in the sequence. Numbers were then attributed to each base pair within the genome so that No. 1 is the first base pair of chromosome 1 and No. 245,522,848 is the first base pair of chromosome 2 and so forth, with 3,019,080,196 corresponding to the last base pair of chromosome X. Each of the randomly generated numbers was therefore converted to a chromosomal position. In addition, the orientation of the position was randomly assigned to the positive or negative strand. Once chromosomal positions were assigned, they were used to search and extract 1 kb of downstream sequence (as this was considered the maximum amount of sequence that could be amplified *in vitro* during the proviral–genomic junction enrichment stage; this was substantiated experimentally) from the human genome database at the UCSC Web site. The extracted 1 kb of sequence was searched for *Nla*III sites and cut at the site closest to the integration point. At this stage, a proportion of the fragments were discarded as no *Nla*III site was located within 1 kb of the random site. This is a true reflection of the *in vivo* situation (fragments that do not possess an *Nla*III site would not be digested and therefore would not be mapped). The "digested" fragments were re-inputted into the human genome database and the following information was collated: the position of integration, whether the integration site was within a gene/intergenic region, the Blat score, the returned size relative to the query size, the percentage identity, and the span. This process of re-inputting sequence into Blat enabled us to account for the various limitations of Blat. We were able to account for "false positives"—sites that were assigned a different position in the random data following a second Blat search. In addition "false negatives" were accounted for; these occurred when the original randomly generated position matched the sequence returned after digestion but other potential chromosomal positions were also identified with a close sequence match. The false positive and false negative sequences were discounted from the random dataset. This process accounted for repeat regions and reflects what would happen with the *in vitro* dataset if a repeat region was targeted, i.e., the exact chromosomal position could not be determined. Sequences that were fewer than 18 nucleotides in length following the predicted *Nla*III "digestion" were also discarded, as these sequences were too small to lead to an unambiguous mapped position on the chromosome. After this process, 7861 random sites remained. This method for generating an *in silico* dataset was similar to that used by Miller *et al*. [33] in a large-scale analysis of adeno-associated virus vector integration sites.

Determination of HEK293T transcriptional activity. Cellular RNA was extracted from HEK293T cells and labeled as described by Affymetrix (Santa Clara, CA, USA). Five micrograms of labeled cRNA was used per Affymetrix HG-focus array chip, on which over 8500 of the best characterized human genes are represented. The data obtained using three replicate chips were collated and the gene expression for a given gene was calculated from the average of the three readings.

Statistical analysis. The univariate and multivariate logistic regressions were generated using SAS version 8.2.

RECEIVED FOR PUBLICATION APRIL 13, 2006; REVISED JUNE 9, 2006; ACCEPTED JUNE 18, 2006.

REFERENCES

1. Coffin, J. M., Hughes, S. H., and Varmus, H. E. (1997). *Retroviruses*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
2. Kim, V. N., Mitrophanous, K., Kingsman, S. M., and Kingsman, A. J. (1998). Minimal requirement for a lentivirus vector based on human immunodeficiency virus type 1. *J. Virol.* **72:** 811–816.
3. Naldini, L., *et al.* (1996). In vivo gene delivery and stable transduction of nondividing cells by a lentiviral vector. *Science* **272:** 263–267.
4. Mitrophanous, K., *et al.* (1999). Stable gene transfer to the nervous system using a non-primate lentiviral vector. *Gene Ther.* **6:** 1808–1818.
5. Olsen, J. C. (1998). Gene transfer vectors derived from equine infectious anemia virus. *Gene Ther.* **5:** 1481–1487.
6. Azzouz, M., *et al.* (2002). Multicistronic lentiviral vector-mediated striatal gene transfer of aromatic l-amino acid decarboxylase, tyrosine hydroxylase, and GTP cyclohydrolase I induces sustained transgene expression, dopamine production, and functional improvement in a rat model of Parkinson's disease. *J. Neurosci.* **22:** 10302–10312.
7. Bienemann, A. S., *et al.* (2003). Long-term replacement of a mutated nonfunctional CNS gene: reversal of hypothalamic diabetes insipidus using an EIAV-based lentiviral vector expressing arginine vasopressin. *Mol. Ther.* **7:** 588–596.
8. Schmidt, M., *et al.* (2001). Detection and direct genomic sequencing of multiple rare unknown flanking DNA in highly complex samples. *Hum. Gene Ther.* **12:** 743–749.
9. Schmidt, M., *et al.* (2002). Polyclonal long-term repopulating stem cell clones in a primate model. *Blood* **100:** 2737–2743.
10. Schroder, A. R., Shinn, P., Chen, H., Berry, C., Ecker, J. R., and Bushman, F. (2002). HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110:** 521–529.
11. Mitchell, R. S., *et al.* (2004). Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol.* **2:** e234.
12. Crise, B., *et al.* (2005). Simian immunodeficiency virus integration preference is similar to that of human immunodeficiency virus type 1. *J. Virol.* **79:** 12199–12204.
13. Narezkina, A., *et al.* (2004). Genome-wide analyses of avian sarcoma virus integration sites. *J. Virol.* **78:** 11656–11663.
14. Wu, X., Li, Y., Crise, B., and Burgess, S. M. (2003). Transcription start regions in the human genome are favored targets for MLV integration. *Science* **300:** 1749–1751.
15. Bushman, F., *et al.* (2005). Genome-wide analysis of retroviral DNA integration. *Nat. Rev. Microbiol.* **3:** 848–858.
16. Bushman, F. D. (2003). Targeting survival: integration site selection by retroviruses and LTR-retrotransposons. *Cell* **115:** 135–138.
17. Karolchik, D., *et al.* (2003). The UCSC genome browser database. *Nucleic Acids Res.* **31:** 51–54.
18. Lewinski, M. K., and Bushman, F. D. (2005). Retroviral DNA integration—mechanism and consequences. *Adv. Genet.* **55:** 147–181.
19. Wu, X., Li, Y., Crise, B., Burgess, S. M., and Munroe, D. J. (2005). Weak palindromic consensus sequences are a common feature found at the integration target sites of many retroviruses. *J. Virol.* **79:** 5211–5214.
20. Katzman, M., and Sudol, M. (1996). Influence of subterminal viral DNA nucleotides on differential susceptibility to cleavage by human immunodeficiency virus type 1 and visna virus integrases. *J. Virol.* **70:** 9069–9073.
21. Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. (2005). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **33:** D54–D58.
22. Ciuffi, A., *et al.* (2005). A role for LEDGF/p75 in targeting HIV DNA integration. *Nat. Med.* **11:** 1287–1289.
23. De Palma, M., *et al.* (2005). Promoter trapping reveals significant differences in integration site selection between MLV and HIV vectors in primary hematopoietic cells. *Blood* **105:** 2307–2315.
24. Bird, A. P. (1986). CpG-rich islands and the function of DNA methylation. *Nature* **321:** 209–213.
25. Laufs, S., Nagy, K. Z., Giordano, F. A., Hotz-Wagenblatt, A., Zeller, W. J., and Fruehauf, S. (2004). Insertion of retroviral vectors in NOD/SCID repopulating human peripheral blood progenitor cells occurs preferentially in the vicinity of transcription start regions and in introns. *Mol. Ther.* **10:** 874–881.
26. Hematti, P., *et al.* (2004). Distinct genomic integration of MLV and SIV vectors in primate hematopoietic stem and progenitor cells. *PLoS Biol.* **2:** e423.
27. Montes de Oca, R., Lee, K. K., and Wilson, K. L. (2005). Binding of barrier-to-autointegration factor (BAF) to histone H3 and selected linker histones including H1.1. *J. Biol. Chem.*
28. Ciuffi, A., *et al.* (2006). Integration site selection by HIV-based vectors in dividing and growth-arrested IMR-90 lung fibroblasts. *Mol. Ther.* **13:** 366–373.
29. Rohll, J. B., *et al.* (2002). Design, production, safety, evaluation, and clinical applications of nonprimate lentiviral vectors. *Methods Enzymol.* **346:** 466–500.
30. Miskin, J., *et al.* (2006). A replication competent lentivirus (RCL) assay for equine infectious anaemia virus (EIAV)-based lentiviral vectors. *Gene Ther.* **13:** 196–205.
31. Kotsopoulou, E., Kim, V. N., Kingsman, A. J., Kingsman, S. M., and Mitrophanous, K. A. (2000). A Rev-independent human immunodeficiency virus type 1 (HIV-1)-based vector that exploits a codon-optimized HIV-1 gag–pol gene. *J. Virol.* **74:** 4839–4852.
32. Siapati, E. K., *et al.* (2005). Comparison of HIV-1 and EIAV-based vectors on their efficiency in transducing murine and human hematopoietic repopulating cells. *Mol. Ther.* **12:** 537–546.
33. Miller, D. G., Trobridge, G. D., Petek, L. M., Jacobs, M. A., Kaul, R., and Russell, D. W. (2005). Large-scale analysis of adeno-associated virus vector integration sites in normal human cells. *J. Virol.* **79:** 11434–11442.