Integration target site selection by a resurrected human endogenous retrovirus

Troy Brady,[1] Young Nam Lee,[2] Keshet Ronen,[1] Nirav Malani,[1] Charles C. Berry,[3] Paul D. Bieniasz,[2,4] and Frederic D. Bushman[1,5]

[1]Department of Microbiology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA; [2]Aaron Diamond AIDS Research Center, Rockefeller University, New York, New York 10065, USA; [3]Department of Family/Preventive Medicine, University of California, San Diego School of Medicine, San Diego, California 92093, USA; [4]Howard Hughes Medical Institute and Laboratory of Retrovirology, Rockefeller University, New York, New York 10065, USA

At least 8% of the human genome was formed by integration of retroviral DNA sequences. Here we analyze the forces directing the accumulation of human endogenous retroviruses (HERVs) by comparing de novo HERV integration targeting with the distribution of fixed HERV elements in the human genome. All known genomic HERVs are inactive due to mutation, but we were able to study integration targeting using a reconstituted consensus HERV-K (designated HERV-K_{Con}). We found that HERV-K_{Con} integrated preferentially in transcription units, in gene-rich regions, and near features associated with active transcription units and associated regulatory regions. In contrast, genomic HERV-K proviruses are found preferentially outside transcription units. The minority of genomic HERVKs present inside transcription units are in opposite transcriptional orientation relative to the host gene, the orientation predicted to be minimally disruptive to host mRNA synthesis, but de novo HERV-K_{Con} integration within transcription units showed no orientation bias. We also found that the youngest HERV-K elements in the human genome showed a distribution intermediate between de novo HERV-K_{Con} integration sites and older fixed HERV-Ks. These findings indicate that accumulation of HERVs in the human germline is a two-step process: integration targeting biases direct initial accumulation, then purifying selection leads to loss of proviruses disrupting gene function.

[*Keywords:* HERV-K; HML2; genome construction; integration; positive selection]

Supplemental material is available at http://www.genesdev.org.

Received November 12, 2008; revised version accepted January 22, 2009.

The genomes of most organisms are infested with repeat sequences derived from genomic parasites. These parasites, though selfish DNA (Dawkins 1976; Orgel and Crick 1980), create genetic variation that provides a substrate for natural selection, resulting in new gene formation, altered host gene expression, and facilitated recombination (Bushman 2001; Lander 2001; Craig et al. 2002). In humans, repeat elements constitute ~45% of the total genome. About 8% of these are contributed by the class of elements that replicate via an RNA intermediate and contain long terminal repeats (LTRs), which includes human endogenous retroviruses (HERVs) (Smit 1999; Lander 2001; Bannert and Kurth 2004). These HERVs resemble known exogenous retroviruses—Class I HERVs are most homologous to gammaretroviruses, Class II to betaretroviruses, and Class III to spumaretroviruses (Medstrand et al. 2002).

The Class II HERVs are a collection of viruses that are thought to use $tRNA^{Lys}$ to prime reverse transcription, and so are also known as HERV-K. Class II HERVs are divided into subfamilies from HML-1 through HML-10. These proviruses began to appear in the germline of Old World primates ~30–35 million years ago (Medstrand and Mager 1998; Barbulescu et al. 1999; Costas 2001). The HML-2 subfamily of HERV-Ks is the only HERV family in which some elements are polymorphic within the human population, suggesting integration into the germline after the divergence of humans and chimpanzees, perhaps as recently as a few hundred thousand years ago for some elements (Costas 2001; Turner et al. 2001; Belshaw et al. 2004, 2005; Hughes and Coffin 2004). Despite the evidence of recent activity, no replication-competent endogenous HERVs have been identified in the human genome so far.

The HERV elements that are fixed in the human genome show a different distribution than has been seen for integration by any previously studied exogenous retrovirus. Previous studies of retroviral integration have shown that different retroviral genera favor integration in different regions of vertebrate genomes (for review, see Bushman et al. 2005; Engelman 2005; Berry et al. 2006). HIV favors integration within active transcription units (Schroder et al. 2002; Wu et al. 2003; Barr et al. 2005; Ciuffi et al. 2005, 2006; Lewinski et al. 2005, 2006), MLV

[5]**Corresponding author.**
E-MAIL bushman@mail.med.upenn.edu; FAX (215) 573-4856.
Article is online at http://www.genesdev.org/cgi/doi/10.1101/gad.1762309.

favors integration at transcription start sites and CpG islands (Wu et al. 2003; Lewinski et al. 2006; Aiuti et al. 2007; Wang et al. 2007), and ASLV shows nearly random integration (Mitchell et al. 2004; Narezkina et al. 2004; Barr et al. 2005), only weakly favoring transcription units and CpG islands.

In contrast, HERVs are enriched outside transcription units, a pattern rarely seen for de novo retroviral integration. Moreover, the minority of HERV elements that are present within transcription units show a strong orientation bias, such that the viral genome is usually oriented opposite to the direction of host gene transcription (Smit 1999; Medstrand et al. 2002; van de Lagemaat et al. 2006). Some integrating elements are known to show strong orientation biases (e.g., phage lambda) (Hendrix et al. 1983), but no such orientation bias has been reported for retroviruses. Both of these trends in HERV positions could potentially be explained as resulting from purifying selection, in which deleterious HERV proviruses within transcription units were subject to negative selection at the organismic level and lost during evolution. According to this idea, the opposite transcriptional orientation for HERVs within genes is the result of selection for minimal disruption of mRNA synthesis, since the HERV splicing and polyA addition signals are present in antisense orientation and hence inactive (Mager 1999; Smit 1999; van de Lagemaat et al. 2006). However, it was also possible that these biases arose during initial HERV integration.

Because HERV-K is an extinct retrovirus, the mechanisms underlying its differential distribution in the human genome could not previously be addressed experimentally. However, two groups recently reconstituted active HERV-K elements (Dewannieux et al. 2006; Lee and Bieniasz 2007). Each group determined a consensus sequence for the youngest group of HERV elements, HERV-K (HML2), then synthesized a DNA copy. Transfection of DNA encoding the consensus HERV (termed "HERV-K_{Con}" in Lee and Bieniasz 2007) into cells yielded viral particles capable of infecting new cells. Using these reconstituted HERV-Ks, replication was shown to require reverse transcription and integration, the cell type specificity of infection was determined, and the response to restriction factors was analyzed (Esnault et al. 2008; Lee et al. 2008).

Here we investigated the forces dictating the accumulation of HERV elements in the human genome by analyzing integration target site selection by HERV-K_{Con}. Specifically, we report the isolation and sequencing of 1565 de novo integration junctions between HERV-K_{Con} proviruses and genomic DNA, and analysis of their placement on the human genome. This allowed us to assess the relationship between initial integration site selection by HERV-K and long-term accumulation of HERV-K sequences in the human genome.

Results

HERV-K_{Con} infection and integration site analysis

To generate integration sites for analysis, 293T and HT1080 cell lines were infected with HERV-K_{Con}. These cell lines were chosen based on an initial screen for cell lines that support efficient infection. The HERV-K_{Con} genome was engineered to express a GFP gene to allow convenient quantification of infection (Fig. 1A; Lee and Bieniasz 2007). In order to distinguish new integration events from pre-existing HERV-Ks present in the human genome, 53 base pairs (bp) of HIV LTR-derived sequence was inserted ~35 bp from the 5′-end of the HERV-K_{Con} U3 region (Fig. 1A), allowing selective PCR amplification of newly integrated HERV-K_{Con} proviruses. Typical infections yielded 10%–20% GFP-positive cells. Cells were harvested 2 d after infection, and genomic DNA was isolated. Integration site sequences were recovered by digesting genomic DNA with restriction enzymes, ligating linkers to the resulting fragment ends, and PCR-amplifying virus–host DNA junctions using LTR and linker-specific primers. Amplification products were subjected to pyrophosphate sequencing (Margulies et al. 2005), and sites of integration were mapped onto the human genome sequence (Lander 2001; Venter 2001). A total of 1565 unique integration sites were identified by this method (Table 1).

For statistical analysis, it is useful to compare the distribution of integration sites to randomly chosen genomic locations. We thus generated random controls in silico, but used a matching procedure to control for the bias that results from the use of restriction enzyme cleavage during integration site recovery (Wang et al. 2008). A large library of random sites was generated in silico, then the distances to restriction enzyme recognition sites were scored. Each experimental site was matched with three random control sites that were positioned the same number of nucleotides from a restriction site for the enzyme used to isolate the experimental site. That is, if an integration site was isolated after cleavage with ApoI, and the distance from the ApoI site to the edge of the HERV-K_{Con} sequence was 80 bp, then three random control sites were drawn from the pool that were also 80 bp from an ApoI site. Integration sites and matched random controls were annotated for proximity to genomic features, and the distributions were compared.

During mapping of integration sites, we noticed a substantial fraction of sequence reads contained HERV sequences abutting the HERV viral DNA end. Some could be identified as autointegration events, in which the ends of the HERV-K_{Con} DNA used internal HERV sequences as integration targets. Of 25,102 HERV sequence reads analyzed, 3784 (15%) had the HERV-K_{Con} viral DNA end abutting internal HERV DNA sequences. Another set of sequences showed the U3 end of the LTR joined to U5 LTR sequences in the orientation expected for 2-LTR circles (2017 sequence reads or 8% of the total). Autointegration and 2-LTR circle formation have been detected for all retroviruses studied (Lee and Coffin 1990; Farnet and Haseltine 1991; Lee and Craigie 1994), but are newly reported for HERV-K_{Con} here.

Comparison of HERV-K_{Con} integration site distributions in different cell types

HERV-K_{Con} integration sites were analyzed first by comparing the 293T and HT1080 data sets to each other. An

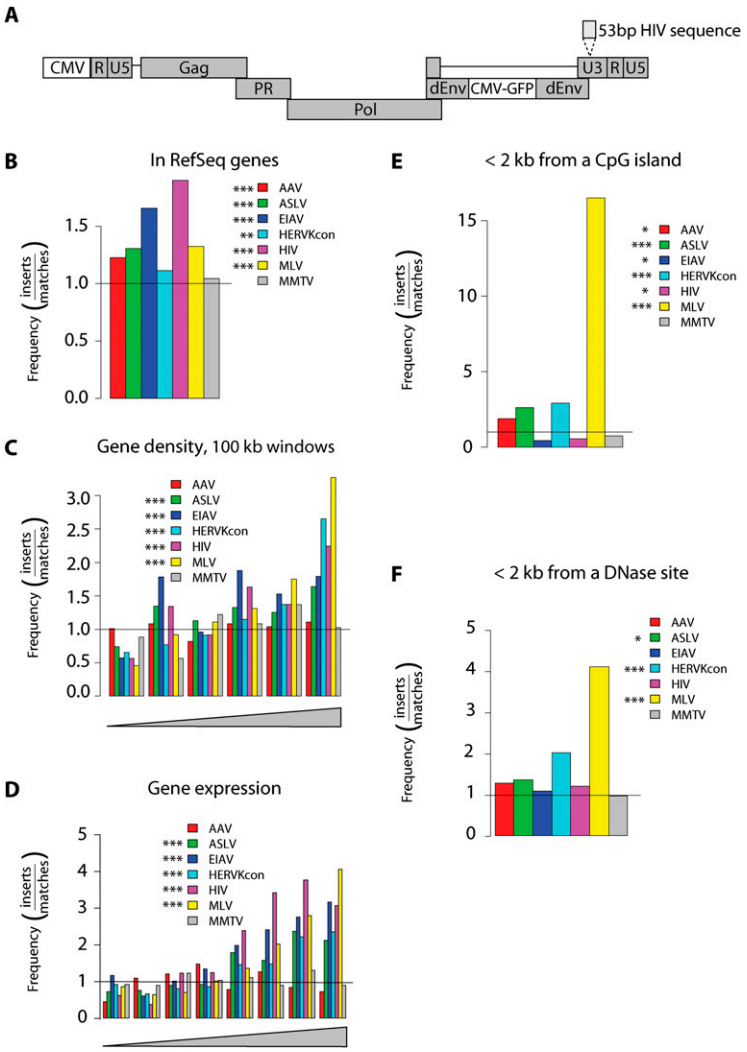

Figure 1. Integration target site selection of HERVK-con compared with other retroviruses. (*A*) Diagram of the modified HERVKcon used for integration targeting studies showing the insertion of a DNA tag in the U3 region of the 3′ LTR. For *B–F*, values are reported as the proportion of integration events divided by random events. The bar at 1.0 represents the expected random distribution. The statistical significance of differences from the matched random controls is shown by the asterisks next to the legends. (*) $0.05 > P > 0.01$; (**) $0.01 > P > 0.001$; (***) $P < 0.001$. (*B*) Integration frequency within RefSeq genes. (*C*) Integration frequency as a function of gene density. The *X*-axis shows six bins of increasing gene density from lowest (*left*) to highest (*right*). (*D*) Integration frequency relative to gene expression. All genes tested in 293T cells using the Affymetrix 133 array were divided into eight equal bins, then the proportions of integration sites in genes at each activity level were quantified and compared with random. The *X*-axis shows bins of increasing expression rank from lowest (*left*) to highest (*right*). (*E*) Integration frequency relative to CpG islands, scored as the proportion of integration sites within 2 kb of an annotated CpG island. (*F*) Integration frequency relative to sites of DNase I cleavage (Crawford et al. 2004), scored as the proportion of integration sites within 2 kb of an annotated cleavage site.

automated comparison was carried out over many types of genomic annotation, revealing no strong differences between cell types (data not shown). The 293T cells are female, while HT1080 is male, so there were differences in integration frequency in the sex chromosomes, but other forms of annotation showed no consistent strong differences. Therefore, data from both cell types were pooled except where noted in the ensuing analysis.

HERV-K_{Con} integration frequency relative to genomic features

The HERV-K_{Con} integration site distribution was compared with distributions of five other retroviruses (ASLV, EIAV, HIV, MLV, and MMTV), the parvovirus AAV, or matched random controls. MMTV is a betaretrovirus, the genus to which HERV-K_{Con} belongs. HIV-1 and EIAV are lentiviruses, ASLV is an alpharetrovirus, and MLV is a gammaretrovirus. The data sets are from multiple different cell types (Table 1), but each was selected to be representative of the relatively consistent pattern seen for each type of retrovirus over diverse cell types (e.g., Berry et al. 2006). Each of the comparison integration site data sets has been analyzed previously, and trends mentioned below for the comparison sets were reported previously except where noted (Mitchell et al. 2004; Bushman et al.

Table 1. *Integration data sets used in this study*

Set	Size	Cell type	Enzyme	Reference
AAV	436	MHF2	MfeI, AvrII	Miller et al. 2005
ASLV	557	293T	MseI	Mitchell et al. 2004
EIAV	747	SupT1	MseI	Marshall et al. 2007
HERVKcon	1064	293T	MseI	This study
HERVKcon	501	HT1080	MseI, ApoI	This study
HIV	729	293T	MseI	Ciuffi et al. 2005
MLV	1588	293T	MseI	This study
MMTV	236	578T	MseI	Faschinger et al. 2008
ERV2	10,573	NA	NA	Lander 2001
HML2(85)	402	NA	NA	This study

2005; Ciuffi et al. 2005; Miller et al. 2005; Berry et al. 2006; Marshall et al. 2007; Faschinger et al. 2008).

In the following analyses, the frequency of retroviral integration near the indicated feature is divided by the frequency in the matched random controls. Random integration thus has a value of one and is marked by a horizontal line, allowing departures from random to be visualized (Fig. 1B–F,P; values show significance compared with random).

HERV-K_{Con} integration was modestly favored within transcription units (Fig. 1B, $P < 0.01$). The lentiviruses HIV and EIAV showed a strong tendency to integrate within transcription units; ASLV, MLV, and AAV showed a significant tendency to integrate within transcription units; but MMTV showed no such preference. Additionally HERV-K_{Con} integration site density was positively correlated with gene density (Fig. 1C, $P < 0.001$). The other retroviruses also showed a tendency to integrate in gene-dense regions except MMTV and AAV, which showed no such tendency. HERV-K_{Con} and the other viruses (except AAV and MMTV) integrated more frequently in genes that were actively expressed as measured by Affymetrix microarrays (Fig. 1D).

A variety of genomic features are positively correlated with high gene density in the human genome, and several of these features were also positively correlated with HERV-K_{Con} integration frequency. For example, highly expressed genes (Fig. 1D) tend to reside in more gene-dense regions. Other correlated features include CpG islands (which often mark regulatory sites), DNase I cleavage sites, and characteristic forms of histone post-translational modification. HERV-K_{Con} integration near CpG islands and DNase cleavage sites was favored by two fold to 2.5-fold over random ($P < 0.001$ for both). In contrast, MMTV integration frequency was unaffected by CpG islands or DNase sites (Fig. 1E,F). MLV strongly favored integration near CpG islands (>15-fold) and DNase sites (Wu et al. 2003; Berry et al. 2006; Lewinski et al. 2006). As shown previously, the lentiviruses HIV and EIAV showed negative correlations between integration frequency and proximity to CpG islands (Fig. 1E, $P < 0.05$ for both, analyzed over a 2-kb window) (Mitchell et al. 2004; Berry et al. 2006) despite the positive correlation with integration in active transcription units.

Integration of HERV-K_{Con} and other retroviruses near sites of histone methylation and chromatin-bound proteins

To probe the relationship between HERV-K_{Con} integration frequency and chromatin structure, we quantified integration by HERV-K_{Con} and other retroviruses relative to sites of epigenetic modification and chromatin-bound proteins. We compared the density of integration sites with the density of 20 forms of histone post-translational methylation and three chromatin-bound proteins (Pol II, H2AZ, and CTCF), which had been mapped using chromatin immunoprecipitation and Solexa sequencing ("ChIP-Seq" method) (Barski et al. 2007). Each ChIP-Seq data set contained between one and 16 million sequence tags characterizing the distribution of each type of modification. Detailed information on the roles of each of these epigenetic marks can be found in Barski et al. (2007) and Taverna et al. (2007).

The associations between integration frequency and modification density were quantified and expressed as a heat map (Fig. 2) using the ROC area method described in Berry et al. (2006). The comparisons were carried out over three different interval sizes surrounding each integration site (5 kb, 10 kb, and 50 kb), since previous studies have shown that the interval sizes chosen for comparison can influence the conclusions (Berry et al. 2006). In this study, results were similar for each interval size examined (data not shown), so only the data for 50-kb intervals are presented. Results of statistical tests comparing the distributions of integration sites to the matched random controls are summarized as asterisks on each tile of the heat map.

HERV-K_{Con} integration was less frequent near histone methylation marks negatively associated with transcription, including H3K9 me2, H3K9 me3, and H3K27 me2. H3K9 me2 and me3 are associated with pericentromeric heterochromatin and were negatively correlated with integration by all the elements studied except MMTV and AAV. HERV-K_{Con} integration was more frequent near epigenetic marks associated with active transcription, H3K4 me1 and me2, H3K9 me1, H3K36 me3, and H4K20 me1, and also bound RNA Pol II. MLV and the lentivirus integration showed a stronger positive association with chromatin features linked to active transcription, consistent with the stronger biases away from random for their integration site distributions. H2AZ, a histone variant associated with promoters, was positively associated with MLV integration frequency but negatively associated with HIV-1 integration frequency, paralleling the integration frequencies of the two viruses near CpG islands. The presence of H2AZ did not have a detectable effect on HERV-K_{Con} integration frequency.

HERV-K_{Con} integration targeting and the distribution of HERV-K sequences in the human genome

To investigate the forces regulating accumulation of HERV elements in the human genome, we compared the distribution of the de novo HERV-K_{Con} integration sites described above with the distribution of Class II HERVs resident in the human genome. Two data sets of endogenous HERV-K sequences were studied (Table 1). For the first set, RepeatMasker (http://www.repeatmasker.org) was used to generate a large set of all Class II HERV-related sequences in the human genome (ERV2 data set; 10,573 sequences). The ERV2 set combines all subfamilies of the HERV-K superfamily (HML1 through HML10), including both old and young ERV2s. For the second set, we collected sequences with 85% matches to the HERV-K_{Con} LTR sequences, marking the most recently acquired and evolutionarily youngest HERV-K (HML2) elements [termed "HML2(85)"; 402 sequences]. The 85% cutoff was determined by comparison of the nucleotide percent similarity among HML2 elements, which ranges from 99% to ~85%, to the percent similarity of the next closest

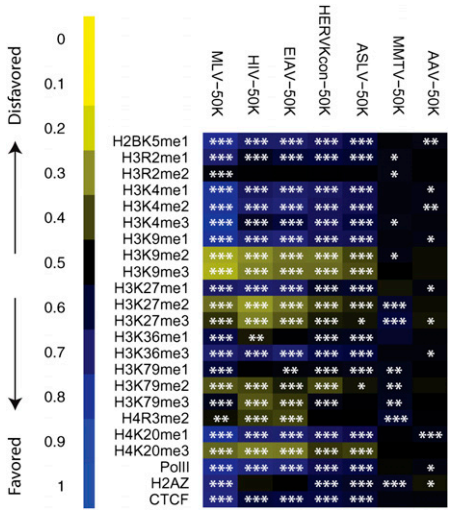

Figure 2. Integration frequency near sites of epigenetic modification and bound chromosomal proteins. Associations of integration with histone methylation and chromatin-bound proteins were quantified using ROC curve areas (Berry et al. 2006). In each case, the association of the experimental integration site data set was compared with the frequency in the matched random controls. Negative correlations between the genomewide annotation and integration frequency are shown by shades of yellow, with increasing intensity indicating stronger effects. Positive correlations are shown similarly but colored blue. Statistical tests for significant differences in distribution compared with the matched random control are summarized by asterisks on each tile of the heat map: (*) $0.05 > P > 0.01$; (**) $0.01 > P > 0.001$; (***) $P < 0.001$. The data on epigenetic modifications and bound proteins was from Barski et al. (2007). The viruses studied are marked above each column. CTCF is a DNA-binding protein proposed to be associated with chromatin boundaries, H2AZ a histone variant associated with promoters.

HERV-K subfamily, HML1, which ranges between 70% and 80% (Medstrand and Blomberg 1993).

We compared the distribution of HERV-K_{Con}, ERV2, and HML2(85) versus a variety of genomic features. As described above, de novo HERV-K_{Con} integration is modestly favored within transcription units. Both ERV2 and HML2(85) were less frequently present within transcription units than expected by chance (Fig. 3A, *P*-values show significance for comparisons between data sets), as reported previously.

For a variety of chromosomal features, the HML2(85) data set showed a distribution intermediate between HERV-K_{Con} and ERV2. HERV-K_{Con} proviruses were integrated more often than genomic HERVs near CpG islands [Fig. 3B, $P < 0.05$ vs. HML2(85), $P < 0.001$ vs. ERV2], but HML2(85) was more frequently found near CpG islands than ERV2. HERV-K_{Con} integration was most strongly favored near DNase I cleavage sites of the three sets, but HML2(85) was more often near DNase sites than ERV2 (Fig. 3C). HERV-K_{Con} integration sites were found more frequently near regions of higher G/C content than were ERV2 and HML2(85) sequences, but HML2(85) integration sites were more often in regions of higher G/C than ERV2 (Fig. 3D). Similarly for both integration in gene-dense regions (Fig. 3E) and in active genes (Fig. 3F), HERV-K_{Con} was found most frequently in these regions, HML2(85) was intermediate, and ERV2 was found the least frequently.

HERV distributions on the Y chromosome have been reported to have unique features (Kim et al. 2004; Villesen et al. 2004), so we compared the distributions of our three data sets on Y. HERV-K_{Con} integration on the Y chromosome was less frequent than expected by chance, although not statistically different from random (Fig. 3G). HML2(85) sequences were present on Y about as often as expected for random integration, while ERV2 was considerably enriched on Y (Fig. 3G, $P < 0.001$ for ERV2 compared with random). Here too we found that the two genomic HERV data sets also differed significantly from each other, with HML2(85) intermediate between HERV-K_{Con} and ERV2. Initial integration on Y may be disfavored due to its heterochromatic status, but proviruses, once integrated, will often be minimally deleterious because Y is gene-sparse. In addition, because Y lacks a homolog, newly integrated DNA cannot be removed by homologous recombination. These findings are consistent with a picture in which the Y chromosome is largely a graveyard for mobile DNA sequences (Lander 2001), and the ERV2 group is enriched on Y because it has been accumulating for longer than the HML2(85) subset.

Provirus orientation within transcription units: comparison of HERV-K_{Con} and resident HERVs

Lastly, we compared the orientation of HERV-K_{Con} proviruses within transcription units. Resident HERV proviruses in the human genome show a strong orientation bias relative to host gene transcription, accumulating preferentially in the opposite transcriptional orientation (Smit 1999; Medstrand et al. 2002; Villesen et al. 2004). We analyzed the orientation of HERV-K_{Con}, ERV2s, and HML2(85) elements within genes relative to the direction of host gene transcription (Fig. 4). Both the genomic ERV2s and the HML2(85) elements were found predominantly in the opposite transcriptional orientation relative to the gene into which they had integrated. However, the HERV-K_{Con} integration-site data set showed no significant departure from random orientation. Thus the orientation bias observed for endogenous HERVs is not a result of favoring of a specific orientation relative to host gene transcription during initial integration.

Discussion

Here we report a study of integration target site selection by HERV-K_{Con} and its relationship to the distribution of fixed HERV sites in the human genome. Sites of HERV-K_{Con} integration were slightly enriched in transcription units, in gene-dense regions, and in a collection of features associated with gene activity. The endogenous HERV-K elements, ERV2 and HML2(85), showed a very different distribution and were enriched outside genes.

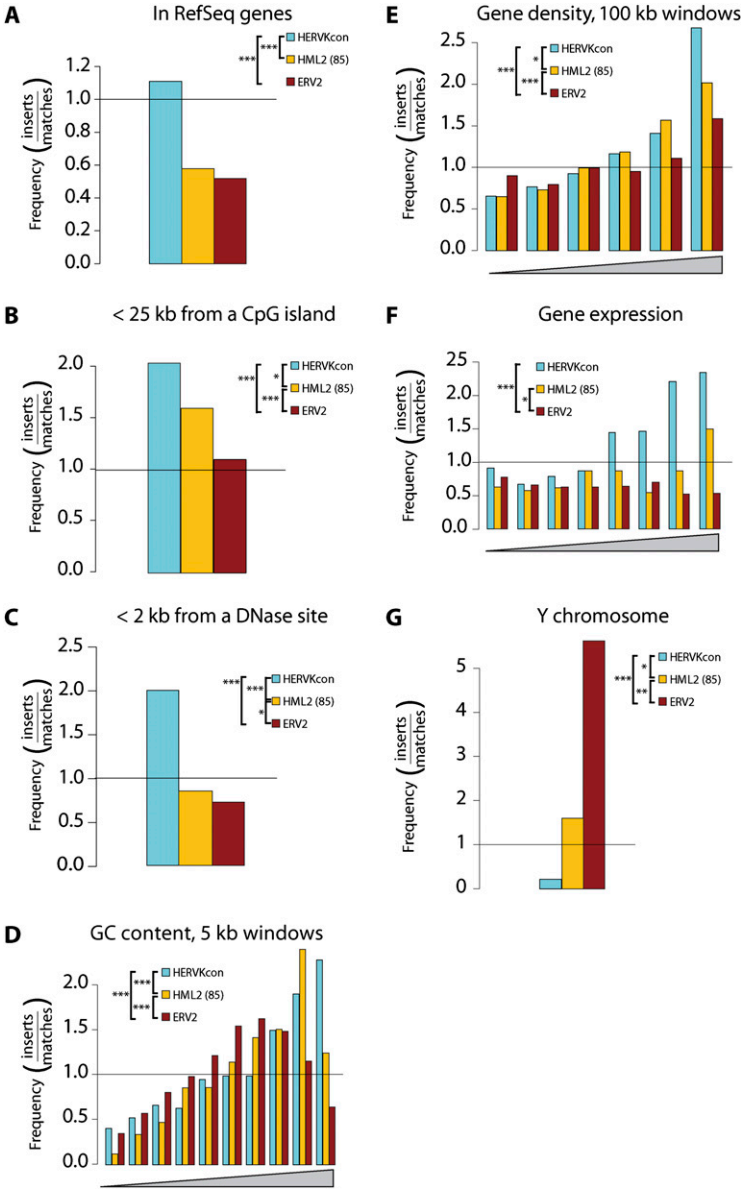

Figure 3. Integration of HERVKcon versus resident ERV2 elements. Values are reported as the proportion of integration events divided by random events. The bar at 1.0 represents the expected random distribution. The statistical significance of differences between data sets is shown by the asterisks next to the legends: (*) $0.05 > P > 0.01$; (**) $0.01 > P > 0.001$; (***) $P < 0.001$. (*A*) Integration frequency relative to transcription units as defined by the RefSeq database. (*B*) Integration frequency relative to CpG islands. (*C*) Integration frequency relative to DNAse I cleavage sites, 2-kb windows. (*D*) Integration frequency relative to G/C content, 5-kb windows. (*E*) Integration frequency relative to gene density. (*F*) Integration frequency relative to gene activity. In this plot, Affymetrix microarray analysis was used to rank the activity of all genes queried, then the ranks were distributed into eight bins. The genes hosting integration events were then distributed into the bins and the frequencies compared with matched random controls. (*G*) Integration site distribution on the Y chromosome. Only the HT1080 data set was used in this analysis, since it is from a male cell line.

The youngest endogenous HERVs, HML2(85), showed a pattern intermediate between the older ERV2 sites and the de novo integrated HERV-K_{Con}. These data support a two-step model for accumulation of fixed HERV elements in the human genome, in which integration targeting preferences dictated the initial placement of integration sites, while subsequent purifying selection eliminated the majority of insertions because they were deleterious to host viability. According to this idea, the youngest HERVs, HML2(85), are only part way along this progression.

Previous studies have reported differing distributions among the distinct HERV element families and investigated the mechanisms that mediate gene disruption upon integration (Mager 1999; Smit 1999; Lander 2001; van de Lagemaat et al. 2006). These studies support the idea that strong splice sites and poly(A) sites within HERV elements can disrupt gene transcription, as has been seen with other genomic parasites (Britten 1996; Jordan et al. 2003; van de Lagemaat et al. 2003). However, these studies did not identify a distinctive distribution pattern for the most recently integrated HERV sequences. Our study used homology searching to form a collection of the most similar, hence youngest, genomic HERV-K(HML2) elements, and it was by analyzing this collection that we were able to detect that the distribution of HML2(85) sequences is intermediate between the older ERV2 elements and newly integrated HERV-K_{Con}. In contrast to the relationship to genomic features, the orientation bias within genes is evident and similar for both the HML2(85) and ERV2 data sets. This is consistent with the idea that particularly disruptive proviruses integrated within genes may be removed relatively quickly by purifying selection.

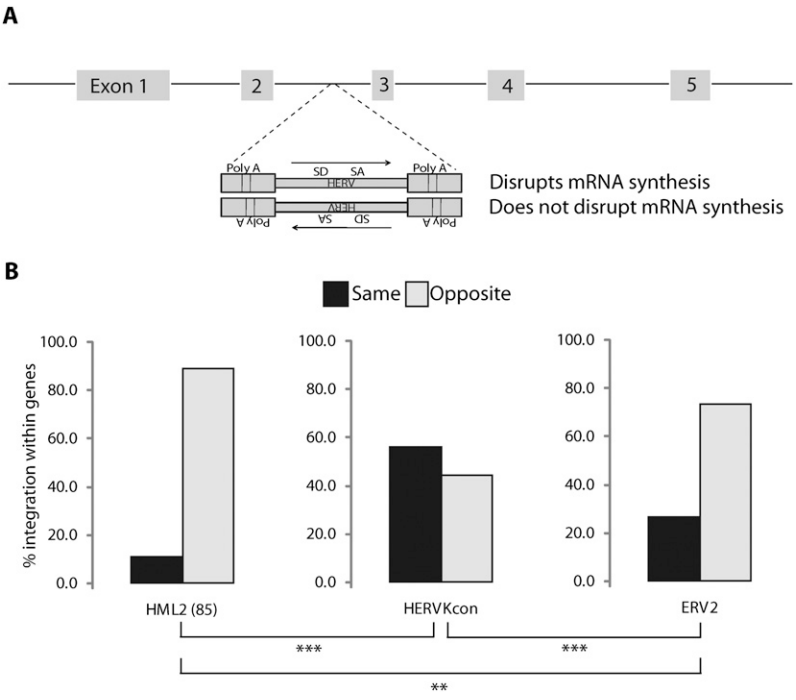

Figure 4. Proviral orientations for newly integrated HERVKcon versus resident HERVs. (*A*) Diagram showing proviral orientations and the potential for transcriptional disruption by provirus-encoded transcription signals. (SD) Splice donor; (SA) splice acceptor; (PolyA) polyA signal. (*B*) Transcriptional orientation of ERV2, HML2(85), and HERVKcon sequences found within gene-coding regions as defined by the RefSeq database.

These findings support the use of endogenous retroviruses as phylogenetic markers (Shimamura et al. 1997; Takahashi et al. 2001; Salem et al. 2003; Singer et al. 2003). One requirement for using endogenous retrovirus sequence for lineage tracing is that many genomic sites must be capable of hosting integration events, so that the observed precise coincidence of integration site locations between taxa can be interpreted as evidence of common descent. Our data indicate that a large number of sites in the genome are capable of hosting integration events, strengthening the idea that endogenous retroviral positions can be useful phylogenetic markers (Bushman 2001; Craig et al. 2002; Kazazian 2004).

Previous studies of retroviral integration targeting have shown that retroviruses from the same genus tend to share the same targeting patterns, but HERV-K_{Con} appears to break that trend. HERV-K_{Con} is most closely related to exogenous betaretroviruses, and its integrase protein sequence clusters with MMTV, a prototype betaretrovirus rather than integrases from other retroviral genera (Supplemental Material 2). As discussed above, HERV-K_{Con} integration is more frequent in gene-rich regions, and near genomic features associated with active transcription, somewhat resembling ASLV (Mitchell et al. 2004; Narezkina et al. 2004), human T-cell leukemia virus (Derse et al. 2007; Meekings et al. 2008), and spumaretroviruses (Nowrouzi et al. 2006; Trobridge et al. 2006). Surprisingly, the reported MMTV distribution is almost perfectly random (Faschinger et al. 2008). The only other data set with such a random distribution is AAV, but AAV is believed to become integrated at cellular DNA double-strand breaks by the action of cellular DNA repair enzymes (Rutledge and Russell 1997; Song et al. 2001). MMTV, in contrast, encodes an integrase protein, and MMTV integration events show the usual sequence features associated with retroviral integration. It will be useful to obtain more data on integration site distributions from the betaretrovirus genus to clarify this puzzling observation. Another previously noted surprising difference between the members of betaretroviridae is the location of assembly (Dewannieux et al. 2006; Lee and Bieniasz 2007). Assembly of the betaretrovirus Mason Pfizer monkey virus (MPMV) takes place at a perinuclear region (Rhee and Hunter 1987), whereas HERV-K assembly takes place at the plasma membrane. These two phenotypic differences (assembly and integration targeting) within betaretroviridae suggest that the genus may not be monophyletic. Thus, although our integration targeting data for HERV-K_{Con} seems likely to model trends for all of the HERV-K elements, it is uncertain to what extent, if any, the data for HERV-K_{Con} models the other HERV families that most closely resemble exogenous retroviruses of other genera.

The two-step model for HERVs accumulation is likely operating on endogenous retroviruses and other integrating elements of many vertebrates (Bushman 2001; Craig et al. 2002; Kazazian 2004; Han and Boeke 2005). In a previous study, Barr et al. (2005) compared de novo ASLV integration events in chicken cells to fixed proviruses in the chicken germline that were derived from the same retroviral group. They found that de novo ASLV integration showed a modest preference for transcription units, while fixed ASLVs in the germline accumulated outside of transcription units. Fixed ASLVs in the germline also showed an orientation bias, so that proviruses within genes tended to accumulate in opposite transcriptional orientation relative to the host gene, while the de novo integration events showed no such bias. Similar

biases in endogenous provirus accumulation have also been observed in mouse and rat (Barr et al. 2005; van de Lagemaat et al. 2006). These findings suggest that purifying selection is operating similarly on the endogenous retroviruses inhabiting the genomes of many vertebrates (Barr et al. 2005; Brookfield 2005; Cutter et al. 2005; Roy-Engel et al. 2005; Lowe et al. 2007).

Materials and methods

Cell lines and transfection

293T and HT1080 cells were maintained in DMEM supplemented with 10% fetal calf serum and gentamycin. 293T cells were seeded at $\sim 7 \times 10^6$ cells in 10-cm plates and transfected with polyethylenimine the following day with 9 μg of CCGBX-P, 2.5 μg of pCRVI/Con GP, 2.5 μg of pCR3.1/K108 Rec, and 1 μg of VSVG as previously described (Lee and Bieniasz 2007). Five hours after transfection, supernatant was replaced with fresh media containing 5 μM sodium butyrate. Supernatant was collected after an additional 40 h, filtered via 0.2-μm filter, and treated with 0.1 U/μL DNAse I (Roche) for 1 h at 37°C, supplemented with 10 mM $MgCl_2$, to eliminate residual transfected plasmid DNA.

Infection and recovery of integration sites

For infection, 293T and HT1080 cells were seeded at 2.5×10^5 and 1.5×10^5 cells per well, respectively, in six-well plates the previous day. Cells were spinoculated with the DNAse-treated HERV-K_{Con} virus at 2000 rpm for 2 h at room temperature. Total DNA was collected 48 h post-infection.

Recovery of integration sites was performed as described (Wang et al. 2007). Two micrograms of genomic DNA were digested overnight with MseI or ApoI, ligated to linkers overnight at 16°C, and digested a second time with PstI and DpnI. Nested PCR was then carried out under stringent conditions using LTR primers complementary to HERV U3 sequences. Oligonucleotides used in this study are listed in Supplemental Table 1. DNA barcodes were included in the second-round PCR primers in order to track sample origin (Hoffmann et al. 2007). Amplification products were gel-purified and sequenced by massively parallel pyrophosphate sequencing. Only sequences that uniquely aligned to the human genome by BLAT (hg18, version 36.1, >98% match score) and began within 3 bp of the LTR end were used in downstream analyses. Integration sites sequences have been deposited in GenBank under the accession numbers FI497131–FI498695.

Of the 25,102 sequences analyzed, 6873 showed a high-quality match to the HERV-K_{Con} vector using BLAT. Sequences were classified as 2-LTR circle if there was a match to the U5 LTR end in the expected orientation, while allowing indels of 100 bp. One-thousand-fifty-eight were an internal fragment derived from the internal U3 LTR and flanking sequences. A total of 3784 sequences showed the viral DNA end abutting internal HERV-K_{Con} sequences and were classified as autointegration products. Another 14 sequence reads had complex structures and were not included in the above categories.

Analysis of other retroviruses and genomic HERVs

Integration site data sets published previously (Table 1) were analyzed using the bioinformatics pipeline mentioned above. Discrepancies in data set sizes likely result from differences in quality-control thresholds compared with the original publications. The ERV2 data set was generated using RepeatMasker and the human genome (hg18, version 36.1). For the HML2(85) data set, the HERV-K_{Con} LTR sequence was used as a query to search for sequences 85% or higher in nucleotide similarity and longer than 600 bp using the Ensembl BLASTN (http://www.ensembl.org/Multi/blastview). The remaining sequences were organized by chromosomal location, and LTRs <9000 bp apart were manually determined as either solo-LTRs or LTRs of the same provirus based on the LTR flanking sequence and identification of target site duplication sequence. Duplicate hits due to genome duplications or belonging to the same provirus were condensed into a single entry.

Analysis of integration site distributions

Analyses were carried out as described (Berry et al. 2006; Marshall et al. 2007). A detailed account of the statistical methods used can be found in Supplemental Material 1, and the methods for forming and analyzing heat maps using ROC curves in Supplemental Material 3.

Analyses of gene expression used data from 293T cells, with expression measured using the Affymetrix HU133 plus 2.0 gene chip array. Expression values were ranked and divided into eight bins according to rank. Consensus sequence analysis at the point of integration was performed using WebLogo (http://weblogo.Berkeley.edu/logo.cgi), and the primary sequence features found to match those expected for HERV-K integration (data not shown).

$CD4^+$ T cells were used to generate ChIP-Seq data (Barski et al. 2007), differing from the cell types studied here. However, genome-wide surveys of modification densities in different cell types from the ENCODE project show that a substantial fraction of epigenetic marks are common to most cell types analyzed probably because a large fraction of transcription is from "housekeeping genes." For example, for HIV data sets, there is no stronger correlation with epigenetic marks measured in T cells than for integration site data sets from T cells than from other cell types (C.C. Berry, T.L. Brady, F.D. Bushman, and K. Ronen, unpubl.). Furthermore, differences due to experimental error were generally greater than differences due to cell type (ENCODE Project Consortium 2004). Thus we believe that the data from Barski et al. (2007) represent a useful approximation to the cell types studied here.

Acknowledgments

We are grateful to members of the Bieniasz and Bushman laboratories for help and suggestions. This work was supported by NIH grant AI52845, the University of Pennsylvania Center for AIDS Research, and the Penn Genome Frontiers Institute with a grant with the Pennsylvania Department of Health to F.D.B. and by NIH grant AI064003 to P.D.B.

References

Aiuti, A., Cassani, B., Andolfi, G., Mirolo, M., Biasco, L., Recchia, A., Urbinati, F., Valacca, C., Scaramuzza, S., Cazzola, M., et al. 2007. Multilineage hematopoietic reconstitution without clonal selection in ADA-SCID patients treated with stem cell gene therapy. *J. Clin. Invest.* **117:** 2233–2240.

Bannert, N. and Kurth, R. 2004. Retroelements and the human genome: New perspectives on an old relation. *Proc. Natl. Acad. Sci.* **101:** 14572–14579.

Barbulescu, M., Turner, G., Seaman, M.I., Deinard, A.S., Kidd, K.K., and Lenz, J. 1999. Many human endogenous retrovirus K (HERV-K) proviruses are unique to humans. *Curr. Biol.* **26:** 861–868.

Barr, S.D., Leipzig, J., Shinn, P., Ecker, J.R., and Bushman, F.D. 2005. Integration targeting by avian sarcoma-leukosis virus

and human immunodeficiency virus in the chicken genome. *J. Virol.* **79:** 12035–12044.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129:** 823–837.

Belshaw, R., Pereira, V., Katzourakis, A., Talbot, G., Paces, J., Burt, A., and Tristem, M. 2004. Long-term reinfection of the human genome by endogenous retroviruses. *Proc. Natl. Acad. Sci.* **101:** 4894–4899.

Belshaw, R., Dawson, A.L., Woolven-Allen, J., Redding, J., Burt, A., and Tristem, M. 2005. Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K(HML2): Implications for present-day activity. *J. Virol.* **79:** 12507–12514.

Berry, C., Hannenhalli, S., Leipzig, J., and Bushman, F.D. 2006. Selection of target sites for mobile DNA integration in the human genome. *PLoS Comput. Biol.* **2:** e157. doi: 10.1371/journal.pcbi.0020157.

Britten, R.J. 1996. DNA sequence insertion and evolutionary variation in gene regulation. *Proc. Natl. Acad. Sci.* **93:** 9374–9377.

Brookfield, J.F. 2005. The ecology of the genome—Mobile DNA elements and their hosts. *Nat. Rev. Genet.* **6:** 128–136.

Bushman, F.D. 2001. *Lateral DNA transfer: Mechanisms and consequences*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Bushman, F., Lewinski, M., Ciuffi, A., Barr, S., Leipzig, J., Hannenhalli, S., and Hoffmann, C. 2005. Genome-wide analysis of retroviral DNA integration. *Nat. Rev. Microbiol.* **3:** 848–858.

Ciuffi, A., Llano, M., Poeschla, E., Hoffmann, C., Leipzig, J., Shinn, P., Ecker, J.R., and Bushman, F. 2005. A role for LEDGF/p75 in targeting HIV DNA integration. *Nat. Med.* **11:** 1287–1289.

Ciuffi, A., Mitchell, R.S., Hoffmann, C., Leipzig, J., Shinn, P., Ecker, J.R., and Bushman, F.D. 2006. Integration site selection by HIV-based vectors in dividing and growth-arrested IMR-90 lung fibroblasts. *Mol. Ther.* **13:** 366–373.

Costas, J. 2001. Evolutionary dynamics of the human endogenous retrovirus family HERV-K inferred from full-length proviral genomes. *J. Mol. Evol.* **53:** 237–243.

Craig, N.L., Craigie, R., Gellert, M., and Lambowitz, A.M. 2002. *Mobile DNA II*. ASM Press, Washington, DC.

Crawford, D.H., Holt, I.E., Mullikin, J.C., Tai, D., Blakesley, R., Bouffard, G., Young, A., Masiello, C., Green, E.D., Wolfsberg, T.D., et al. National Institutes of Health Intramural Sequencing Center. 2004. Identifying gene regulatory elements by genome-wide recovery of DNAse hypersensitive sites. *Proc. Natl. Acad. Sci.* **101:** 922–927.

Cutter, A.D., Good, J.M., Pappas, C.T., Saunders, M.A., Starrett, D.M., and Wheeler, T.J. 2005. Transposable element orientation bias in the *Drosophila melanogaster* genome. *J. Mol. Evol.* **61:** 733–741.

Dawkins, R. 1976. *The selfish gene.*. Oxford University Press, Oxford.

Derse, D., Crise, B., Li, Y., Princler, G., Lum, N., Stewart, C., McGrath, C.F., Hughes, S.H., Munroe, D.J., and Wu, X. 2007. Human T-cell leukemia virus type 1 integration target sites in the human genome: Comparison with those of other retroviruses. *J. Virol.* **81:** 6731–6741.

Dewannieux, M., Harper, F., Richaud, A., Letzelter, C., Ribet, D., Pierron, G., and Heidmann, T. 2006. Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. *Genome Res.* **16:** 1548–1556.

ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia of DNA elements) project. *Science* **306:** 636–640.

Engelman, A. 2005. The ups and downs of gene expression and retroviral DNA integration. *Proc. Natl. Acad. Sci.* **102:** 1275–1276.

Esnault, C., Priet, S., Ribet, D., Heidmann, O., and Heidmann, T. 2008. Restriction by APOBEC3 proteins of endogenous retroviruses with an extracellular life cycle: Ex vivo effects and in vivo 'traces' on the murine IAPE and human HERV-K elements. *Retrovirology* **5:** 75.

Farnet, C.M. and Haseltine, W.A. 1991. Circularization of human immunodeficiency virus type 1 DNA in vitro. *J. Virol.* **65:** 6942–6952.

Faschinger, A., Rouault, F., Sollner, J., Lukas, A., Salmons, B., Gunzburg, W.H., and Indik, S. 2008. Mouse mammary tumor virus integration site selection in human and mouse genomes. *J. Virol.* **82:** 1360–1367.

Han, J.S. and Boeke, J.D. 2005. LINE-1 retrotransposons: Modulators of quantity and quality of mammalian gene expression? *Bioessays* **27:** 775–784.

Hendrix, R.W., Roberts, J.W., Stahl, F.W., and Weisberg, R.A. 1983. *Lambda II.* Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

Hoffmann, C., Minkah, N., Leipzig, J., Wang, G., Arens, M.Q., Tebas, P., and Bushman, F.D. 2007. DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. *Nucleic Acids Res.* **35:** e91. doi: 10.1093/nar/gkm435.

Hughes, J.F. and Coffin, J.M. 2004. Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: Implications for human and viral evolution. *Proc. Natl. Acad. Sci.* **101:** 1668–1672.

Jordan, I.K., Rogozin, I.B., Glazko, G.V., and Koonin, E.V. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet.* **19:** 68–72.

Kazazian, H.H. 2004. Mobile elements: Drivers of genome evolution. *Science* **303:** 1626–1632.

Kim, T.H., Jeon, Y.J., Yi, J.M., Kim, D.S., Huh, J.W., Hur, C.G., and Kim, H.S. 2004. The distribution and expression of HERV families in the human genome. *Mol. Cells* **18:** 87–93.

Lander, E. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Lee, Y.N. and Bieniasz, P.D. 2007. Reconstitution of an infectious human endogenous retrovirus. *PLoS Pathog.* **3:** e10. doi: 10.1371/journal.ppat.0030010.

Lee, Y.M.H. and Coffin, J.M. 1990. Efficient autointegration of avian retrovirus DNA in vitro. *J. Virol.* **64:** 5958–5965.

Lee, M.S. and Craigie, R. 1994. Protection of retroviral DNA from autointegration: Involvement of a cellular factor. *Proc. Natl. Acad. Sci.* **91:** 9823–9827.

Lee, Y.N., Malim, M.H., and Bieniasz, P.D. 2008. Hypermutation of an ancient human retrovirus by APOBEC3G. *J. Virol.* **82:** 8762–8770.

Lewinski, M., Bisgrove, D., Shinn, P., Chen, H., Verdin, E., Berry, C.C., Ecker, J.R., and Bushman, F.D. 2005. Genome-wide analysis of chromosomal features repressing HIV transcription. *J. Virol.* **79:** 6610–6619.

Lewinski, M.K., Yamashita, M., Emerman, M., Ciuffi, A., Marshall, H., Crawford, G., Collins, F., Shinn, P., Leipzig, J., Hannenhalli, S., et al. 2006. Retroviral DNA integration: Viral and cellular determinants of target-site selection. *PLoS Pathog.* **2:** e60. doi: 10.1371/journal.ppat.0020060.

Lowe, C.B., Bejerano, G., and Haussler, D. 2007. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proc. Natl. Acad. Sci.* **104:** 8005–8010.

Mager, D.L. 1999. Human endogenous retroviruses and pathogenicity: Genomic considerations. *Trends Microbiol.* **7:** 431.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen,

Z., et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376–380.

Marshall, H., Ronen, K., Berry, C., Llano, M., Sutherland, H., Saenz, D., Bickmore, W., Poeschla, E., and Bushman, F. 2007. Role of PSIP1/LEDGF/p75 in lentiviral infectivity and integration targeting. *PLoS One* **2:** e1340. doi: 10.1371/journal.pone.0001340.

Medstrand, P. and Blomberg, J. 1993. Characterization of novel reverse transcriptase encoding human endogenous retroviral sequences similar to type A and type B retroviruses: Differential transcription in normal human tissues. *J. Virol.* **67:** 6778–6787.

Medstrand, P. and Mager, D.L. 1998. Human-specific integrations of the HERV-K endogenous retrovirus family. *J. Virol.* **72:** 9782–9787.

Medstrand, P., van de Lagematt, L.N., and Mager, D.L. 2002. Retroelement distributions in the human genome: Variations associate with age and proximity to genes. *Genome Res.* **12:** 1483–1495.

Meekings, K.N., Leipzig, J., Bushman, F.D., Taylor, G.P., and Bangham, C.R. 2008. HTLV-1 integration into transcriptionally active genomic regions is associated with proviral expression and with HAM/TSP. *PLoS Pathog.* **4:** e1000027. doi: 10.1371/journal.ppat.1000027.

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.

Mitchell, R.S., Beitzel, B.F., Schroder, A.R., Shinn, P., Chen, H., Berry, C.C., Ecker, J.R., and Bushman, F.D. 2004. Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol.* **2:** E234. doi: 10.1371/journal.pbio.0020234.

Narezkina, A., Taganov, K.D., Litwin, S., Stoyanova, R., Hayashi, J., Seeger, C., Skalka, A.M., and Katz, R.A. 2004. Genome-wide analyses of avian sarcoma virus integration sites. *J. Virol.* **78:** 11656–11663.

Nowrouzi, A., Dittrich, M., Klanke, C., Heinkelein, M., Rammling, M., Dandekar, T., von Kalle, C., and Rethwilm, A. 2006. Genome-wide mapping of foamy virus vector integrations into a human cell line. *J. Gen. Virol.* **87:** 1339–1347.

Orgel, L.E. and Crick, F.H.C. 1980. Selfish DNA: The ultimate parasite. *Nature* **284:** 604–607.

Rhee, S.S. and Hunter, E. 1987. Myristylation is required for intracellular transport but not for assembly of D-type retrovirus capsids. *J. Virol.* **61:** 1045–1053.

Roy-Engel, A.M., El-Sawy, M., Farooq, L., Odom, G.L., Perepelitsa-Belancio, V., Bruch, H., Oyeniran, O.O., and Deininger, P.L. 2005. Human retroelements may introduce intragenic polyadenylation signals. *Cytogenet. Genome Res.* **110:** 365–371.

Rutledge, E.A. and Russell, D.W. 1997. Adeno-associated virus vector integration junctions. *J. Virol.* **71:** 8429–8436.

Salem, A., Ray, D.A., Xing, J., Callinan, P.A., Myers, J.S., Hedges, D.J., Garber, R.K., Witherspoon, D.J., Jorde, L.B., and Batzer, M.A. 2003. Alu elements and hominid phylogenetics. *Proc. Natl. Acad. Sci.* **100:** 12787–12791.

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.

Shimamura, M., Yasue, H., Ohshima, K., Abe, H., Kato, H., Kishiro, T., Goto, M., Munechika, I., and Okada, N. 1997. Molecular evidence from retroposons that whales form a clade within even-toed ungulates. *Nature* **388:** 666–670.

Singer, S.S., Schmitz, J., Schwiegk, C., and Zischler, H. 2003. Molecular cladistic markers in New World monkey phylogeny (Platyrrhini, Primates). *Mol. Phylogenet. Evol.* **26:** 490–501.

Smit, A.F. 1999. Interspersed repeats and other momentos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9:** 657–663.

Song, S., Laipis, P.J., Berns, K.I., and Flotte, T.R. 2001. Effect of DNA-dependent protein kinase on the molecular fate of the rAAV2 genome in skeletal muscle. *Proc. Natl. Acad. Sci.* **98:** 4084–4088.

Takahashi, K., Nishida, M., Yuma, M., and Okada, N. 2001. Retroposition of the AFC family of SINEs (short interspersed repetitive elements) before and during the adaptive radiation of cichlid fishes in Lake Malawi and related inferences about phylogeny. *J. Mol. Evol.* **53:** 496–507.

Taverna, S.D., Li, H., Ruthenburg, A.J., Allis, C.D., and Patel, D.J. 2007. How chromatin-binding modules interpret histone modifications: Lessons from professional pocket pickers. *Nat. Struct. Mol. Biol.* **14:** 1025–1040.

Trobridge, G.D., Miller, D.G., Jacobs, M.A., Allen, J.M., Kiem, H., Kaul, R., and Russell, D.W. 2006. Foamy virus vector integration sites in normal human cells. *Proc. Natl. Acad. Sci.* **103:** 1498–1503.

Turner, G., Barbulescu, M., Su, M., Jensen-Seaman, M.I., Kidd, K.K., and Lenz, J. 2001. Insertional polymorphisms of full-length endogenous retroviruses in humans. *Curr. Biol.* **11:** 1531–1535.

van de Lagemaat, L.N., Landry, J., Mager, D.L., and Medstrand, P. 2003. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet.* **19:** 530–536.

van de Lagemaat, L.N., Medstrand, P., and Mager, D.L. 2006. Multiple effects govern endogenous retrovirus survival patterns in human gene introns. *Genome Biol.* **7:** R86. doi: 10.1186/gb-2006-7-9-r86.

Venter, J.C. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.

Villesen, P., Aagaard, L., Wiuf, C., and Pedersen, F.S. 2004. Identification of endogenous retroviral reading frames in the human genome. *Retrovirology* **1:** 32. doi: 10.1186/1742-4690-1-32.

Wang, G.P., Ciuffi, A., Leipzig, J., Berry, C.C., and Bushman, F.D. 2007. HIV integration site selection: Analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Res.* **17:** 1186–1194.

Wang, G.P., Garrigue, A., Ciuffi, A., Ronen, K., Leipzig, J., Berry, C., Lagresle-Peyrou, C., Benjelloun, F., Hacein-Bey-Abina, S., Fischer, A., et al. 2008. DNA bar coding and pyrosequencing to analyze adverse events in therapeutic gene transfer. *Nucleic Acids Res.* **36:** e49. doi: 10.1093/nar/gkn125.

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.