Cell, Vol. 110, 521–529, August 23, 2002, Copyright ©2002 by Cell Press

HIV-1 Integration in the Human Genome Favors Active Genes and Local Hotspots

**Astrid R.W. Schröder,[1] Paul Shinn,[2]
Huaming Chen,[2] Charles Berry,[3] Joseph R. Ecker,[2]
and Frederic Bushman[1,4]**
[1]Infectious Disease Laboratory
[2]Genomic Analysis Laboratory
The Salk Institute
10010 North Torrey Pines Road
La Jolla, California 92037
[3]Department of Family/Preventive Medicine
School of Medicine
University of California, San Diego
San Diego, California 92093

Summary

A defining feature of HIV replication is integration of the proviral cDNA into human DNA. The selection of chromosomal targets for integration is crucial for efficient viral replication, but the mechanism is poorly understood. Here we describe mapping of 524 sites of HIV cDNA integration on the human genome sequence. Genes were found to be strongly favored as integration acceptor sites. Global analysis of cellular transcription indicated that active genes were preferential integration targets, particularly genes that were activated in cells after infection by HIV-1. Regional hotspots for integration were also found, including a 2.4 kb region containing 1% of sites. These data document unexpectedly strong biases in integration site selection and suggest how selective targeting promotes aggressive HIV replication.

Introduction

The early steps of retroviral replication involve reverse transcription of the viral RNA genome to make a cDNA copy, then integration of that cDNA copy into a chromosome of the host cell. The integration reaction requires specific sequences at the ends of the viral cDNA, which bind the viral-encoded integrase and other proteins to form preintegration complexes (PICs). The cellular DNA sequences that serve as integration target sites, however, show no strong primary sequence preferences. Here we investigate targeting of integration in the cellular chromosomes in vivo, where the target DNA is packaged in chromatin, revealing strong biases imposed by the in vivo environment.

Studies using integration in vitro have clarified factors influencing integration site selection in simplified models. DNA binding proteins bound to target DNA can block integration by obstructing access of integration complexes to target DNA (Bushman, 1994; Pryciak and Varmus, 1992). In contrast, DNA bending proteins such as nucleosomes can actually promote integration (Pruss et al., 1994a, 1994b; Pryciak et al., 1992; Pryciak and Varmus, 1992). On the nucleosome, the positions of maximal DNA distortion were particularly favored for integration (Pruss et al., 1994a, 1994b), probably because integrase distorts its DNA substrates during the reaction cycle (Bushman and Craigie, 1992; Katz et al., 1998, 2001; Scottoline et al., 1997), so prior distortion of integration target DNA facilitates catalysis. Thus, wrapping of DNA in nucleosomes alone does not inhibit integration.

In cells, nucleosomal DNA is assembled into higher-order chromatin, with unknown consequences for integration site selection. Studies in vivo can potentially address the influence of chromatin on integration targeting, but previous studies have led to diverse and sometimes conflicting proposals (for reviews see Bushman, 2001; Coffin et al., 1997). Early studies suggested that integration may be favored near DNase I hypersensitive sites (Panet and Cedar, 1977; Rohdewohld et al., 1987; Vijaya et al., 1986) or active genes (Mooslehner et al., 1990; Scherdin et al., 1990). However, a recent report suggested that high-level transcription actually interfered with integration by avian leukosis virus (ALV) (Weidhaas et al., 2000). Heterochromatic regions of chromosomes have been proposed to be disfavored for HIV integration in cell culture (Carteau et al., 1998), but a study of HTLV integration sites from patients did not see this trend (Leclercq et al., 2000). Another early study suggested that ALV strongly favored integration at hotspots comprised of specific base pairs in an avian genome (Shih et al., 1988), but a reexamination of these sites using another method did not reveal such a strong bias (Withers-Ward et al., 1994). Many of the above studies were limited by analysis of relatively few integration events and none analyzed integration site placement on a complete genome sequence.

Here we report an analysis of integration target selection by HIV-1 in the human genome. We infected a human lymphoid cell line with HIV or an HIV-based vector and cloned 524 junctions between viral and cellular DNA. The sequences were then determined and mapped on the human genome sequence. As a control, 111 sites were generated by integration in vitro into naked human DNA and their genomic distribution compared with the in vivo integration sites. Genes were clearly preferential integration targets for the in vivo-targeted set but not for the in vitro control. Transcriptional profiling revealed a strong correlation between gene activity and integration targeting, particularly for genes that were active in cells after infection with the HIV vector. Hotspots for integration were also detected, including a 2.5 kb region that contained 1% of integration events. These data reveal unexpected specificity in integration targeting by HIV and begin to elucidate the mechanism of site choice in vivo.

Results

Cloning Sites of HIV Integration

To investigate HIV integration targeting in the human genome, we infected a human T cell line (SupT1) with

[4]Correspondence: bushman@salk.edu

HIV or an HIV-based vector (Carteau et al., 1998; Follenzi et al., 2000), then isolated chromosomal DNA containing integrated proviruses after 48 hr. The short time was chosen to minimize possible selection at the cellular level due to integration, thus preserving the initial distribution of sites (in this work, "site" is used to indicate both a short sequence in the genome and that same sequence flanking a provirus after integration). Purified cellular DNA was cleaved with restriction enzymes, linkers were ligated onto the DNA ends, and then sequences were amplified with primers complementary to the linker and the HIV cDNA end. DNA fragments containing junctions between integrated HIV proviruses and cellular DNA were then cloned and sequenced, yielding 524 different integration target sequences.

As a control, integration products were generated in vitro using naked genomic DNA from SupT1 cells as an integration target. HIV preintegration complexes (PICs) were used as a source of integration activity (Brown et al., 1987; Ellison et al., 1990; Farnet and Haseltine, 1990; Hansen et al., 1999). PICs are replication intermediates that can be isolated from infected cells and which contain the viral cDNA, integrase, and other viral and cellular proteins (Bukrinsky et al., 1993; Farnet and Bushman, 1997; Gallay et al., 1995; Miller et al., 1997). PICs were incubated with purified genomic DNA in vitro to allow integration, then junctions between viral and cellular DNA sequences were cloned and sequenced, yielding 111 control integration sequences. Comparison of sequences generated by integration in vivo and in vitro allows possible biases due to the cloning protocol to be detected and highlights the effects of the intracellular environment.

HIV Integration Is Favored in Genes

We mapped the integration site sequences on the draft human genome (Lander et al., 2001; Venter et al., 2001) and asked (1) whether integration sites were correlated with features mapped on the genome sequence, and (2) whether integration sites were correlated with each other. The distribution of integration sites is shown mapped on the human chromosomes in Figure 1.

Analysis of the placement of the integration sites made by infection in vivo revealed that 69% resided in transcription units, a highly significant departure from random placement ($p < 10^{-11}$; a detailed description of our statistical methods is available in the Supplemental Data at http://www.cell.com/cgi/content/full/110/4/521/DC1). The populations of sites made with HIV-1 or the HIV-derived vector could be analyzed separately, revealing that each favored integration in genes (86% for HIV-1 analyzed alone and 67% for the HIV-based vector; the difference is not statistically significant). For the control in vitro integration sites, 35% were in transcription units. The human genome is about 33% transcription units (Lander et al., 2001; Venter et al., 2001), so the frequency of integration in genes in vitro is not significantly different from the frequency expected for random placement of sites ($p = 0.76$). For the in vivo population of integration sites, gene-dense regions and gene-rich chromosomes were also favored for integration, and even those integration events outside of genes tended to be in gene-rich regions. No functional class of genes was obviously favored or disfavored for integration.

Within genes, integration was favored in introns over exons, likely a result of the greater size of introns compared to exons (Lander et al., 2001; Venter et al., 2001). More detailed analysis was not attempted due to the incomplete information available on the structure of the human genes. All targeted genes were predicted to be transcribed by RNA polymerase II. There was no significant correlation between the direction of viral transcription and the direction of transcription in genes that hosted integration events ($p = 0.3$ for the hypothesis of correlation). Preliminary studies of integration sites in human peripheral blood mononuclear cells also showed favoring of integration in genes, indicating that this tendency is not unique to the SupT1 cell line studied here (R. Mitchell, A.R.W.S., P.S., H.C., C.B., J.R.E., and F.B., unpublished data).

Gene Activity and HIV Integration

To investigate whether integration site selection was influenced by transcriptional activity, we analyzed the SupT1 target cells by transcriptional profiling. RNA was harvested from SupT1 cells, labeled, and analyzed using the Affymetrix U95A chip, which assays about 12,000 human genes. Of the 524 integration sites studied in the in vivo population, 358 were in genes. Of those, 179 were in genes that were assayed on the chip. The median expression level of this group of genes, using the "average difference" expression metric, was 1300. In contrast, the median for all the genes tested on the chip was 700. For the in vitro control, only 35% of integration sites were in genes (39 total), and only 18 of these were assayed by the chip. The median average difference for this group was 616, close to the chip average.

These data suggested that transcription may be correlated with favored integration, but a statistically rigorous assessment was needed. All of the genes assayed by the chip were divided into eight equal sets by relative expression level in SupT1 cells. The genes targeted for integration in vivo were then distributed into the same "bins" and summed (Figure 2A). A strong trend was found, in which genes hosting integration events were scarce in the lowest expression categories and enriched in the highest categories ($p < 0.0001$), indicating that transcriptional activity correlates with integration site selection. No significant trend was seen with the control in vitro integration sites, though the number analyzed was lower.

We next assessed the correlation between integration site placement and gene activity after infection of SupT1 cells with the HIV vector. Cells are known to change their transcriptional programs within 30 min after HIV infection (Arendt and Littman, 2001; Corbeil et al., 2001; Geiss et al., 2000), well before integration can take place (Butler et al., 2001). How infection initiates signaling and leads to transcriptional changes is unclear. Some possibilities include: (1) signaling due to engagement of CD4 and coreceptor by the viral envelope protein (Davis et al., 1997; Popik and Pitha, 2000), (2) activation of signaling pathways by nef (Simmons et al., 2001), (3) activation of the interferon response (Corbeil et al., 2001), (4) activation of the stress response (Corbeil et al., 2001), and (5) activation of the DNA damage response accompanying synthesis of the unintegrated viral DNA (Temin et al., 1980; Li et al., 2001).

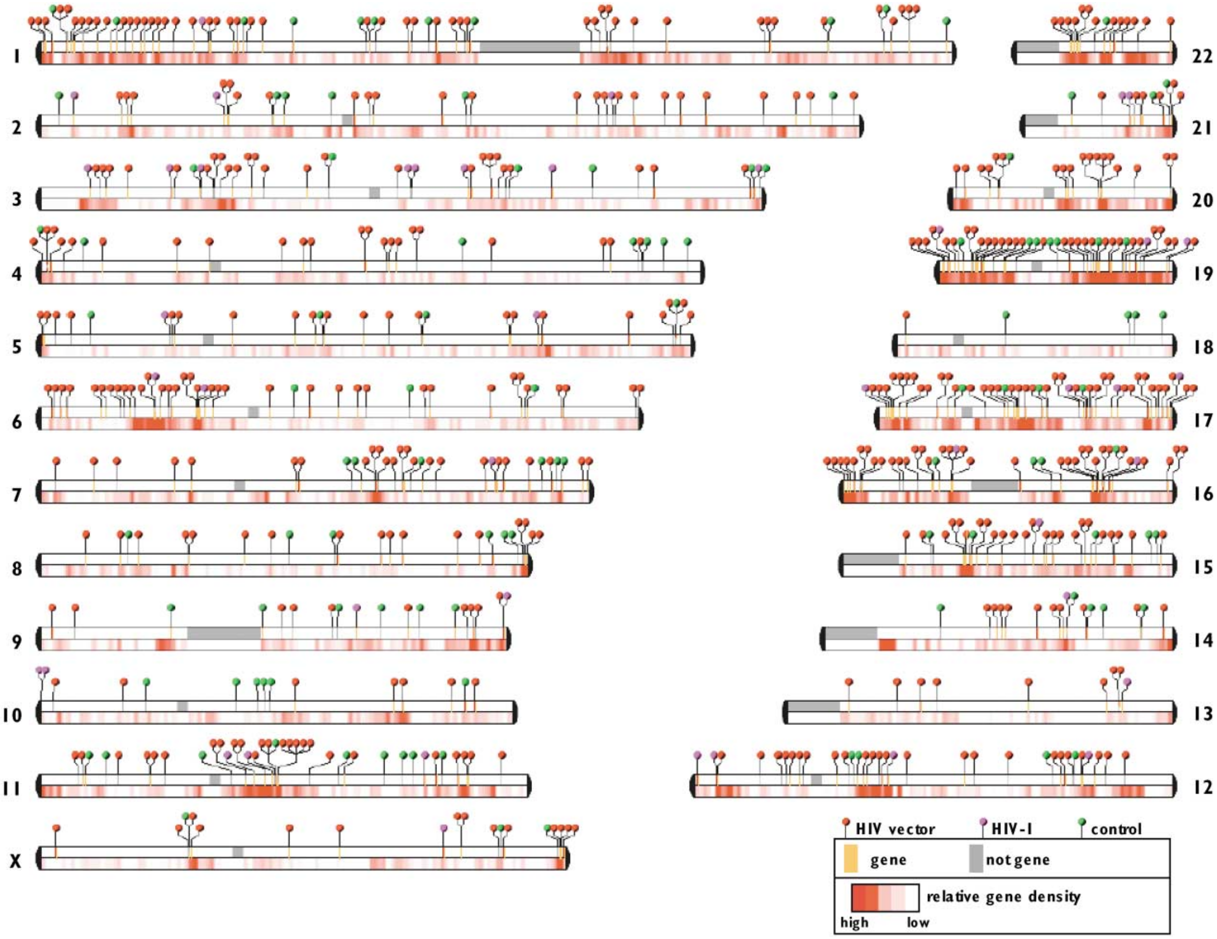

Figure 1. Sites of HIV-1 cDNA Integration in the Human Genome

Locations of chromosomal sequences matching HIV-1 integration site clones are shown as "lollipops" above the linear chromosomes. Purple indicates HIV-1; red, HIV-based vector; and green, PIC (in vitro control). The human chromosomes are shown numbered. For each chromosome, the color of the dashes on the upper bar indicates integration within genes (gold) or outside genes (gray). The lower bar indicates relative gene density, with more-gene-dense regions shown as a more intense red. Centromeres are shown by the gray rectangles. Karyotype analysis showed that the Y chromosome is not present in the SupT1 cells studied and the representation of chromosomes was roughly equal in the cells analyzed (data not shown).

We thus asked whether the infection-induced pattern of transcription correlated more strongly with integration than the uninfected cell pattern (Figure 2B). Cells were infected with the HIV-based vector, RNA was harvested 48 hr later (the same time at which DNA was harvested for cloning of integration sites), and samples were assayed using the Affymetrix U95A chip. Extensive transcriptional changes were seen after infection, with about 8% of genes showing a 2-fold or greater increase in expression and 9% showing as similar decrease. Genes involved in transcription, DNA repair, signaling, and metabolism were notable among those affected. Comparison with a previous study of transcription in CEM cells infected by replication-competent HIV-1 (Corbeil et al., 2001) allowed a set of genes to be identified that were affected in both experiments, possibly representing a "core" set of genes responding to the infection process (a list of these genes is available in the Supplemental Data at http://www.cell.com/cgi/content/full/110/4/521/DC1).

Statistical analysis revealed a correlation between integration site placement and transcription in infected cells that was even stronger than the correlation with data from transcription in uninfected cells ($p < 0.0001$). The difference in the trends in infected versus uninfected cells was also highly significant ($p < 0.0001$). The scatter plot in Figure 2C highlights the differences in gene activity between infected and uninfected cells. This analysis illustrates that the average expression rank is increased for genes targeted for integration compared to the SupT1 population as a whole and that the average expression rank for targeted genes is higher in infected cells than in uninfected cells.

Regional Hotspots for HIV Integration

Regional hotspots for integration were also detected. The most favored region found was an intergenic locus in chromosome 11q13, which contained five independent integration sites within 2.4 kb (Figure 3A). To document clustering in the full data set, we compared the distribution of lengths of DNA segments between integration sites to the distribution expected under homoge-

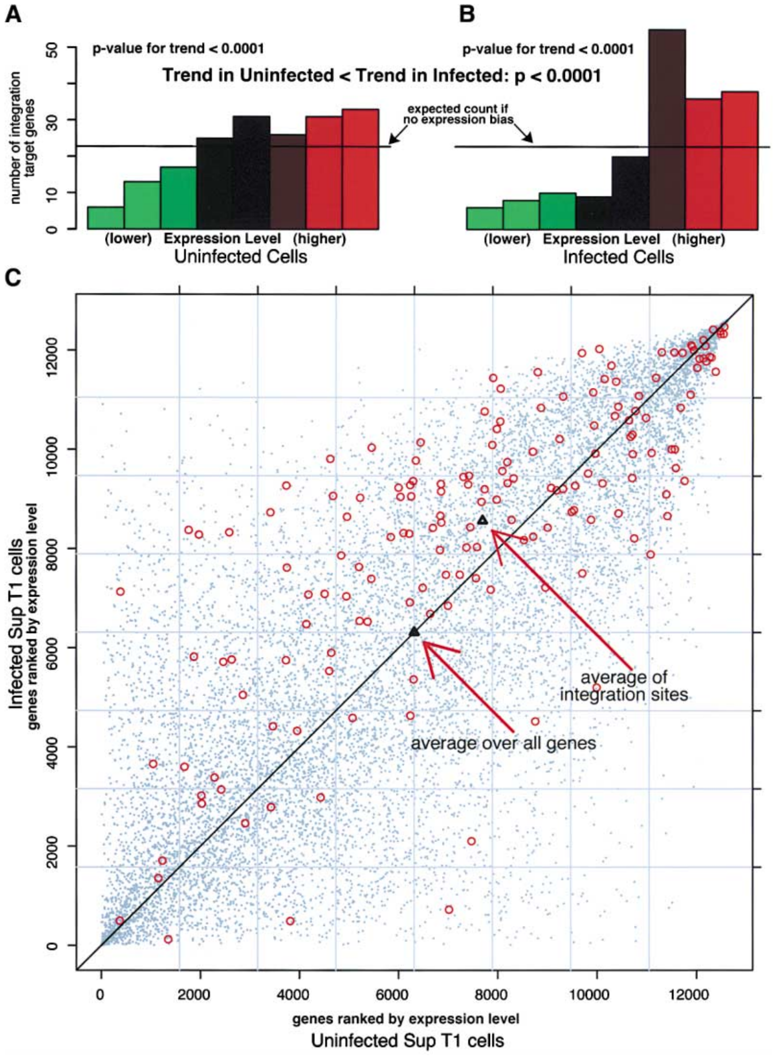

Figure 2. Genes Targeted for Integration In Vivo Analyzed by Transcriptional Profiling

(A) Analysis of expression levels of genes targeted for integration in uninfected SupT1 cells. Expression levels were scored by the average difference value as defined in the Affymetrix Microarray Suite 4.0.1 software package. The roughly 12,000 genes assayed were distributed into eight equal "bins" (1500 genes per bin) by relative expression levels. That is, each bin is defined by a span of average difference values calculated to include one-eighth of all genes assayed on the chip. The bin with the lowest average expression is at the left and the highest at the right. Genes used as integration targets were then distributed into the same bins based on their expression levels and summed. The vertical axes indicate the numbers of genes hosting integration events in each bin. The horizontal line indicates the value expected with no bias, which would be reached if one-eighth of all the genes analyzed were placed in each bin. (B) The same as in (A) but using data from cells analyzed 48 hr after infection with the HIV vector. The trends in (A) and (B) show a highly significant difference ($p < 0.0001$).

(C) Scatter plot comparing results from uninfected and infected cells. All 12,000 genes tested in each experiment were ranked by expression level. Each gene is shown as a gray point, with the position determined by the rank in the uninfected cell data (horizontal axis) and the infected cell data (vertical axis). Red circles indicate genes used as integration targets. Triangles indicate the grand average for all genes and for genes hosting integration events. The average expression level of targeted genes is displaced to the right and upward of the average of all genes, indicating that gene activity was correlated with integration targeting both before and after infection. The average for genes hosting integration events is above the diagonal that designates equal expression in both experiments, indicating that integration is favored in those genes that increased in expression after infection.

neous (random) integration (Figure 3B). The in vivo population of integration sites contained many more short intersegment distances than expected, indicative of clustering. No significant clustering was found in the in vitro control data.

To assess the relationship between gene activity and favored loci, we cataloged and analyzed regions of <100 kb containing three or more integration sites. This yielded seven regional hotspots, four of which contained a gene (data not shown). High local gene density correlated with all regional hotspots, providing a partial explanation for the bias. Regional hotspot function was further probed by quantifying expression of the four targeted genes by fluorescence-monitored RT-PCR. The targeted genes in all four cases were found to be active, and all increased in activity after infection by 2- to 3-fold (data not shown). Four control genes not targeted for integration were tested similarly and found not to increase in activity ($p = 0.03$, Fisher's exact test). These data support the conjecture that some aspect of the activation process itself may promote formation of regional hotspots, though other factors likely contribute as well.

HIV Integration and Human Endogenous Retroviruses (HERVs)

Human endogenous retrovirus (HERV) sequences, which account for 8% of the human genome (Lander et al., 2001), are found predominantly in intergenic regions (Smit, 1999). Genomic positions of human endogenous retroviruses are negatively correlated with sites of HIV integration ($p = 0.0002$ versus $p = 0.24$ for the in vitro control; Table 1). For HIV integration sites within genes, HERV sequences are very infrequent at the site of integration, while for integration sites outside of genes, the trend is much weaker (data not shown). Thus, the negative correlation between HERVs and HIV integration sites is largely explained by favoring of HIV integration in genes.

HERVs and HIV proviruses located within genes also

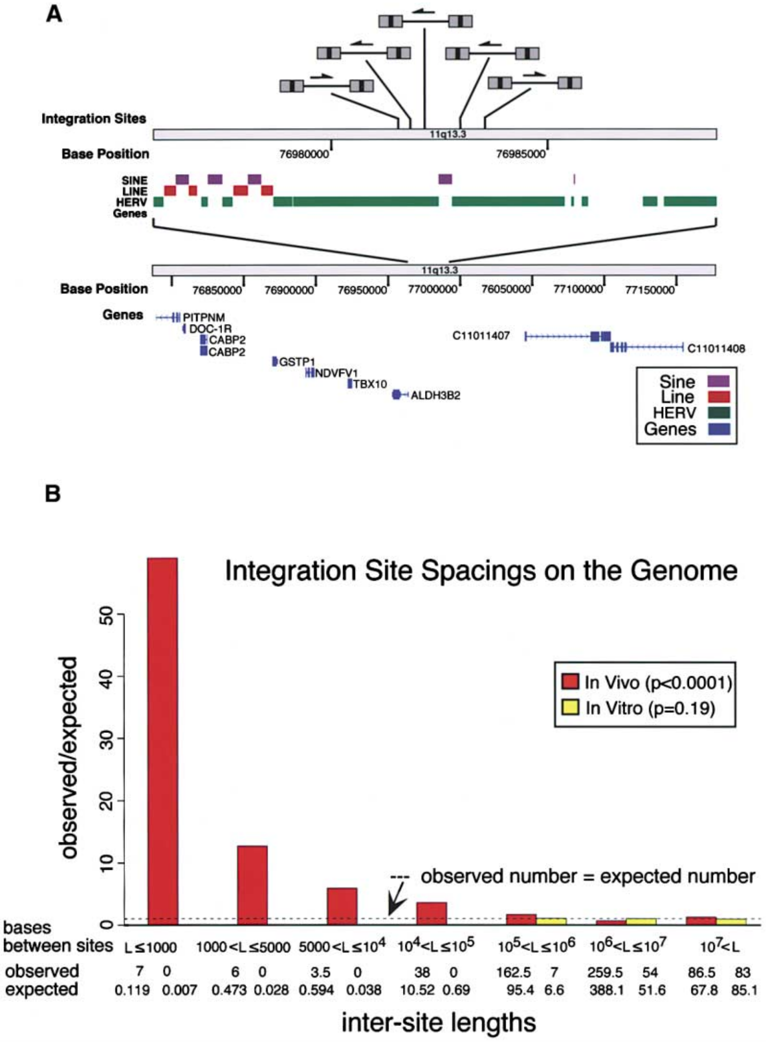

Figure 3. Regional Hotspots for Integration

(A) The 11q13 regional hotspot. Integration sites are shown at the top, with the direction of HIV transcription indicated by the arrow. The upper coordinate indicates the base position (August 1, 2001, freeze of the sequence, UCSC) and local repeated sequences. The lower coordinate indicates nearby genes.

(B) Analysis of integration site spacing on the human genome. Integration site clustering was assessed by comparing the spacing between integration sites to the same number of uniformly distributed (random) sites. Distances between sites are collected in eight length "bins," with the shortest intersite lengths to the left and the longest to the right. To highlight differences between the experimental data sets and the calculated random data set, the data are plotted as the ratio of the observed count for each spacing category divided by the expected count under uniform distribution. The distances between sites and the numbers of sites are as indicated below the bar graph. Red indicates in vivo integration sites and yellow the in vitro control. If no clustering of integration sites was detected, then the observed would equal the expected, and the bar height would be 1 (dotted line); the greater height of the red bars at shorter distances between sites indicates clustering.

differ by their orientation relative to gene transcription. The minority of HERV sequences found within genes are predominantly in reverse orientation relative to the direction of gene transcription, which is expected to prevent transcription termination at HERV poly(A) addition sites (Smit, 1999). In contrast, no significant directional bias is seen for integrated HIV proviruses in genes ($p = 0.3$ for the hypothesis of correlation). The modern HERV distribution probably resulted from selection against host genomes containing proviruses that inter-

Table 1. Chromosomal Features Associated with HIV-1 Integration Sites

Chromosomal Feature	Percent in Human Genome	Percent at In Vivo Integration Sites	Percent at In Vitro Integration Sites
Transcription units	∼33%[a]	69% ($p < 0.0001$)	35% ($p = 0.76$)
SINES			
Alu	10.6%	15.9% ($p = 0.001$)	13.2% ($p = 0.43$)
MIR	2.2%	0.7% ($p = 0.03$)	0.8% ($p = 0.47$)
DNA elements	2.8%	2.2% ($p = 0.46$)	0.8% ($p = 0.29$)
LTR elements (HERV)	8.3%	3.7% ($p = 0.0002$)	6.6% ($p = 0.61$)
LINE	20%	17.0% ($p = 0.10$)	16.5% ($p = 0.4$)
Satellite			
alpha Satellite	UN	0.4%	1.7%
beta Satellite	UN	0%	1.7%

The integration sites studied included those mapped to unique locations on the genome and those in identifiable repeats. p values are for comparison of each integration site population to the human genome.

Abbreviation: UN, unknown.

[a] Estimated value

fered with gene function, accounting both for the accumulation outside genes and the orientational bias. This comparison emphasizes the different requirements for persistence in the primate lineage characteristic of HERVs versus aggressive replication by HIV.

HIV Integration and Other Repeated Sequences

HIV integration was favored in Alu elements ($p = 0.001$), potentially because Alu elements are enriched in gene-rich regions (Table 1; Lander et al., 2001; Stevens and Griffith, 1996; Venter et al., 2001). No bias was seen in favor of LINE elements (contrary to Stevens and Griffith, 1994). MIR elements were underrepresented at integration sites in vivo for unknown reasons.

The frequency of satellite sequences was lower in the in vivo data set than in the in vitro set ($p = 0.012$). Satellite DNA at centromeres and telomeres is known to be packaged in distinctive heterochromatin. As suggested previously, wrapping of target DNA in heterochromatin probably disfavors integration (Carteau et al., 1998).

None of the data on integration in repeated sequences for the control in vitro data set showed a significant departure from the genome average (Table 1). Thus, integration in vitro apparently sampled the naked chromosomal target DNA without detectable biases. The observation of apparently random integration in vitro argues strongly against possible artifactual biases introduced during the isolation and analysis of integration sites. This finding provides important support for the significance of the strong biases detected in the in vivo integration data.

Discussion

We report that sites of HIV integration in the human genome are not randomly distributed but instead are enriched in active genes and regional hotspots. The availability of the human genome sequence was crucial for this study, allowing a much more straightforward and quantitative analysis of integration site selection than has been possible previously. Going forward, as new chromosomal features are mapped onto the genome sequence, it will be possible to assess their possible influence on HIV integration by comparison with the data set reported here.

How does gene activity favor integration? Integration may be promoted by increased chromatin accessibility in transcribed regions, thereby removing inhibitory effects of an unfavorable chromatin environment. Alternatively, integration may be promoted at active genes by favorable interactions between PICs and locally bound transcription factors, as has been suggested for integration targeting by yeast retrotransposons (Ji et al., 1993; Kirchner et al., 1995). A further possibility is that the intranuclear environment of active genes is conducive to integration. Whether transcription promotes integration directly or is correlated indirectly is unclear. Our finding of favored integration in active genes comes as a surprise, since the prevailing view in the field has been that active transcription disfavors integration (Weidhaas et al., 2000). The Weidhass study, however, examined integration by ALV into a single model gene expressed at different levels, whereas the study presented here assayed targeting by HIV genome-wide. Possibly the different experimental approaches emphasized different aspects of target selection. For example, the tendency seen in the HIV study might be due to favorable interactions between PIC components and transcription initiation proteins, as suggested by the studies of yeast retrotransposons (Ji et al., 1993; Kirchner et al., 1995; Boeke and Devine, 1998). However, polymerase passage itself might actually disfavor integration, possibly by steric collision with PICs attempting integration, explaining the ALV data. Alternatively, ALV may differ from HIV in its biases for chromosomal target sites. Further studies will be needed to address these issues.

Integrating into active genes may have evolved to facilitate efficient HIV gene expression after infection. Verdin and coworkers have reported that integration of HIV at different chromosomal loci correlates with quite different levels of gene expression (Jordan et al., 2001). Differences could be attributed to the local chromatin environment—thus, integration targeting to active genes may be important for efficient expression of the HIV genome.

Some aspects of the molecular nature of regional integration hotspots are suggested by our data. Regional hotspots lie in regions enriched for active genes, indicating that those forces favoring integration in genes may favor integration in regional hotspots as well. For genes at 100 kb regions hosting three or more integration events, all were activated by infection with HIV, suggesting further that the activation process itself may be favorable. For example, proteins may bind to genes during the activation process that promote integration. However, the 11q13 hotspot is in an intergenic region, and the favoring of integration in gene-rich regions does not fully account for the quantitative "risk" of serving as a regional hotspot, emphasizing that additional factors likely play a role.

Eukaryotic retrotransposons are known that target Pol I or Pol III transcribed genes for integration, but HIV favors Pol II. Pol II genes are the most abundant class, limiting the conclusions on targeting that can be drawn, but there clearly is not a dominant bias in favor of Pol I or Pol III genes for HIV. These differences in targeting reflect differences in retroelement replication strategies. The Ty1–4 retrotransposons of *Saccharomyces cerevisiae* integrate predominantly upstream of Pol III transcribed genes, which are benign sites because Pol III transcription is not disrupted by integration (Boeke and Devine, 1998). Similarly, the R1 and R2 non-LTR retrotransposons of insects target Pol I transcription units, which are again benign targets because the ribosomal genes are highly repeated (Burke et al., 1993). The Ty, R1, and R2 elements maintain evolutionarily stable relationships with the host organism. HIV, in contrast, targets Pol II transcribed genes, which may help maximize gene expression but at the expense of increased toxicity for the host. Consistent with this idea, HERV elements, which have persisted long-term in the primate lineage, have accumulated primarily in benign sites outside genes, and those HERVs that are in genes are in the antisense orientation, thereby avoiding termination of gene transcription at HERV polyA addition sites (Smit, 1999).

The findings reported here suggest possible means for optimizing retrovirus-based technology. Retroviruses are widely used as insertional mutagens (Gaiano et al., 1996; Hartung et al., 1986; Zijlstra et al., 1989), so mutagenic spectra might be widened if retroviruses with different target preferences can be identified. Recent studies raise the possibility that ALV may have a different integration preference than does HIV—if confirmed this might be exploited for improved mutagenesis. The safety of retroviral vectors in human gene therapy may be increased by taking into account the integration target preferences described here. For example, data on preferred integration sites could guide the choice of gene-delivery vectors to minimize possible toxicity from integration and inform surveillance for possible malignancy due to integration at characteristic hotspots.

Experimental Procedures

Cloning Sites of HIV-1 Integration

Infection was carried out with a VSV-G pseudotyped HIV vector (multiplicity of infection about 1 as determined by quantitative Alu-PCR) (Butler et al., 2001). The HIV-vector particles used were produced from the cell line SODk1CG2 (described in Kafri et al., 1999; Hansen et al., 1999). Note that the gag-pol component of the vector system used (pPTK) encoded all of the HIV-1 auxiliary genes (e.g., vif, vpr, vpu, nef, tat, rev), so any effects of these on integration targeting were maintained in vector infections. The HIV-1 virus stocks used were derived by transfection of a plasmid encoding the R9 strain (Carteau et al., 1998; Swingler et al., 1997). Infected cell DNA was purified using the DNeasy Tissue Kit (Qiagen, Valencia, CA) and cleaved with restriction enzymes that do not cut within the HIV-based vector genome (AvrII, SpeI, and NheI). Linkers were ligated onto the cleaved DNA and sequences were amplified using one primer that bound to the linker DNA and one that bound to the HIV cDNA. PCR products were diluted 1:1000 and amplified with nested primers, and then amplification products were gel-isolated. The structure of the linker forced the PCR to initiate in the HIV sequences, suppressing amplification of DNAs lacking integrated HIV. Methods are essentially as described (GeneWalker Kit, Clontech, Palo Alto, CA) using oligonucleotides described in Supplemental Table S1 at http://www.cell.com/cgi/content/full/110/4/521/DC1. Integration sites from infections with HIV-1 are as in Carteau et al. (1998).

Sequence matches (identified using BLAT, UCSC Human Genome Project Working Draft, December 2000 freeze) were judged to be authentic only if a match to the human genome (1) started at the junction with the HIV terminal (5′-CA-3′) sequence, (2) extended over the length of the high-quality sequence with average identity >98%, and (3) yielded a unique best hit in the BLAT ranking. Identical sequences from different clones were judged to represent multiple isolates of a single integration event. Of 642 sequences analyzed for the in vivo infections, 524 could be placed on the genome, 16 showed matches to multiple locations in the human genome, and 102 sequences did not yield a high-quality match to the genome and were excluded as low-quality sequence reads, sequences too short to determine a unique placement, or integration events in parts of the human genome that are still unsequenced.

An integration target sequence was scored as a part of a transcription unit if it was (1) a member of the Refseq set of well-studied genes (http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html) or (2) if it was predicted to be a transcription unit by the ENSEMBLE (http://www.ensembl.org) or Fgenesh++ (http://www.softberry.com/Help/fgeneshplus2.htm) programs and if that assignment was supported by mRNA or spliced EST sequence evidence. Repeated sequences were identified using RepeatMasker analysis of the December 2000 genome draft. Gaps in the human genome sequences were removed and flanking sequences joined to facilitate statistical analysis of integration site placement (discussed in the Supplemental Data at http://www.cell.com/cgi/content/full/110/4/521/DC1). Integration site sequences in the human genome have been deposited in GenBank (accession numbers BH609398–BH610085).

Preparation of the Control Population of In Vitro Integration Sites

Purification of vector-derived HIV PICs used for the in vitro control was carried out as described (Hansen et al., 1999). In vitro integration was achieved by incubating 250 μl of PIC extract with 1 μg of SupT1 genomic DNA for 45 min at 37°C. The integration product was recovered by incubating with proteinase K in 0.5% sodium dodecyl sulfate followed by extraction with phenol-chloroform and ethanol precipitation. Cloning, sequencing, and analysis were as for the in vivo integration site population.

Microarray Analysis

RNA was harvested from SupT1 cells in log phase growth. For the analysis of infected cells, infections were carried out with the HIV-based vector (stocks were generated by transfection as in Kafri et al., 1999; Hansen et al., 1999; Follenzi et al., 2000) at a multiplicity of 1, and RNA was harvested 48 hr later. Labeling of RNA was performed as described by Affymetrix (Santa Clara, CA). Ten micrograms of cRNA were used per Affymetrix HU95A array, which assays about 12,000 human transcripts (specifically 12,625 transcripts, but including some duplicates and controls). Two chips were used for each experimental condition and the average used for subsequent analysis (using GeneSpring.4.0, Silicon Genetics, Redwood City, CA). For the in vivo integration sites, we analyzed 179 integration sites in 166 genes that were present on the chip (two genes had three hits, and nine genes had two hits). For genes in the in vitro set, 18 integration sites could be analyzed, the low number being due to the relatively lower frequency of integration in genes the in vitro collection. Six genes analyzed on the chips were also tested by quantitative PCR, which showed good agreement with data from transcriptional profiling. Raw data is available upon request.

Statistical Analysis

A detailed description of the statistical methods used in this study is available as Supplemental Data at http://www.cell.com/cgi/content/full/110/4/521/DC1.

Acknowledgments

We thank Joanne Chory, John Coffin, Leslie Orgel, Inder Verma, Detlef Weigel, and members of the Salk Institute Infectious Disease Laboratory for helpful discussions. We also thank Richard Mitchell for help analyzing gene expression data. This work was supported by NIH grants AI52845 and AI34786 to F.B., the James B. Pendleton Charitable Trust, the Berger Foundation, the Fritz B. Burns Foundation, the Gertrude E. Skelly Foundation, and Cornelia Mackey. A.R.W.S. was supported by a grant from the Deutsche Forschungsgemeinschaft.

Received: May 31, 2002
Revised: July 17, 2002

References

Arendt, C.W., and Littman, D.R. (2001). HIV: master of the host cell. Genome Biol. *2*, reviews 1030.1–1030.4.

Boeke, J.D., and Devine, S.E. (1998). Yeast retrotransposons: finding a nice quiet neighborhood. Cell *93*, 1087–1089.

Brown, P.O., Bowerman, B., Varmus, H.E., and Bishop, J.M. (1987). Correct integration of retroviral DNA in vitro. Cell *49*, 347–356.

Bukrinsky, M.I., Sharova, N., McDonald, T.L., Pushkarskaya, T., Tarpley, G.W., and Stevenson, M. (1993). Association of integrase, matrix, and reverse transcriptase antigens of human immunodeficiency virus type 1 with viral nucleic acids following acute infection. Proc. Natl. Acad. Sci. USA *90*, 6125–6129.

Burke, W.D., Eickbush, D.G., Xiong, Y., Jakubczak, J., and Eickbush, T.H. (1993). Sequence relationship of retrotransposable elements R1 and R2 within and between divergent insect species. Mol. Biol. Evol. *10*, 163–185.

Bushman, F.D. (1994). Tethering human immunodeficiency virus 1 integrase to a DNA site directs integration to nearby sequences. Proc. Natl. Acad. Sci. USA *91*, 9233–9237.

Bushman, F.D. (2001). Lateral DNA Transfer: Mechanisms and Consequences (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press).

Bushman, F.D., and Craigie, R. (1992). Integration of human immunodeficiency virus DNA: adduct interference analysis of required DNA sites. Proc. Natl. Acad. Sci. USA *89*, 3458–3462.

Butler, S., Hansen, M., and Bushman, F.D. (2001). A quantitative assay for HIV cDNA integration in vivo. Nat. Med. *7*, 631–634.

Carteau, S., Hoffmann, C., and Bushman, F.D. (1998). Chromosome structure and HIV-1 cDNA integration: centromeric alphoid repeats are a disfavored target. J. Virol. *72*, 4005–4014.

Coffin, J.M., Hughes, S.H., and Varmus, H.E. (1997). Retroviruses (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press).

Corbeil, J., Sheeter, D., Genini, D., Rought, S., Leoni, L., Du, P., Ferguson, M., Masys, D.R., Welsh, J.B., Fink, J.L., et al. (2001). Temporal gene regulation during HIV-1 infection of human CD4+ T cells. Genome Res. *11*, 1198–1204.

Davis, C.B., Dikic, I., Unutmaz, D., Hill, C.M., Arthos, J., Siani, M.A., Thompson, D.A., Schlessinger, J., and Littman, D.R. (1997). Signal transduction due to HIV-1 envelope interactions with chemokine receptors CXCR4 or CCR5. J. Exp. Med. *186*, 1793–1798.

Ellison, V.H., Abrams, H., Roe, T., Lifson, J., and Brown, P.O. (1990). Human immunodeficiency virus integration in a cell-free system. J. Virol. *64*, 2711–2715.

Farnet, C.M., and Haseltine, W.A. (1990). Integration of human immunodeficiency virus type 1 DNA in vitro. Proc. Natl. Acad. Sci. USA *87*, 4164–4168.

Farnet, C., and Bushman, F.D. (1997). HIV-1 cDNA integration: requirement of HMG I(Y) protein for function of preintegration complexes in vitro. Cell *88*, 1–20.

Follenzi, A., Ailes, L.E., Bakovic, S., Gueuna, M., and Naldini, L. (2000). Gene transfer by lentiviral vectors is limited by nuclear translocation and rescued by HIV-1 pol sequences. Nat. Genet. *25*, 217–222.

Gaiano, N., Amsterdam, A., Kawakami, K., Allende, M., Becker, T., and Hopkins, N. (1996). Insertional mutagenesis and rapid cloning of essential genes in zebrafish. Nature *383*, 829–832.

Gallay, P., Swingler, S., Song, J., Bushman, F., and Trono, D. (1995). HIV nuclear import is governed by the phosphotyrosine-mediated binding of matrix to the core domain of integrase. Cell *17*, 569–576.

Geiss, G.K., Bumgarner, R.E., An, M.C., Agy, M.B., van't Wout, A.B., Hammersmark, E., Carter, V.S., Upchurch, D., Mullins, J.I., and Katze, M.G. (2000). Large-scale monitoring of host cell gene expression during HIV-1 infection using cDNA microarrays. Virology *266*, 8–16.

Hansen, M.S.T., Smith, G.J.I., Kafri, T., Molteni, V., Siegel, J.S., and Bushman, F.D. (1999). Integration complexes derived from HIV vectors for rapid assays in vitro. Nat. Biotechnol. *17*, 578–582.

Hartung, S., Jaenisch, R., and Breindl, M. (1986). Retrovirus insertion inactivates mouse a1(I) collagen gene by blocking initiation of transcription. Nature *320*, 365–367.

Ji, H., Moore, D.P., Blomberg, M.A., Braiterman, L.T., Voytas, D.F., Natsoulis, G., and Boeke, J.D. (1993). Hotspots for unselected Ty1 transposition events on yeast chromosome III are near tRNA genes and LTR sequences. Cell *73*, 1–20.

Jordan, A., Defechereux, P., and Verdin, E. (2001). The site of HIV-1 integration in the human genome determines basal transcriptional activity and response to Tat transactivation. EMBO J. *20*, 1726–1738.

Kafri, T., van Praag, H., Ouyang, L., Gage, F.H., and Verma, I.M. (1999). A packaging cell line for lentiviral vectors. J. Virol. *73*, 576–584.

Katz, R.A., Gravuer, K., and Skalka, A.M. (1998). A preferred target DNA structure for retroviral integrase in vitro. J. Biol. Chem. *273*, 24190–24195.

Katz, R.A., DiCandeloro, P., Kukolj, G., and Skalka, A.M. (2001). Role of DNA end distortion in catalysis by avian sarcoma virus integrase. J. Biol. Chem. *276*, 34213–34220.

Kirchner, J., Connolly, C.M., and Sandmeyer, S.B. (1995). In vitro position-specific integration of a retroviruslike element requires Pol III transcription factors. Science *267*, 1488–1491.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

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

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

Miller, M.D., Farnet, C.M., and Bushman, F.D. (1997). Human immunodeficiency virus type 1 preintegration complexes: studies of organization and composition. J. Virol. *71*, 5382–5390.

Mooslehner, K., Karls, U., and Harbers, K. (1990). Retroviral integration sites in transgenic Mov mice frequently map in the vicinity of transcribed DNA regions. J. Virol. *64*, 3056–3058.

Panet, A., and Cedar, H. (1977). Selective degradation of integrated murine leukemia proviral DNA by deoxyribonucleases. Cell *11*, 933–940.

Popik, W., and Pitha, P.M. (2000). Exploitation of cellular signaling by HIV-1: unwelcome guests with master keys that signal their entry. Virology *276*, 1–6.

Pruss, D., Bushman, F.D., and Wolffe, A.P. (1994a). Human immunodeficiency virus integrase directs integration to sites of severe DNA distortion within the nucleosome core. Proc. Natl. Acad. Sci. USA *91*, 5913–5917.

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

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

Pryciak, P., Muller, H.-P., and Varmus, H.E. (1992). Simian virus 40 minichromosomes as targets for retroviral integration in vivo. Proc. Natl. Acad. Sci. USA *89*, 9237–9241.

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

Scherdin, U., Rhodes, K., and Breindl, M. (1990). Transcriptionally active genome regions are preferred targets for retrovirus integration. J. Virol. *64*, 907–912.

Scottoline, B.P., Chow, S., Ellison, V., and Brown, P.O. (1997). Disruption of the terminal base pairs of retroviral DNA during integration. Genes Dev. *11*, 371–382.

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

Simmons, A., Aluvihare, V., and McMichael, A. (2001). Nef triggers a transcriptional program in T cells imitating single-signal T cell activation and inducting HIV virulence mediators. Immunity *14*, 763–777.

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

Stevens, S.W., and Griffith, J.D. (1994). Human immunodeficiency virus type 1 may preferentially integrate into chromatin occupied by L1Hs repetitive elements. Proc. Natl. Acad. Sci. USA *91*, 5557–5561.

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

Swingler, S., Gallay, P., Camaur, D., Song, J., Abo, A., and Trono, D. (1997). The Nef protein of human immunodeficiency virus type 1 enhances serine phosphorylation of the viral matrix. J. Virol. *71*, 4372–4377.

Temin, H.M., Keshet, E., and Weller, S.K. (1980). Correlation of transient accumulation of linear unintegrated viral DNA and transient cell killing by avian leukosis and reticuloendotheliosis viruses. Cold Spring Harb. Symp. Quant. Biol. *44*, 773–778.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science *291*, 1304–1351.

Vijaya, S., Steffan, D.L., and Robinson, H.L. (1986). Acceptor sites for retroviral integrations map near DNaseI-hypersensitive sites in chromatin. J. Virol. *60*, 683–692.

Weidhaas, J.B., Angelichio, E.L., Fenner, S., and Coffin, J.M. (2000). Relationship between retroviral DNA integration and gene expression. J. Virol. *74*, 8382–8389.

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

Zijlstra, M., Li, E., Sajjadi, F., Subramani, S., and Jaenisch, R. (1989). Germ-line transmission of a disrupted beta 2-microglobulin gene produced by homologous recombination in embryonic stem cells. Nature *342*, 435–438.