GENOME-WIDE ANALYSIS OF RETROVIRAL DNA INTEGRATION

Frederic Bushman, Mary Lewinski, Angela Ciuffi, Stephen Barr, Jeremy Leipzig, Sridhar Hannenhalli and Christian Hoffmann

Abstract | Retroviral vectors are often used to introduce therapeutic sequences into patients' cells. In recent years, gene therapy with retroviral vectors has had impressive therapeutic successes, but has also resulted in three cases of leukaemia caused by insertional mutagenesis, which has focused attention on the molecular determinants of retroviral-integration target-site selection. Here, we review retroviral DNA integration, with emphasis on recent genome-wide studies of targeting and on the status of efforts to modulate target-site selection.

One of the defining features of retroviral replication is the covalent integration of viral DNA into the host-cell genome (FIG. 1), a process with profound consequences for both the virus and the host (for reviews see REFS 1,2). Most primary DNA sequences can function as acceptor sites for retroviral integration *in vitro*[3–5] and, *in vivo*, the influence of the primary DNA sequence of the target site on integration is weak, thus target-site selection is primarily not sequence-specific[6–9]. However, chromosomal integration-site selection of retroviruses *in vivo* is not random[10–12]. Now that near-complete sequences are available for several vertebrate genomes, it has become possible to analyse integration targeting in a statistically rigorous manner by sequencing junctions between proviral and host-cell DNA. Surprisingly, the three viruses studied in detail so far — HIV, avian sarcoma-leukosis virus (ASLV) and murine leukaemia virus (MLV) — show different patterns of favoured and disfavoured target sites.

Understanding the mechanisms that dictate integration-site selection is important to both the understanding of basic retrovirology and to its clinical applications. In the field of HIV research, an understanding of the molecular mechanisms that control integration-site selection could reveal new targets for antiretroviral drugs. In the gene-therapy field, it is now unfortunately clear that integration of therapeutic retroviral vectors can activate proto-oncogenes in patients. The insertional activation of proto-oncogenes has long been studied in animal models[1,2]. More recently, in an otherwise successful gene-therapy trial in France that used retroviral vectors to treat patients with severe immunodeficiency (X-linked severe combined immunodeficiency; X-SCID), three children developed leukaemia, and one has since died. In two of these cases, in the leukaemic cells, the therapeutic MLV vector was found to have integrated in the 5′ region of the *LMO2* oncogene, which probably contributed to neoplastic transformation[13,14]. So far, the location of integration sites for the third case has not been reported. Therefore, the gene-therapy field is focused on the question of where DNA integration takes place in the targeted genome, and what can be done to minimize the risks of insertional mutagenesis.

This review summarizes the work on retroviral DNA-integration targeting that has been published to date, with an emphasis on genome-wide studies. These studies provide a new route to understanding integration mechanisms, and hint at approaches to modulate integration target-site selection *in vivo*.

University of Pennsylvania School of Medicine, Department of Microbiology, 3610 Hamilton Walk, Philadelphia, Pennsylvania 19104-6076, USA.
Correspondence to F.B.
e-mail: bushman@mail.med.upenn.edu
doi:10.1038/nrmicro1263
Published online
19 September 2005

Early studies of integration targeting

Early sequencing studies of integrated retroviral genomes revealed that there were no strong similarities in the cellular DNA sequences used as integration acceptor sites[1]. Studies of MLV integration sites in cultured cell lines led to the proposal that integration was favoured near accessible chromatin regions in

and around transcription units, as associated features such as DNase I HYPERSENSITIVE SITES[15–17], CPG ISLANDS[18] or transcribed sequences[18] were apparently enriched nearby. However, several factors complicate the simple interpretation of these studies, including the small numbers of sites studied, uncontrolled effects of selection during cell isolation, and lack of information on the frequency of these features in vertebrate genomes.

An early survey of the integration sites favoured by ASLV in bird genomes led to the conclusion that certain integration sites were preferred, and that integration took place repeatedly at these preferred sites in avian DNA[19]. However, this result has not been confirmed using other methods[20] and is now deemed unlikely to be correct. In the second study, a PCR-based method was used to analyse integration at different chromosomal regions, and several different genomic regions were found to be favoured with similar frequency[20]. It is unknown how representative these regions were of the entire genome.

Integration target-site selection *in vitro*

The establishment of *in vitro* assays for covalent integration by purified integrase proteins[4,21–23] allowed a detailed assessment of integration. Some of the main conclusions of these studies are summarized below.

Most DNA sequences can act as integration acceptor sites, although more detailed analysis reveals some effects of primary sequence on integration[5,7,24–26]. So far, it seems probable that the sequence of the integration-target DNA has a minor influence on site selection, although detailed bioinformatic studies of *in vivo* sites might reveal new influences.

The presence of a DNA-binding protein on target DNA can block access of integration complexes, creating regions that are refractory to integration[27–29]. Such simple steric hindrance can also be seen in integration *in vivo*[30].

Distortion of the integration-target DNA can strongly promote integration. *In vivo*, target DNA is not naked, but is incorporated into chromatin. Incorporating DNA into nucleosomes *in vitro* does not reduce integration, as might have been expected from a steric hindrance model, but instead creates new 'hot spots' for integration. Analysis of these integration hot spots indicates that these are sites at which DNA is probably distorted owing to the wrapping of DNA around nucleosomes[25,27,31,32]. The distortion of DNA in several other model protein–DNA complexes has also been shown to favour integration[29,33]. These findings are consistent with the idea that DNA distortion is involved in the integrase mechanism, so that predistorting the target DNA favours the integration reaction[34,35].

The nucleosomal templates used in the above examples are all presumed to be present as 10-nm 'beads on a string'. The consequences for integration of incorporating such structures into the 30-nm fibres, or the still higher-order structures that comprise chromosomes, are unknown.

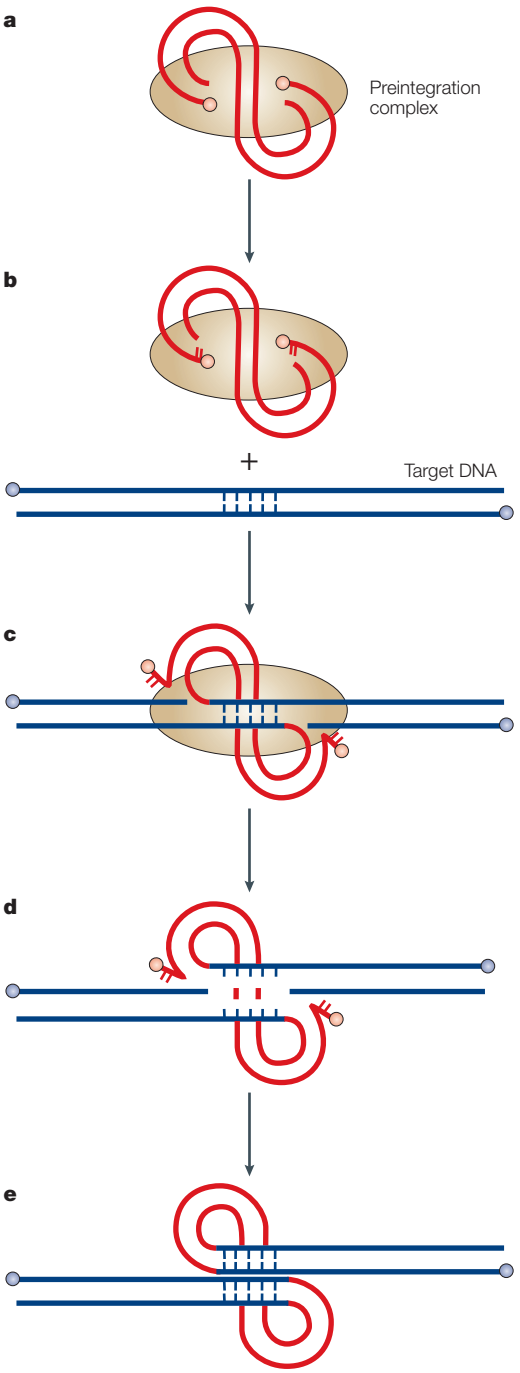


Figure 1 | **The DNA breaking and joining reactions that mediate integration.** The beige oval represents the viral-encoded integrase and other proteins of the preintegration complex, the curved red lines indicate the viral DNA, the straight blue line the chromosomal target DNA. **a** | The immediate product of reverse transcription is a double-stranded complementary DNA (cDNA) molecule with mostly blunt ends. **b** | Prior to integration, the integrase enzyme removes two nucleotides from each 3′ end. **c** | The recessed 3′ hydroxyl groups are then joined to protruding 5′ ends of the target DNA. A single-step transesterification reaction mediates the target-cleavage and joining steps. **d** | Unpairing of the integration intermediate yields gaps at each host–virus DNA junction. **e** | Gap repair, probably by host-cell gap-repair enzymes, completes formation of the integrated provirus.

DNase I HYPERSENSITIVE SITES
DNA sites in chromosomes that show increased sensitivity to digestion by DNase I. These sites probably represent regions of the chromosome that are nucleosome-free, and often correspond to gene-control regions.

CPG ISLANDS
Regions in chromosomes that are enriched in the rare CpG dinucleotide. They often correspond to gene-control regions.

HUMAN ENDOGENOUS RETROVIRUSES
(HERVs). Sequences in human DNA that are derived from infection of the human germ line by retroviruses. They account for about 8% of the human genome sequence.

LONG INTERSPERSED NUCLEAR ELEMENTS
(LINEs). Non-long-terminal-repeat retrotransposons. These comprise the only known type of active transposon in the human genome.

ALU ELEMENTS
Repeated DNA sequences that contain recognition sites for the *Alu* restriction enzyme.

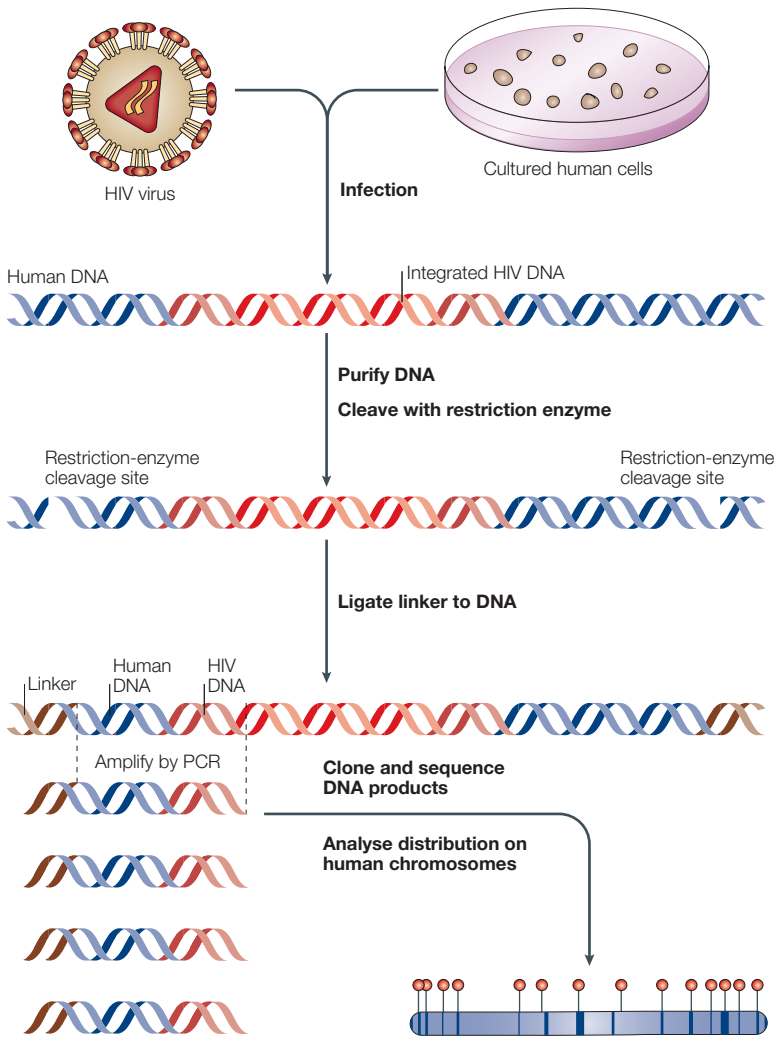
CENTROMERIC HETEROCHROMATIN
The distinctive type of protein–DNA complexes that are found at centromeres.

Box 1 | **Analysing retroviral integration sites in the human genome**

To analyse retroviral integration sites in the human genome, cultured cells are infected with HIV or another retrovirus. DNA from infected cells is isolated, cleaved with restriction enzymes and ligated to DNA linkers. Integration sites are then amplified using one primer that binds to the viral DNA end and another primer that binds to the DNA linker (see the figure). Amplification is carried out a second time with nested primers, and the PCR products, which contain host–virus DNA junctions, are cloned and sequenced[10–12]. Integration sites are mapped on the draft human genome sequence (FIG. 2), and local features at integration sites are quantified.

Various control sites can be used for comparison in these experiments. Probably the best type of control takes advantage of integration *in vitro*. Schroder and colleagues purified DNA from uninfected SupT1 cells and used this DNA as a target for preintegration-complex integration *in vitro*[10]. Reaction products were purified and integration sites were cloned, as for the *in vivo* sites. The analysis then compared the *in vivo* and *in vitro* populations. Statistical analysis revealed that the distribution of integration sites in the *in vitro* population was indistinguishable from random sampling of the human genome[10], supporting the idea that the cloning and analytical methods used did not bias the analysis.

Another type of control takes advantage of random locations in the human genome that are generated computationally. These are used in statistical comparisons with the experimental population. A more sophisticated variation of this approach uses random sites that are selected in a way that takes into account the possible influence of the distribution of restriction-enzyme recognition sites used in the cloning of experimental integration sites (see REF. 12 for details).



Genome-wide studies of integration targeting

Several groups have investigated integration targeting *in vivo* by sequencing junctions between viral and human DNA and analysing their positions in the human genome. So far, the sequences of several thousand retroviral integration sites have been reported. Most were generated by the acute infection of cultured cells with retroviruses or retroviral vectors, followed by mapping and quantification of local features in the draft human-genome sequence (BOX 1; FIG. 2).

HIV favours integration in transcription units

Following large-scale sequence analysis, the position of integration sites in the human genome can be compared with the position of other annotated features (FIG. 2). In the first such study, the distribution of HIV integration sites in the chromosomes of a human lymphoid cell line, SupT1, was investigated[10]. This study showed that genes were favoured targets for HIV integration, and later studies of HIV integration in other cell types reached the same conclusion[11,12]. For example, if the well characterized RefSeq genes are used for comparison, the human genome contains 31.1% genes, whereas HIV integration-site data sets showed frequencies of integration in genes that ranged from 66.1% in SupT1 cells to 73.4% in Jurkat cells (both cell lines originated as tumours of human T cells, which are the natural target of HIV infection). Similar results were obtained using other catalogues of human genes for analysis ($P<0.0001$ for all comparisons).

Further studies investigated whether there were any preferences in the location of HIV integration sites along the length of transcription units[10–12]. No biases were found, although strong biases were seen with MLV (discussed below)[11]. Evidently, the positive influence of transcription units on HIV integration extends across their entire length.

Gene-rich regions in the human genome are depleted in certain classes of repeat DNA sequences (such as HUMAN ENDOGENOUS RETROVIRUSES (HERVs) and LONG INTERSPERSED NUCLEAR ELEMENTS (LINEs)) and enriched in others (such as ALU ELEMENTS). Therefore, a bias in the rates of integration in these different repeat classes is expected. Indeed, HIV integration is disfavoured in HERV elements, which is consistent with favoured HIV integration in genes. Furthermore, integration in Alu elements is favoured in some data sets[10,12]. HIV integration is strongly disfavoured in alphoid repeats (which, in humans, are composed of α-satellite DNA). This indicates that CENTROMERIC HETEROCHROMATIN, the location of most of the α-satellite-containing DNA, is an unfavourable target for HIV integration[7,10]. Centromeric heterochromatin is known to be wrapped tightly by distinctive DNA-binding proteins, and this chromatin environment is unfavourable for the expression of most genes (including those expressed by HIV, see below). We therefore infer that the packing of DNA into centromeric heterochromatin renders it less accessible, and so disfavours integration.

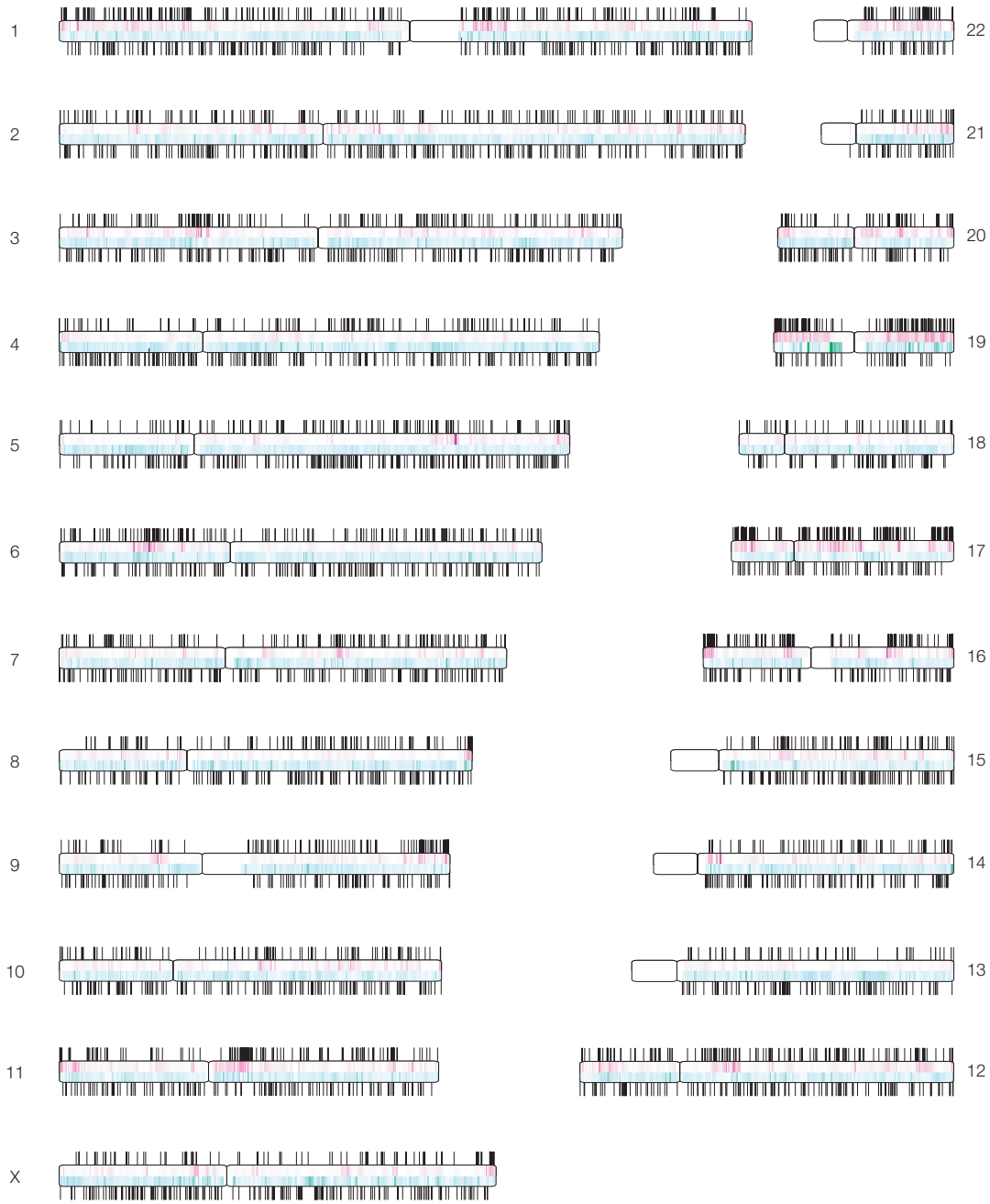


Figure 2 | **Sites of integration of HIV or HIV-based vectors in the human genome.** The human chromosomes are shown as horizontal bars and are numbered. Sites of HIV integration are shown along the top, and control random integration sites generated *in silico* are shown on the bottom. The pink shading in each bar indicates the density of genes, the green shading indicates the density of human endogenous retroviruses (HERVs). Sites of HIV integration correlate with high gene density and low HERV density.

Regions of high gene density in human chromosomes correlate with several other features, including high-level expression per gene, high densities of CpG islands, a GIEMSA-LIGHT BANDING pattern and a high GC content[36–40]. The effects of chromosomal banding patterns and GC content on retroviral integration were analysed statistically, revealing that Giemsa-light banding and a high GC content also correlated with favoured HIV integration[12].

GIEMSA-LIGHT BANDING
Staining of the human chromosomes by the Giemsa procedure results in a pattern of light and dark bands, which roughly corresponds with the relative GC content and gene density.

Effects of transcriptional activity on HIV integration. Transcriptional profiling analysis has been carried out in some of the cell types studied as integration targets, allowing the influence of transcriptional activity on integration-site selection to be assessed. Some of these transcriptional profiling studies were carried out on retrovirus-infected cells[10,12,41], so that the data reflected the influence of infection on cellular gene activity[10,42–44]. Analysis of the microarray

data revealed that the median expression level of genes hosting HIV integration events was consistently higher than the median expression level of all the genes assayed on the microarray. For HIV, the ratios of the expression levels of targeted genes assayed compared to controls ranged from 1.6 to 3.0 (REFS 10,12), and the results were statistically significant for all data sets.

Trancriptional profiling studies were carried out for HIV vector integration in SupT1 cells. This study showed that the trend towards integration in highly expressed genes increased when data from infected cells were used, which indicates that genes that are activated by infection are favoured integration targets[10].

Given that transcriptional activity favours HIV integration in more active genes and that transcription differs in different cell types, to what extent does tissue-specific transcription result in tissue-specific patterns of integration? Mitchell *et al.* compared integration-site data sets from SupT1, peripheral blood mononuclear cells and IMR90 human fibroblasts, for which transcriptional-profiling data were available[12]. Statistical analysis revealed that cell-type-specific transcription did indeed bias integration-site selection, as genes that were relatively more active in a given cell type were more likely to be targeted for HIV integration. However, the differences were quantitatively modest, perhaps because most of the cellular programme of gene activity is shared among many cell types[38,40].

Substructure in chromosomal regions that favour HIV integration. Two lines of evidence indicate that the gene-rich chromosomal regions that are favourable for HIV integration can be subdivided into interspersed favourable and unfavourable segments[12]. In the first study, a computational analysis indicated that relatively short chromosomal regions (100–250 kb) were most favourable for integration, and that these were interspersed with less favourable regions[12]. In the second study, an analysis of HIV integration frequency near CpG islands also indicated interspersed favourable and unfavourable regions. CpG islands are chromosomal regions that are enriched in the rare CpG dinucleotide, and these regions usually correspond to gene-regulatory regions that contain clustered transcription-factor-binding sites. Consequently, CpG islands are enriched in gene-rich regions. For HIV, the region surrounding CpG islands was disfavoured for integration[12] — on average, the several kb that surround CpG islands hosted fewer HIV integration events than expected by chance. Therefore, for HIV, gene-dense regions that favour retroviral integration contain interleaved favourable clusters of active genes and unfavourable regions that include CpG islands. The mechanism by which CpG islands obstruct HIV integration is unclear: there might be specific proteins bound at these sites that block integration, CpG islands might be located in a nuclear compartment that is unfavourable for integration, or some other as-yet-unknown mechanism might be responsible.

Clustering of HIV integration sites. Analysis of integration-site positions in the human genome revealed that HIV integration sites cluster. This is partly a consequence of favoured integration in gene-rich regions. However, some data indicate that there might be additional mechanisms at work. In SupT1 cells, a very 'hot region' was detected in chromosome 11q13, in which five integration sites were found in 2.5 kb (REF. 10). This region is in a gene-rich chromosomal domain, but is not in a known transcription unit. The mechanism of this strong favouring is unknown.

Also, integration events have been recovered several times in a few genes (FIG. 3). An analysis of pooled data on 2,969 HIV integration sites revealed that the CREB-binding-protein gene (*CREBBP*) hosted 11 independent integration events, the F-box and leucine-rich-repeat protein-11 gene (*FBXL11*) hosted 8 independent integration events, and the DNA (cytosine-5-)methyltransferase-1 gene (*DNMT1*) hosted 8 independent integration events. It is not clear whether genes that host many integration sites have special characteristics, or whether they are at the extremes of a Gaussian distribution — this issue is not easy to evaluate statistically[10–12]. Analysis published so far has not indicated that any specific functional class of gene is favoured for integration by HIV, although it is possible that future studies might disclose common features.

HIV integration at the chromosomal level. HIV strongly favours integration in gene-rich human chromosomes[10,12], but there are statistically significant variations in the selection of specific chromosomes in different cell types[12]. This finding hints at possible mechanisms of integration target-site selection that operate at the level of whole chromosomes. For example, it has been suggested that chromosomes occupy non-random positions in the nucleus[45] — if this varies among cell types, then the intranuclear position of chromosomes might influence integration targeting.

Integration targeting by MLV and ASLV

MLV and ASLV show different integration-target preferences in the human genome. In human HeLa cells, 903 sites of MLV integration were characterized and compared with data for HIV[11]. Eighty percent of MLV integration sites were distributed in the genome in a near-random fashion, but 20% were within 5 kb of DNA that contained the 5′ end of a transcription unit[11,12]. In contrast to HIV, integration was also favoured near CpG islands (16.8% of sites within 1 kb). A subsequent paper that examined the integration of MLV-based vectors in haematopoietic stem cells found a similar pattern of preferred sites[46,47].

These findings are significant in evaluating the potential hazards of using MLV as a gene-therapy vector. MLV-derived retroviral vectors are the most commonly used vectors in human gene therapy and the most commonly used integration system for stable gene transfer. The first two adverse events in the French X-SCID gene-therapy trial involved integration of an MLV-derived vector near the 5′ end of the

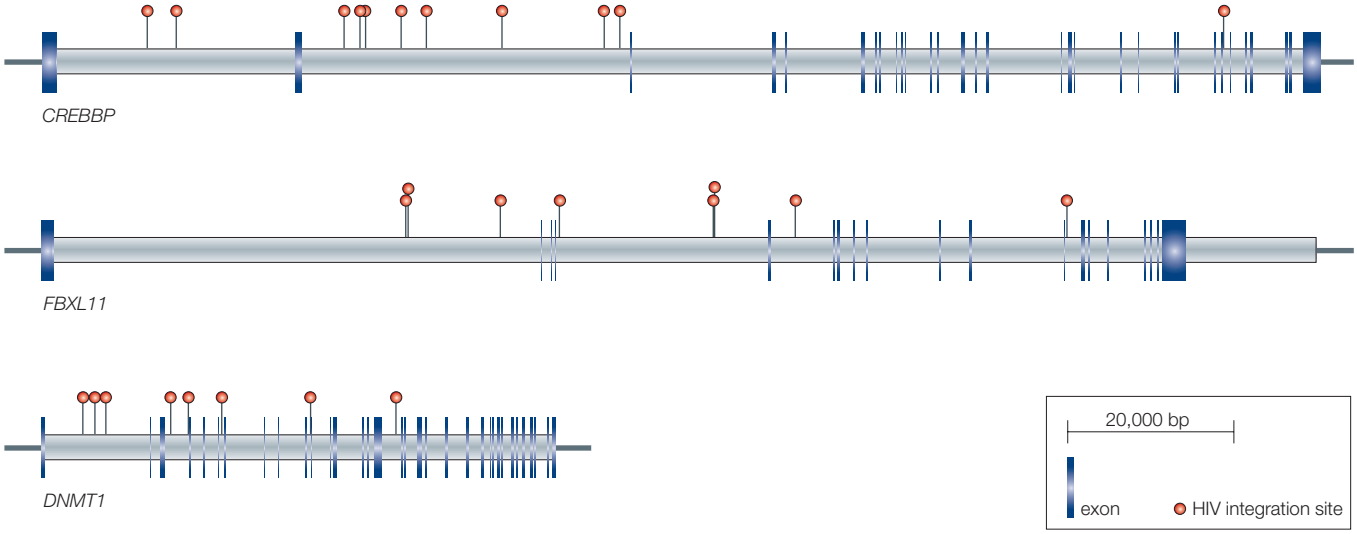


Figure 3 | **Three human genes that have hosted multiple HIV integration events.** The three genes are drawn to the same scale. An analysis of pooled data on HIV integration sites revealed that these genes had hosted several integration events. In this analysis, the gene that encodes CREB-binding protein (*CREBBP*) hosted 11 HIV integration events, the F-box and leucine-rich-repeat protein-11 gene (*FBXL11*) hosted 8 independent integration events, and the DNA (cytosine-5-)methyltransferase-1 gene (*DNMT1*) hosted 8 independent integration events (F.B., unpublished results).

LMO2 gene[13,14,48] — the first integration event was in the 5′ promoter region, the second in the first intron; both were within the window of favoured integration defined by Wu *et al.*[11] This raises the disturbing possibility that MLV might be particularly prone to insertional activation of oncogenes.

ASLV shows a third pattern of preferred integration sites. For this virus, 469 sites in 293T cells[12] and 226 sites in HeLa cells[49] have been sequenced and analysed. ASLV has the most random distribution of integration sites in the human genome. Only a slight preference for integration in transcription units and a weak bias in favour of active genes could be detected[12]. This is in contrast to studies of ASLV integration in two genes in quail cells, in which integration was reported to be disfavoured by high levels of transcription[30,50]. The reason for the different conclusions about the effects of transcription on ALSV integration in these studies is unclear — possibly the cell type studied or the experimental methods used resulted in the different outcomes. Further experimental work will be helpful in resolving this issue. These studies raise the question of whether ASLV-based vectors might be attractive for use in human gene therapy, as their target-site specificity might be the least toxic of the three viruses studied so far.

Genome-wide studies in other vertebrates

So far, only one study has evaluated integration in a non-human system by large-scale sequencing of integration sites. This work focused on integration in the haematopoietic stem cells of rhesus macaques, a cell type that is relevant to gene-therapy applications[51]. Vectors derived from MLV and simian immunodeficiency virus (SIV) were compared. The SIVs are members of the lentivirus family, which also includes HIV. Some complications arose because the rhesus genome is not fully sequenced, but integration sites could be mapped on the similar human genome. The analysis revealed that the SIV vectors favoured integration in transcription units, and that MLV favoured integration near transcription start sites. Therefore, the SIV integration pattern paralleled that seen previously for HIV, which indicates that the determinants of integration targeting might be conserved in the lentivirus group. Although the primate studies examined integration sites years after the initial infection of cells, this did not seem to have a large effect on the global distribution of integration sites. MLV showed the same integration-target preferences in rhesus-macaque and human cells, which indicates cross-species conservation of the cellular determinants of targeting.

Integration-site selection by tethering?

What mechanisms direct retroviral-integration target-site selection in human chromosomes (FIG. 4)? It has been proposed that retroviral integration is favoured in open chromatin, which might be more accessible to the integration apparatus[15]. This notion is supported by genome-wide studies, as integration in transcription units is favoured in all the data sets for retroviruses and other integrating elements[8,10,12,41,52]. By contrast, integration in centromeric heterochromatin, which is identified by alphoid repeats, is disfavoured[7,10]. However, because the integration-target preferences of HIV, MLV and ASLV are so different, it seems unlikely that the accessibility of DNA is the only mechanism that determines target-site selection.

Studies of retrotransposons in yeast provide another candidate mechanism[53–56]. The Ty elements replicate by cycles of transcription, reverse transcription and integration that are similar to retroviruses, and these elements encode reverse transcriptase and integrase

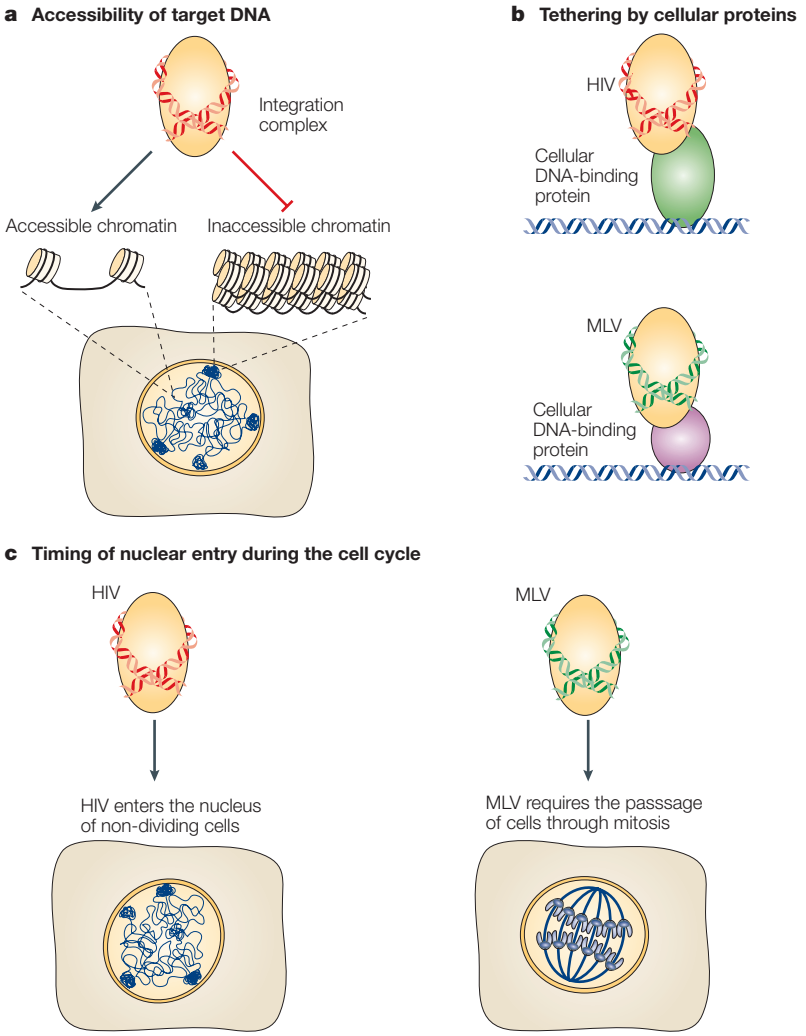


Figure 4 | **Candidate mechanisms that direct integration-site selection by retroviruses.** **a** | Regulation of integration by accessibility. According to this model, chromosomal DNA is relatively inaccessible to integration complexes when packed in nucleosomes and other proteins. Exposure of target sequences promotes integration. **b** | Regulation of integration by tethering. This model proposes that specific interactions between integration complexes and cellular proteins bound locally on target DNA promote integration at nearby sites. **c** | Timing of nuclear entry influences integration targeting. To enter the nucleus, murine leukaemia virus (MLV) requires the passage of cells through mitosis, whereas HIV can enter cells at different stages of the cell cycle. If the state of chromosomes differs at different points in the cell cycle, this could influence integration targeting.

enzymes that are also similar to their retroviral counterparts. The Ty elements differ from retroviruses in that the replication cycle has no extracellular stage, which makes these elements retrotransposons instead of retroviruses. Because Ty elements never leave their host cell, they have a strong motive to preserve the host genome — lethal mutagenesis would be suicidal for both the Ty element and the host. These elements show a strongly biased target-site selection. Ty3, for example, favours integration at the 5′ ends of RNA-polymerase-III-transcribed genes, which does not disrupt gene expression. Ty5 favours integration at telomeres and the silent information loci that are involved in mating-type switching — integration at these targets also does not seem to adversely affect the yeast host cell. In each of these cases, there is evidence that Ty integration complexes are tethered to their favoured sites by interactions with specific cellular proteins, and these interactions mediate local integration. Such a tethering mechanism might also operate for retroviruses[56].

For retroviral integration complexes, any viral or cellular protein in the complex could act as a binding partner in a tethering interaction. Several cellular proteins have been proposed to bind to viral preintegration complexes, including Ku, the high mobility group protein HMGA, the barrier to autointegration factor (BAF) and the transcriptional co-activator LEDGF[57–60] — these proteins might serve as tethering factors for integration targeting. Of particular interest are cellular proteins that have been reported to bind integrase, as these could act as tethering factors similar to the findings in yeast. Two well studied examples are the integrase interactor-1 INI-1 (REF. 61) and LEDGF[62–64]. With available techniques, it should soon be possible to determine whether these proteins are involved in targeting retroviral integration *in vivo*.

Targeting integration* in vitro *with fusion of integrase to sequence-specific DNA-binding domains. Studies of purified retroviral integrase proteins *in vitro* indicate that tethering can also affect integrase targeting in model systems. In the first such study, HIV integrase was fused to the DNA-binding domain of phage λ repressor[28,65–67]. Such a fusion showed preferential integration into target DNA near λ-repressor-binding sites. Further studies have documented that other combinations of integrase proteins and DNA-binding domains can result in targeted integration *in vitro*[68,69]. Although these studies have not yet yielded modified retroviral derivatives capable of targeted integration *in vivo*, they provide proof of principle that tethering of retroviral integrases at target sites can promote nearby integration.

HIV integration and oncogenic transformation

It is well documented that oncoretroviruses such as MLV and ASLV can cause insertional activation of proto-oncogenes in animals, resulting in cancer. HIV accumulates to high levels in patients, so why does it not cause transformation by insertional activation? There are characteristic malignancies associated with AIDS, but the transformed cells do not contain integrated HIV DNA. Two reports have indicated that there might be insertional activation of cellular proto-oncogenes by HIV integration in rare AIDS-associated cancers[70,71], but the data published so far are incomplete, and decisive data have not been forthcoming, leaving the field doubtful about the occurrence of HIV-associated insertional activation even in these rare cases.

For HIV, it seems probable that insertional activation does not take place because the rapid cytopathicity of infection kills potentially transformed cells. The rare long-lived cells that harbour integrated HIV proviruses are usually quiescent (for example, macrophages) and seem to be poor targets for transformation.

However, the non-cytopathic HIV-based vectors used in gene therapy might pose a greater risk of transformation, as gene-therapy experiments are usually designed to maximize the persistence of transduced cells. Analysis of HIV integration sites has shown that integration in proto-oncogenes is readily detected (F.B., unpublished data). Proto-oncogenes are even slightly favoured over other genes as integration targets, possibly because they are more highly expressed than other genes in the cell types studied. However, it is important to note that the consequences of these integration events cannot be deduced from these types of data alone. HIV integration could activate a gene, inactivate a gene or have no effect on its transcription. In the case of tumour-suppressor genes, a second mutation would be needed to inactivate the second allele for an integration event to be carcinogenic. The proportion of HIV vector-integration events in proto-oncogenes and tumour-suppressor genes that affect gene activity is unknown.

Transcription of the integrated proviral genome

The chromosomal location of integration can also affect the efficiency of retroviral gene expression after integration. In particular, the discovery of HIV latency has focused attention on factors that down-modulate HIV gene expression[72–74], including the role of integration-site location. The presently available highly active antiretroviral therapy (HAART) can suppress HIV replication to undetectable levels and can maintain this suppression for years, but unfortunately patients are not cured by this treatment. Long-lived quiescent cells persist, and if drug treatment is stopped, proviruses can emerge from these reservoirs and recommence high-level viral replication. One reservoir is provided by resting T cells, which can be reactivated by stimulation with an appropriate antigen and so produce HIV. These long-lived, latently infected cells are probably the single biggest obstacle to clearing HIV from patients.

Latently infected cells are particularly difficult to study owing to their rarity and to masking by the many inactivated proviruses that accumulate in patients' cells. A study by Siliciano and co-workers analysed 74 integration sites from resting T cells of patients that had been successfully treated with HAART for prolonged periods[74]. They found that most of the sequenced integration sites were in transcription units, similar to the *de novo* integration by HIV in tissue-culture cells. However, a technical issue complicated the interpretation. Most of the HIV proviruses in the T cells of these patients harboured inactivating mutations — indeed, that is probably why the infected cells had survived. Authentic latent (but replication-competent) HIV proviruses comprised only about 1% of HIV sequences in the cell population analysed. In the study by Siliciano and colleagues, it is probable that most of the sequences were from mutationally inactivated proviruses. Thus, the relationship between latency and integration sites unfortunately remains uncertain.

Mechanisms of latency due to the site of HIV integration have thus been studied in cultured cell models, in which populations of inactive but inducible HIV proviruses can be purified and more easily studied. Several studies have taken advantage of a model in which Jurkat cells (a transformed human T-cell line) are infected with HIV-based vectors that transduce green fluorescent protein (GFP)[75,76]. Sorting for GFP-positive cells allows cells that harbour highly expressed and poorly expressed proviruses to be purified by fluorescence-activated cell sorting. On stimulation with tumour-necrosis factor-α (TNFα), some of the 'GFP-dark' cells could be induced to express GFP. In this model, the TNFα treatment mimics the reactivation of latent T cells after contact with antigen that might 'reseed' infection in patients.

In one recent study using the Jurkat model, the sequences of 971 junctions of proviral DNA and flanking cellular DNA were determined, and their distributions were compared in initially bright and inducible populations of proviruses[41]. The relationship between viral transcription and cellular gene activity was assessed by transcriptional profiling of the Jurkat target cells. This study revealed that three types of chromosomal locations were unfavourable for high-level HIV gene expression in these cells. The first, as suggested previously by Verdin and co-workers[76], was centromeric heterochromatin. As discussed above, initial integration in centromeric heterochromatin is disfavoured. However, such events can be detected, and proviruses in these sites are inefficiently transcribed[41,75,76]. The second genomic environment that was overrepresented in the inducible population was highly expressed cellular genes. This indicates that host-cell gene transcription can repress HIV transcription, and it is thought that high-level transcription across the integrated provirus might inhibit HIV gene expression. Several studies have previously indicated that transcriptional interference can inhibit retroviral transcription[77,78], which supports the idea that transcriptional interference is important in this model. The third type of region that was enriched for poorly expressed proviruses was long intergenic regions (known as 'gene deserts'). Why gene deserts are unfavourable environments for HIV gene expression is unclear, although they might be markers for heterochromatic chromosomal regions.

In summary, these data indicate mechanisms by which the chromosomal environment can influence expression of integrated sequences and provide potential explanations for transcriptional latency in cells from HIV-infected patients.

Integration targeting and evolution

Why have retroviruses evolved these integration-targeting strategies? Any after-the-fact evolutionary speculation is risky, but there are striking data from other systems that indicate some of the possibilities[2,56,79]. In prokaryotes, many bacteriophages integrate into bacterial genomes near transfer RNA genes, which seem to be sites where attachment of the prophage does not disrupt cellular transcription. Similarly, in yeast,

the Ty elements favour integration in benign locations. In these cases, and in many not discussed here, the host–parasite interaction seems to have evolved to preserve the host and thereby allow further rounds of parasite replication.

Replication of many retroviruses does not involve a stable and enduring interaction with the host cell, which might explain some of the differences in their targeting strategies. Most cells that are infected by HIV have short life spans, approximately one to two days[80–82]. The HIV pattern of targeting active transcription units might promote efficient expression of the provirus, allowing the maximum production of daughter virions during the limited lifespan of the infected cells.

A recent study by Naldini and co-workers revealed a possible reason for the MLV targeting strategy[83]. These workers compared 'promoter trapping' by MLV and HIV — they placed marker genes that lacked promoters in the genomes of HIV or MLV, and examined how often integration near host-cell promoters allowed activation of the marker gene. They found that MLV was efficient at promoter trapping — approximately 20% of integrated MLV genomes expressed the marker gene. The percentage was lower for HIV. Similar results were found in both human and murine cells, indicating that the integration-target biases might be the same for cells of both species. These findings support the idea that MLV has evolved to take advantage of nearby cellular promoter sequences to activate its own expression.

It is less obvious how the ASLV integration pattern fits with the biology of the virus. Perhaps ASLV has evolved to minimize the damage to the host-cell genome and does so by integrating more randomly than HIV or MLV. Possibly, ASLV integrates with greater specificity in avian cells. However, an initial study of ASLV integration in chicken-embryo fibroblasts revealed that the chromosomal-integration-target preferences were similar to those seen previously in human cells[84].

A more detailed understanding of integration targeting can also help us to understand the forces that shape contemporary genomes. In humans, about 8% of the genome is composed of HERVs that integrated into germ cells in the lineage leading to humans. In mice, the fraction of the genome that is contributed by proviruses is about 8%, and in chickens, 1.3%. The contemporary pattern of integrated elements is the product of two forces — the mechanisms that specify the initial placement of integration sites and the evolutionary constraints that specify which integrated sequences persist in the species. For endogenous retroviruses, in which the initial integration targeting can be studied experimentally, the pressures of subsequent evolution can be discerned by comparison.

Conclusions and perspectives

Recent genome-wide surveys of retroviral DNA integration-site selection have led to the surprising conclusion that different retroviruses have different target-site preferences. HIV favours integration in active genes, MLV favours integration in the 5′ ends of genes, and ASLV does not strongly favour any obvious host-genome feature. Several mechanisms probably contribute to retroviral integration-site selection (FIG. 4). The relative accessibility of target DNA probably plays some part, because active transcription units are favoured for integration for each retrovirus studied, and centromeric heterochromatin is disfavoured for integration. However, the nature of 'open' and 'closed' chromatin has always been poorly defined, and it remains possible that more specific mechanisms explain these observations. Furthermore, the differences in integration targeting of HIV, MLV and ASLV indicate that there might be virus-specific interactions that mediate targeting, perhaps akin to the tethering systems seen for the Ty retrotransposons. Other host-cell mechanisms could also contribute to retroviral integration-site targeting. For example, HIV infects non-dividing cells, whereas cells must pass through mitosis for MLV integration to occur[85,86] — therefore, viral integration complexes might access chromosomal integration targets at different stages of the cell cycle. If the state of the chromosomes varies during the cell cycle in a manner that influences integration targeting, then this difference in cell-cycle entry point could result in an altered distribution of integration sites.

Looking forward, it seems probable that some of the host proteins that affect integration targeting might soon be identified. It should be possible to establish whether host-cell-targeting factors function by tethering or by other mechanisms. A more detailed understanding of these mechanisms might, in turn, allow greater control over integration specificity in retroviral vectors. For example, it might be possible to mutate the integrase proteins of HIV or MLV to reduce binding to tethering factors, and so reduce integration in or near transcription units. Perhaps such integrase derivatives will be more suitable for use in targeting constructs based on fusions of integrase proteins to sequence-specific DNA-binding domains.

The fact that the human genome contains 3.5 billion base pairs complicates efforts to target integration — any attempt to target specific sites must do so in the face of an overwhelming excess of non-specific competitor DNA. It has long been known that sequence-specific DNA-binding proteins accumulate on non-specific DNA when not bound to their specific sites[87]. For a fusion protein that consists of an integrase bound to a sequence-specific DNA-binding domain, time spent bound to non-specific DNA presents an opportunity for integration in non-specific sites. Thus, competition from the vast excess of non-specific DNA presents an important challenge in designing sequence-specific integration systems.

Retroviral vectors based on ASLV might offer some advantages, owing to their potentially less toxic profile of favoured integration sites. In contrast to HIV, ASLV does not strongly favour integration in active transcription units in human cells, nor does it favour integration near the 5′ ends of genes (in contrast to MLV). However, as at least one-third of the human genome is transcribed, 33% of random integration events will

target transcription units. Regarding insertional activation, it is unclear in most cases what size range of chromosomal segments support insertional activation of a particular proto-oncogene. Inspection of the available data indicates that the crucial windows differ among the proto-oncogenes studied[88–90]. It is probable that ASLV vectors will integrate in such windows less often than MLV, but they clearly will hit such sites occasionally. ASLV is known to activate oncogenes; indeed, ASLV was the first group of retroviruses discovered owing to their ability to induce cancers in chickens[91,92]. Therefore, whereas ASLV vectors might result in diminished insertional activation compared to MLV, it would probably be only a quantitative improvement. On the other hand, this illustrates that it is possible to envisage improvements to the safety of retroviral vectors with existing technology, and further studies are likely to open new possibilities.

1. Coffin, J. M., Hughes, S. H. & Varmus, H. E. *Retroviruses* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1997).
The 'bible' of retrovirology.
2. Bushman, F. D. *Lateral DNA Transfer: Mechanisms and Consequences* (Cold Spring Harbor Laboratory Press, New York, 2001).
Describes retroviral replication in the context of the mobile-DNA field.
3. Brown, P. O., Bowerman, B., Varmus, H. E. & Bishop, J. M. Correct integration of retroviral DNA *in vitro*. *Cell* **49**, 347–356 (1987).
The first demonstration of retroviral integration using preintegration complexes from infected cells.
4. Craigie, R., Fujiwara, T. & Bushman, F. The IN protein of Moloney murine leukemia virus processes the viral DNA ends and accomplishes their integration *in vitro*. *Cell* **62**, 829–837 (1990).
The first demonstration that MLV integrase could form covalent bonds between model viral DNA and target DNA.
5. Bor, Y.-C., Miller, M., Bushman, F. & Orgel, L. Target sequence preferences of HIV-1 integration complexes *in vitro*. *Virology* **222**, 238–242 (1996).
6. Stevens, S. W. & Griffith, J. D. Sequence analysis of the human DNA flanking sites of human immunodeficiency virus type 1 integration. *J. Virol.* **70**, 6459–6462 (1996).
7. Carteau, S., Hoffmann, C. & Bushman, F. D. Chromosome structure and HIV-1 cDNA integration: centromeric alphoid repeats are a disfavored target. *J. Virol.* **72**, 4005–4014 (1998).
8. Yant, S. R. *et al.* High-resolution genome-wide mapping of transposon integration in mammals. *Mol. Cell. Biol.* **25**, 2085–2094 (2005).
9. Holman, A. G. & Coffin, J. M. Symmetrical base preferences surrounding HIV-1, avian sarcoma/leukosis virus, and murine leukemia virus integration sites. *Proc. Natl Acad. Sci. USA* **102**, 6103–6107 (2005).
10. Schroder, A. *et al.* HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110**, 521–529 (2002).
The first genome-wide study of HIV integration, which revealed that HIV favours integration in active transcription units.
11. Wu, X., Li, Y., Crise, B. & Burgess, S. M. Transcription start regions in the human genome are favored targets for MLV integration. *Science* **300**, 1749–1751 (2003).
The first genome-wide study of MLV integration, which showed that this virus differed from HIV by favouring integration near transcription start sites.
12. Mitchell, R. *et al.* Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol.* **2**, E234 (2004).
Bioinformatic survey of integration by HIV, MLV and ASLV, and a comparison of integration by HIV in different cell types.
13. Hacein-Bey-Abina, S. *et al.* A serious adverse event after successful gene therapy for X-linked severe combined immunodeficiency. *N. Engl. J. Med.* **348**, 255–256 (2003).
14. Hacein-Bey-Abina, S. *et al.* LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science* **302**, 400–401 (2003).
References 13 and 14 characterize the adverse events that occurred during gene therapy for X-SCID.
15. Panet, A. & Cedar, H. Selective degradation of integrated murine leukemia proviral DNA by deoxyribonucleases. *Cell* **11**, 933–940 (1977).
16. Rohdewohld, H., Weiher, H., Reik, W., Jaenisch, R. & Breindl, M. Retrovirus integration and chromatin structure: Moloney murine leukemia proviral integration sites map near DNase I-hypersensitive sites. *J. Virol.* **61**, 336 (1987).
17. Vijaya, S., Steffan, D. L. & Robinson, H. L. Acceptor sites for retroviral integrations map near DNase I-hypersensitive sites in chromatin. *J. Virol.* **60**, 683–692 (1986).
18. Scherdin, U., Rhodes, K. & Breindl, M. Transcriptionally active genome regions are preferred targets for retrovirus integration. *J. Virol.* **64**, 907 (1990).
19. Shih, C.-C., Stoye, J. P. & Coffin, J. M. Highly preferred targets for retrovirus integration. *Cell* **53**, 531–537 (1988).
20. Withers-Ward, E. S., Kitamura, Y., Barnes, J. P. & Coffin, J. M. Distribution of targets for avian retrovirus DNA integration *in vivo*. *Genes Dev.* **8**, 1473–1487 (1994).
21. Bushman, F. D., Fujiwara, T. & Craigie, R. Retroviral DNA integration directed by HIV integration protein *in vitro*. *Science* **249**, 1555–1558 (1990).
22. Katz, R. A., Merkel, G., Kulkosky, J., Leis, J. & Skalka, A. M. The avian retroviral IN protein is both necessary and sufficient for integrative recombination *in vitro*. *Cell* **63**, 87–95 (1990).
23. Bushman, F. D. & Craigie, R. Activities of human immunodeficiency virus (HIV) integration protein *in vitro*: specific cleavage and integration of HIV DNA. *Proc. Natl Acad. Sci. USA* **88**, 1339–1343 (1991).
24. Pryciak, P. M., Sil, A. & Varmus, H. E. Retroviral integration into minichromosomes *in vitro*. *EMBO J.* **11**, 291–303 (1992).
An early study of integration into chromatin templates.
25. Pruss, D., Reeves, R., Bushman, F. D. & Wolffe, A. P. The influence of DNA and nucleosome structure on integration events directed by HIV integrase. *J. Biol. Chem.* **269**, 25031–25041 (1994).
26. Fitzgerald, M. L. & Grandgenett, D. P. Retroviral integration: *in vitro* host site selection by avian integrase. *J. Virol.* **68**, 4314–4321 (1994).
27. Pryciak, P. M. & Varmus, H. E. Nucleosomes, DNA-binding proteins, and DNA sequence modulate retroviral integration target site selection. *Cell* **69**, 769–780 (1992).
The first study of integration into chromatin templates.
28. Bushman, F. D. Tethering human immunodeficiency virus 1 integrase to a DNA site directs integration to nearby sequences. *Proc. Natl Acad. Sci. USA* **91**, 9233–9237 (1994).
The first demonstration that fusion of retroviral integrase proteins to sequence-specific DNA-binding domains can direct integration into predetermined target DNA sites *in vitro*.
29. Bor, Y.-C., Bushman, F. & Orgel, L. *In vitro* integration of human immunodeficiency virus type 1 cDNA into targets containing protein-induced bends. *Proc. Natl Acad. Sci. USA* **92**, 10334–10338 (1995).
30. Weidhaas, J. B., Angelichio, E. L., Fenner, S. & Coffin, J. M. Relationship between retroviral DNA integration and gene expression. *J. Virol.* **74**, 8382–8389 (2000).
31. Pryciak, P., Muller, H.-P. & Varmus, H. E. Simian virus 40 minichromosomes as targets for retroviral integration *in vivo*. *Proc. Natl Acad. Sci. USA* **89**, 9237–9241 (1992).
32. Pruss, D., Bushman, F. D. & Wolffe, A. P. Human immunodeficiency virus integrase directs integration to sites of severe DNA distortion within the nucleosome core. *Proc. Natl Acad. Sci. USA* **91**, 5913–5917 (1994).
A study of integration in nucleosomes, which revealed that integration at 'kinked' DNA sites is favoured.
33. Muller, H.-P. & Varmus, H. E. DNA bending creates favored sites for retroviral integration: an explanation for preferred insertion sites in nucleosomes. *EMBO J.* **13**, 4704–4714 (1994).
34. Bushman, F. D. & Craigie, R. Integration of human immunodeficiency virus DNA: adduct interference analysis of required DNA sites. *Proc. Natl Acad. Sci. USA* **89**, 3458–3462 (1992).
35. Scottoline, B. P., Chow, S., Ellison, V. & Brown, P. O. Disruption of the terminal base pairs of retroviral DNA during integration. *Genes Dev.* **11**, 371–382 (1997).
36. Lander, E. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
The first draft of the human genome sequence from the public consortium.
37. Venter, J. C. The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
The first draft of the human genome sequence from Celera.
38. Caron, H. *et al.* The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* **291**, 1289–1292 (2001).
39. Versteeg, R. *et al.* The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res.* **13**, 1998–2004 (2003).
40. Mungall, A. J. & *et al.* The DNA sequence and analysis of human chromosome 6. *Nature* **425**, 805–811 (2003).
41. Lewinski, M. *et al.* Genome-wide analysis of chromosomal features repressing HIV transcription. *J. Virol.* **79**, 6610–6619 (2005).
42. van't Wout, A. B. *et al.* Cellular gene expression upon human immunodeficiency virus type 1 infection of $CD4^+$-T-cell lines. *J. Virol.* **77**, 1392–1402 (2003).
43. Corbeil, J. *et al.* Temporal gene regulation during HIV-1 infection of human $CD4^+$ T cells. *Genome Res.* **11**, 1198–204. (2001).
44. Mitchell, R., Chiang, C., Berry, C. & Bushman, F. D. Global effects on cellular transcription following infection with an HIV-based vector. *Mol. Ther.* **8**, 674–687 (2003).
45. Chubb, J. R. & Bickmore, W. A. Considering nuclear compartmentalization in light of nuclear dynamics. *Cell* **112**, 403–406 (2003).
46. Laufs, S. *et al.* Retroviral vector integration occurs in preferred genomic targets in human bone marrow-repopulating cells. *Blood* **101**, 2191–2198 (2003).
47. Laufs, S. *et al.* Insertion of retroviral vectors in NOD/SCID repopulating human peripheral blood progenitor cells occurs preferentially in the vicinity of transcription start regions and in introns. *Mol. Ther.* **10**, 874–881 (2004).
48. Fischer, A., Abina, S. H., Thrasher, A., von Kalle, C. & Cavazzana-Calvo, M. LMO2 and gene therapy for severe combined immunodeficiency. *N. Engl. J. Med.* **350**, 2526–2527 (2004).
49. Narezkina, A. *et al.* Genome-wide analyses of avian sarcoma virus integration sites. *J. Virol.* **78**, 11656–11663 (2004).
50. Maxfield, L. F., Fraize, C. D. & Coffin, J. M. Relationship between retroviral DNA-integration-site selection and host cell transcription. *Proc. Natl Acad. Sci. USA* **102**, 1436–1441 (2005).
51. Hematti, P. *et al.* Distinct genomic integration of MLV and SIV vectors in primate hematopoietic stem and progenitor cells. *PLoS Biol.* **2**, e423 (2004).
52. Nakai, H. *et al.* AAV serotype 2 vectors preferentially integrate into active genes in mice. *Nature Genet.* **34**, 297–302 (2003).
53. Sandmeyer, S. Integration by design. *Proc. Natl Acad. Sci. USA* **100**, 5586–5588 (2003).
54. Zhu, Y., Dai, J., Fuerst, P. G. & Voytas, D. F. Controlling integration specificity of yeast retrotransposon. *Proc. Natl Acad. Sci. USA* **100**, 5891–5895 (2003).
The demonstration of controlling integration target-site specificity *in vivo* by engineering a yeast retrotransposon.
55. Boeke, J. D. & Devine, S. E. Yeast retrotransposons: finding a nice quiet neighborhood. *Cell* **93**, 1087–1089 (1998).
56. Bushman, F. D. Targeting survival: integration site selection by retroviruses and LTR-retrotransposons. *Cell* **115**, 135–138 (2003).

57. Li, L. *et al.* Role of the non-homologous DNA end joining pathway in retroviral infection. *EMBO J.* **20**, 3272–3281 (2001).
58. Farnet, C. & Bushman, F. D. HIV-1 cDNA integration: requirement of HMG I(Y) protein for function of preintegration complexes *in vitro*. *Cell* **88**, 483–492 (1997).
59. Suzuki, Y. & Craigie, R. Regulatory mechanisms by which barrier-to-autointegration factor blocks autointegration and stimulates intermolecular integration of Moloney murine leukemia virus preintegration complexes. *J. Virol.* **76**, 12376–12380 (2002).
60. Llano, M. *et al.* LEDGF/p75 determines cellular trafficking of diverse lentiviral but not murine oncoretroviral integrase proteins and is a component of functional lentiviral preintegration complexes. *J. Virol.* **78**, 9524–9537 (2004).
61. Kalpana, G. V., Marmon, S., Wang, W., Crabtree, G. R. & Goff, S. P. Binding and stimulation of HIV-1 integrase by a human homolog of yeast transcription factor SNF5. *Science* **266**, 2002–2006 (1994).
62. Cherepanov, P. *et al.* HIV-1 integrase forms stable tetramers and associates with LEDGF/p75 protein in human cells. *J. Biol. Chem.* **278**, 372–381 (2003).
63. Maertens, G. *et al.* LEDGF/p75 is essential for nuclear and chromosomal targeting of HIV-1 integrase in human cells. *J. Biol. Chem.* **278**, 33528–33539 (2003).
64. Turlure, F., Devroe, E., Silver, P. A. & Engelman, A. Human cell proteins and human immunodeficiency virus DNA integration. *Front. Biosci.* **9**, 3187–3208 (2004).
65. Bushman, F. D. Targeting retroviral integration. *Science* **267**, 1443–1444 (1995).
66. Bushman, F. & Miller, M. D. Tethering human immunodeficiency virus type 1 preintegration complexes to target DNA promotes integration at nearby sites. *J. Virol.* **71**, 458–464 (1997).
67. Bushman, F. D. Targeting retroviral integration? *Mol. Therapy* **6**, 570–571 (2002).
68. Goulaouic, H. & Chow, S. A. Directed integration of viral DNA mediated by fusion proteins consisting of human immunodeficiency virus type 1 integrase and *Escherichia coli* LexA protein. *J. Virol.* **70**, 37–46 (1996).
69. Katz, R. A., Merkel, G. & Skalka, A. M. Targeting of retroviral integrase by fusion to a heterologous DNA binding domain: *in vitro* activities and incorporation of a fusion protein into viral particles. *Virology* **217**, 178–190 (1996).
70. Shiramizu, B., Herndier, B. G. & McGrath, M. S. Identification of a common clonal human immunodeficiency virus integration site in human immunodeficiency virus-associated lymphomas. *Cancer Res.* **54**, 2069–2072 (1994).
71. Mack, K. D. *et al.* HIV insertions within and proximal to host cell genes are a common finding in tissues containing high levels of HIV DNA and macrophage-associated p24 antigen expression. *J. Acquir. Immune Defic. Syndr.* **33**, 308–320 (2003).
72. Wong, J. K. *et al.* Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. *Science* **278**, 1291–1295 (1997).
73. Finzi, D. *et al.* Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science* **278**, 1295–1300 (1997).
74. Han, Y. *et al.* Resting $CD4^+$ T cells from human immunodeficiency virus type 1 (HIV-1)-infected individuals carry integrated HIV-1 genomes within actively transcribed host genes. *J. Virol.* **78**, 6122–6133 (2004).
75. Jordan, A., Defechereux, P. & Verdin, E. The site of HIV-1 integration in the human genome determines basal transcriptional activity and response to Tat transactivation. *EMBO J.* **20**, 1726–1738 (2001).
76. Jordan, A., Bisgrove, D. & Verdin, E. HIV reproducibly establishes a latent infection after acute infection of T cells *in vitro*. *EMBO J.* **22**, 1868–1877 (2003).
77. Cullen, B. R., Lomedico, P. T. & Ju, G. Transcriptional interference in avian retroviruses —implications for the promoter insertion model of leukaemogenesis. *Nature* **307**, 241–245 (1984).
78. Greger, I. H., Demarchi, F., Giacca, M. & Proudfoot, N. J. Transcriptional interference perturbs the binding of Sp1 to the HIV-1 promoter. *Nucleic Acids Res.* **26**, 1294–1301 (1998).
79. Craig, N. L., Craigie, R., Gellert, M. & Lambowitz, A. M. *Mobile DNA II* (ASM Press, Washington DC, 2002).
80. Ho, D. D. *et al.* Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature* **373**, 123–126 (1995).
81. Wei, X. *et al.* Viral dynamics in human immunodeficiency virus type 1 infection. *Nature* **373**, 117–122 (1995).
82. Coffin, J. M. HIV population dynamics *in vivo*: implications for genetic variation, pathogenesis, and therapy. *Science* **267**, 483–486 (1995).
83. De Palma, M. *et al.* Promoter trapping reveals significant differences in integration site selection between MLV and HIV vectors in primary hematopoietic cells. *Blood* **105**, 2307–2315 (2005).
84. Barr, S., Leipzig, J., Shinn, P., Ecker, J. & Bushman, F. Integration targeting by ASLV and HIV in the chicken genome. *J. Virol.* (in the press).
85. Roe, T., Reynolds, T. C., Yu, G. & Brown, P. O. Integration of murine leukemia virus DNA depends on mitosis. *EMBO J.* **12**, 2099–2108 (1993).
86. Naldini, L. *et al. In vivo* gene delivery and stable transduction of nondividing cells by a lentiviral vector. *Science* **272**, 263–267 (1996).
87. Lin, S.-Y. & Riggs, A. D. The general affinity of lac repressor for *E. coli* DNA: implications for gene regulation in prokaryotes and eukaryotes. *Cell* **4**, 107–111 (1975).
The classic physicochemical study of the role of non-specific DNA binding in the *in vivo* association of the lac repressor with its operator.
88. Lund, A. H. *et al.* Genome-wide retroviral insertional tagging of genes involved in cancer in Cdkn2a-deficient mice. *Nature Genet.* **32**, 160–165 (2002).
89. Suzuki, T. *et al.* New genes involved in cancer identified by retroviral tagging. *Nature Genet.* **32**, 166–174 (2002).
90. Kim, R. *et al.* Genome-based identification of cancer genes by proviral tagging in mouse retrovirus-induced T-cell lymphomas. *J. Virol.* **77**, 2056–62 (2003).
91. Ellerman, V. & Bang, O. Experimentelle Leukamie bei Huhnern. *Zentralbl. Bakteriol. Parasitenkd. Infectionskr. Hyg. Abt. Orig.* **46**, 595–609 (1908).
The first paper on retroviruses.
92. Rous, P. A sarcoma of the fowl transmissible by an agent separable from the tumor cells. *J. Exp. Med.* **13**, 397–411 (1911).
Describes the discovery and characterization of Rous sarcoma virus.

Acknowledgements

This work was supported by grants from the National Institutes of Health.

Competing interests statement

The authors declare no competing financial interests.

Online links

DATABASES

The following terms in this article are linked online to:

Entrez: http://www.ncbi.nlm.nih.gov/Entrez

ASLV | *CREBBP* | *DNMT1* | *FBXL11* | *LMO2* | MLV | SIV

Infectious Disease Information:

http://www.cdc.gov/ncidod/diseases/index.htm

AIDS

FURTHER INFORMATION

Frederic Bushman's laboratory:

http://microb230.med.upenn.edu

The current state of the AIDS epidemic:

http://www.unaids.org.vn

Mouse Retrovirus Tagged Cancer Gene database:

http://rtcgd.ncifcrf.gov

All the virology on the world wide web:

http://www.virology.net/garryfavwebindex.html

Access to this interactive links box is free online.