Methods 47 (2009) 261–268

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Methods for integration site distribution analyses in animal cell genomes

Angela Ciuffi [b], Keshet Ronen [a], Troy Brady [a], Nirav Malani [a], Gary Wang [a],
Charles C. Berry [c], Frederic D. Bushman [a,*]

[a] Department of Microbiology, University of Pennsylvania School of Medicine, 402 Johnson Pavilion, 3610 Hamilton Walk, Philadelphia, PA 19104-6076, USA
[b] Institute of Microbiology, University Hospital Center and University of Lausanne, Bugnon 48, 1011 Lausanne, Switzerland
[c] Department of Family/Preventive Medicine, University of California, San Diego School of Medicine, San Diego, CA 9209, USA

ARTICLE INFO

Article history:
Accepted 29 October 2008
Available online 25 November 2008

Keywords:
Integration sites
HIV
Retroviral DNA
Genomic

ABSTRACT

The question of where retroviral DNA becomes integrated in chromosomes is important for understanding (i) the mechanisms of viral growth, (ii) devising new anti-retroviral therapy, (iii) understanding how genomes evolve, and (iv) developing safer methods for gene therapy. With the completion of genome sequences for many organisms, it has become possible to study integration targeting by cloning and sequencing large numbers of host–virus DNA junctions, then mapping the host DNA segments back onto the genomic sequence. This allows statistical analysis of the distribution of integration sites relative to the myriad types of genomic features that are also being mapped onto the sequence scaffold. Here we present methods for recovering and analyzing integration site sequences.

1. Introduction

Following binding of a retrovirus to a sensitive cell, the viral and cellular membranes fuse and the viral core is released into the cytoplasm. The viral genomic RNA is then reverse transcribed, yielding a double stranded cDNA copy of the viral RNA genome. The complex of viral DNA and proteins (the "preintegration complex" or "PIC") next carries out the covalent attachment of the viral cDNA to host DNA. The integration step completes the formation of a provirus, which contains all the information necessary for the synthesis of the viral RNAs and proteins and formation of new virions (for reviews see [1,2]).

The DNA breaking and joining reactions that mediate retroviral cDNA integration are well understood (reviewed in [1–4]). A linear form of the viral cDNA serves as the immediate precursor of the integrated provirus. Prior to integration, integrase removes in most cases two nucleotides from the 3′ end of each LTR (long terminal repeat), exposing recessed 3′ hydroxyl groups (a minority of viruses use a slight variation of this theme—for example human immunodeficiency virus [HIV] type 2 clips a trinucleotide from its downstream LTR). This terminal cleavage step may serve to remove heterogeneous extra nucleotides occasionally added to the cDNA ends by reverse transcriptase [5,6] and promote the formation of a stable complex [7,8]. Integrase then catalyzes attack by the recessed 3′ hydroxyl groups on phosphodiester bonds on each target DNA strand so as to join each viral DNA 3′ end to protruding 5′ ends in the target. The points of joining on each strand of the target DNA are separated by 4–6 base pairs depending on the retrovirus involved. Unfolding of this integration intermediate yields gaps at each junction between viral and host DNA, and a 5′ two-base flap derived from the viral DNA. Gap repair to connect the remaining DNA strands is probably carried out by host DNA repair systems. The terminal cleavage and first strand transfer reactions can be modeled in vitro with purified integrase [9–15] and the gap repair step modeled with purified DNA repair enzymes [16,17].

The target for HIV DNA integration is the human genome. The roughly 3.4 billion bases of the euchromatic portion of the human genome have now been sequenced [18,19], though the centromere sequences remain inaccessible with available technology. There are believed to be ~30,000 protein coding genes, with the exons comprising ~1.5% of the genome and the transcription units ~33%. However, these numbers remain surprisingly soft: non-protein-coding genes are hard to detect and may be quite abundant, and several reports suggest that low level transcription may take place in what were thought to be intergenic regions [20,21]. It could still turn out that almost all euchromatic human DNA is transcribed at some level, with the recognized transcription units simply transcribed more frequently.

With the completion of the draft human genome sequence in 2001 [18,19], it became possible to study integration target site selection by retroviruses using high-throughput DNA sequencing. In our first genome-wide study of retroviral DNA integration [22], we mapped and analyzed 524 sites of HIV cDNA integration

[*] Corresponding author. Fax: +1 215 573 4856.
E-mail address: bushman@mail.med.upenn.edu (F.D. Bushman).

1046-2023/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ymeth.2008.10.028

in human SupT1 cells. To isolate integration site clones, cells were infected with HIV or an HIV-based vector, and then genomic DNA was harvested 2–3 days after infection. The period of cell growth after infection was minimized to reduce the chances of selection of particular integration sites during cell proliferation. DNA was harvested, cleaved with a restriction enzyme, and linkers were then ligated onto the cleaved DNA ends. Integration site sequences were amplified using primers complementary to the linker and the viral DNA end. PCR products were then cloned into a plasmid, transformed into bacteria, and sequenced in a 96-well plate format. The first survey of HIV integration sites in the SupT1 T-cell line revealed that transcription units were strongly favored as integration targets [22]. Global analysis of cellular transcription using Affymetrix gene expression microarrays indicated that active genes were preferential integration targets, particularly genes that were activated in cells after infection by HIV-1. These data documented unexpectedly strong biases in integration site selection and suggested that HIV had evolved to maximize transcription after integration by targeting active genes.

Next, it was of interest to ask whether similar results would be seen in other human cell types, particularly primary cells. Integration site surveys were thus carried out in many primary cell types and cell lines [22–29]. In all of these data sets, HIV DNA integration was found to be favored in transcription units as well. Integration site selection by retroviruses has been the topic of several recent reviews [4,30–32].

Murine leukemia virus (MLV) integration targeting was initially characterized in a publication from Shawn Burgess's laboratory [33]. They reported the striking discovery that MLV favored integration near gene 5′ ends and CpG islands, and only showed a weak preference for transcription units.

Avian sarcoma-leukosis virus (ASLV) was next characterized and showed the most random distribution of integration sites of the three retroviruses, only weakly favoring transcription units, and not strongly favoring gene 5′ ends or CpG islands [23,34]. Thus, to our surprise, each of the three retroviruses studied showed a unique pattern of integration targeting. Summarizing over many recent studies, integration targeting preferences appear to be consistent within the different retroviral genera but generally different between genera.

A variety of subsequent studies have focused on testing models for possible mechanisms of integration targeting. Three models (which are not mutually exclusive) are as follows:

1.1. Tethering

One simple possibility is that integration complexes bind to cellular proteins that are bound at specific locations on chromosomes. Such tethering interactions are well-documented for the retrovirus-related retrotransposons of yeast [35–37], where binding of the element-encoded integrase proteins to cellular DNA-binding proteins has been shown convincingly to mediate integration targeting. According to this idea, the different favored integration sites for the different retroviruses would be a consequence of binding to different cellular tethering proteins. Additional support for this possibility comes from studies of artificial fusions of integrase proteins to sequence-specific DNA binding domains—in these studies, the fusion integrases have been shown to direct favored integration near DNA recognition sites for the added binding domain [38–44]. In principle, any component of the PIC, viral or cellular, could serve as a docking point for a cellular tethering factor that targets retroviral integration. Among the candidates on the viral side are MA, Vpr, and IN. Cellular factors proposed to be associated with PICs include BAF [45], HMGA1 [46], EED [47], p300 [48], Ini-1 [49,50], and LEDGF/p75 [26,51–59].

Considerable data, presented elsewhere in this volume, indicates that LEDGF/p75 has good credentials as a tethering factor for HIV integration targeting [26,58–60].

1.2. Open chromatin

This idea holds that much of the DNA in human cells is inaccessible due to tight wrapping in chromosomal proteins, so that only those regions that are particularly exposed can serve as integration targets. This is probably the first idea proposed for a mechanism of integration targeting in vivo, based on the apparent proximity of sequenced sites of MLV integration to DNase I hypersensitive sites [61–63]. Considerable data indicates that retroviral integration takes place on nucleosome-bound DNA [64–69]. Contemporary studies confirm that retroviral DNA integration is often favored near DNase I cleavage sites and epigenetic marks positively associated with transcription [29,69,70].

Another type of analysis has also provided tentative support for this idea. HIV integration is significantly less frequent in alphoid repeats than expected by chance [22,71]. Alphoid repeats are mostly found in centromeres and are wrapped in centromeric heterochromatin. This observation suggests that centromeric heterochromatin is disfavored for integration, though the mechanistic basis is not known.

1.3. Cell cycle

Retroviruses differ in the cell cycle timing of integration, which could potentially influence integration targeting. HIV can infect cells arrested at several phases of the cell cycle. MLV, in contrast, requires that cells pass through mitosis to allow integration [72–74]. It is likely that the state of chromatin changes during progression of the cell cycle—thus the point in the cell cycle when integration takes place might influence the selection of integration target sites. Two HIV integration site data sets were generated after infection of growth-arrested cells [27,28]. These showed only slightly different patterns of integration compared to dividing cells, and both showed favored integration in active transcription units. Thus, for the moment, there is not strong data to support differences in targeting due to infection at different cell cycle stages.

Studies of integration targeting are also critical for gene therapy. Gene therapy has successfully restored function in children with inherited immunodeficiencies [75,76]—however, adverse events have also taken place in which the vector containing the therapeutic trans gene integrated near a proto-oncogene 5′ end and activated transcription, thereby contributing to leukemogenesis [77–81]. Thus there is intense interest in integration targeting in the gene therapy field, focusing on devising safer gene therapy [30,81,82].

Given these intriguing questions, many groups have become interested in profiling integration site distributions. Some methods are summarized below.

2. Overview of the method

In a typical experiment, cells are transduced with the integrating vector of interest and incubated for sufficient time to allow integration of the new DNA sequences. For infection of cultured cells with a retroviral vector, two days is typically sufficient time for integration to take place. When longer incubations are allowed, selection for integration sites that promote growth of the host cells may become a factor. In some cases this may be the main question under study, for example in analyzing possible adverse events during gene therapy.

To allow efficient isolation of integration site junction fragments, it is helpful if possible to maximize the number of integrated copies per cell. Of course, for some protocols this will not be feasible, for example in analyzing samples from patients treated by therapeutic gene transfer.

The initial steps of the integration site isolation method are adapted from the Genome Walker Kit (catalog number 1803-1, Clontech, Takara Bio USA, Madison, WI). DNA is extracted from transduced cells, and then cleaved with one or more restriction enzymes (Fig. 1). DNA linkers are then ligated to the cleaved ends. The linkers are specially designed to require PCR amplification to originate within the integrated element, and not at the linker. This is accomplished by attaching a blocking group to one 3′ end within the linker (amino-modifier AmC7-Q), which prevents extension. The 5′ end of the linker contains a single stranded extension. The first amplification primer is identical to the 5′ overhang of the linker, so that the sequence must first be copied by a polymerase primed in the vector. This suppresses linker-to-linker amplification of genomic DNA segments lacking vector sequences.

Considerable care needs to be exercised in the choice of restriction enzyme used for cleaving genomic DNA. For retroviruses or retroviral vectors, the LTR sequence is duplicated. For this reason, primers binding to the LTRs can bind to both copies and allow extension with a polymerase. Special steps are needed to avoid amplification of an "internal fragment" from within the viral vector. This can be achieved by cleaving genomic DNA with a restriction enzyme that does not cleave within the vector, or by using a second enzyme cleavage step after linker ligation to clear out the internal fragment. In addition, restriction cleavage should not be carried out with enzymes that do not yield ends suitable for ligation, nor with enzymes that have the CpG dinucleotide in the recognition sequence (CpG is rare in vertebrate genomes and unevenly distributed).

Two rounds of PCR (nested PCR) are used to recover DNAs containing host–virus DNA junctions. The vector primers need to be designed to lie reasonably close to the host–virus DNA junction, so that sequence reads yield a useful segment of flanking host cell DNA. Amplified vector–host junctions are then cloned and sequenced. Reads are then trimmed to remove plasmid, primer, and viral sequences, aligned with the target genome, and analyzed statistically.

3. Step-by-step protocol

3.1. Objective

A common type of study begins with infecting cells with an HIV-based vector containing a green fluorescent protein (GFP) marker gene, which is pseudotyped with the pantropic envelop glycoprotein G from vesicular stomatitis virus (VSV-G) to provide for a relatively high multiplicity of infection (MOI). The resulting integration sites are then cloned from the infected cells (that is, a segment of host cellular DNA at junctions with HIV) for downstream analyses. To accomplish this, treat viral stocks derived from DNA transfections with RNase-free DNaseI (about 10 U/100 µl viral supernatant) (catalog number 04716728001, Roche Applied Science, Mannheim, Germany) for 1 h at 37 °C (this will decrease plasmid carryover in the subsequent steps), and infect cells with the HIV-derived vector or wild type HIV at high MOI (usually 10; where the MOI was determined on 293T cells, in order to obtain ideally 30–70% of infected cells). Let the cells grow for 48 h to allow integration—this time allows for efficient provirus formation but minimizes selection of cells during growth after integration. Choose a seeding density of cells that will allow you to have a nearly confluent plate after 48 h. Perform fluorescence-activated cell sorting (FACS) with an aliquot of cells to determine the infection rate (proportion of GFP positive cells). Infection rates of 10% or higher are preferred, though it is possible to recover integration sites from samples infected at lower efficiencies.

In the protocol below, besides working up samples of genomic DNA, it is recommended to work up a sample with a model integration site cloned in bacteria, to allow optimization of PCR conditions and genomic DNA cloning steps. In the example below, plasmid SINcPPTeGFP is used for this purpose. We describe a protocol where a cocktail of AvrII, NheI, and SpeI restriction enzymes is used to cleave the genomic DNA.

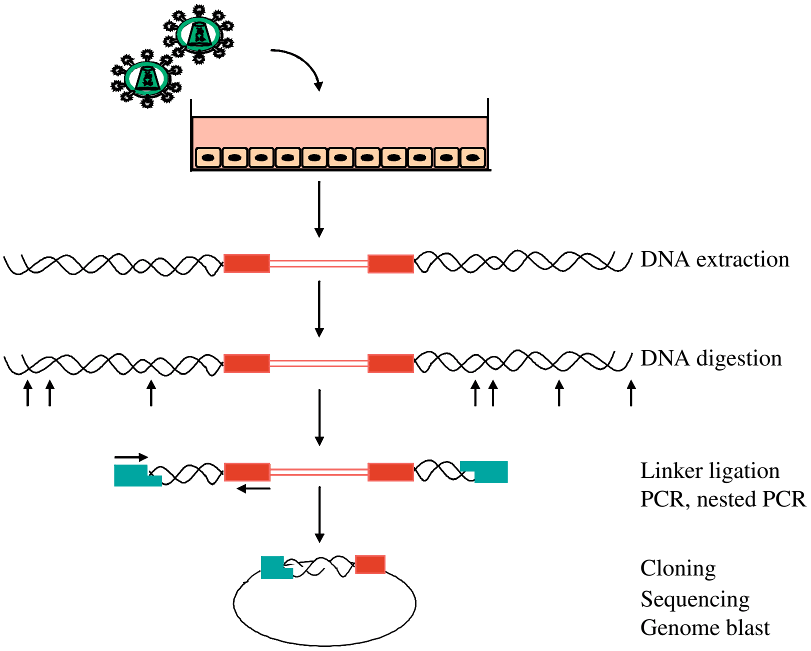

Fig. 1. Diagram of the integration site recovery method. Red boxes, viral LTRs; blue rectangles, linker used during ligation.

3.2. Procedure

3.2.1. Genomic DNA extraction

1. Harvest cells (control or HIV-infected), pellet them, and proceed to genomic DNA extraction, following instructions for the DNeasy tissue kit (catalog number 69506, Qiagen, Valencia, CA). Measure the concentration of the purified DNA by spectrophotometry (OD_{260}).

Check 5 µl on an agarose gel (Fig. 2A).

3.2.2. DNA digestions

2. Genomic DNA digestion (enzymes, reaction buffer, and bovine serum albumin (BSA) from New England Biolabs, Ipswich, MA).

1.5 µg genomic DNA
30 µl NEB2 10× reaction buffer
12 µl AvrII (4 U/µl)
6 µl SpeI (10 U/µl)
6 µl NheI (10 U/µl)
3 µl BSA (100× stock)
→ top up to 300 µl with H_2O
(digest both infected and uninfected DNA)

3. Plasmid DNA digestion

Digest 20 µg of plasmid pSINcPPTeGFP [83,84] as a control/optimization template.

7 µl NEB2 10× reaction buffer
7 µl AvrII (4 U/µl)
→ add H_2O to 70 µl
Digest genomic and plasmid DNAs at 37 °C for 3 h.

Check 5 µl on an agarose gel to ensure proper digestion occurred (Fig. 2B).

3.2.3. DNA purification

Remove proteins via extraction with organic solvents, and recover the digested DNAs by ethanol precipitation.

4. Add 230 µl of H_2O to the control plasmid digestion in order to obtain 300 µl final. Add an equal volume (300 µl) of a 1:1 mixture of phenol/chloroform to both digests. Vortex vigorously and centrifuge for 5 min at 13,000 rpm at room temperature in an Eppendorf centrifuge. Transfer the aqueous phase into a new Eppendorf tube. Add 300 µl of chloroform, vortex vigorously, and repeat centrifugation. Transfer the aqueous phase into a new Eppendorf tube.

Add 1/10th volume (30 µl) of 3 M sodium acetate, pH 5.2, and mix (quick vortex). Add 2.5 volumes (825 µl) of ice cold 100% ethanol. Mix by inversion, and incubate on dry ice for 5 min. Centrifuge at 13,000 rpm for 1 h at 4 °C to pellet DNA. Aspirate the supernatant and invert the Eppendorf tube on a kimwipe. Let air dry for 10–30 min. Resuspend the pellet in 10 µl H_2O, and let sit a moment to help dissolve the DNA.

5. Alternatively, the DNA can be purified using commercially-available spin columns; we prefer the Strataprep PCR purification kit (catalog number 400771; Stratagene, La Jolla, CA). For this, double the genomic DNA digestion to 3 µg DNA in 600 µl final volume. Purify the DNA using the manufacturer's recommendations; elute in

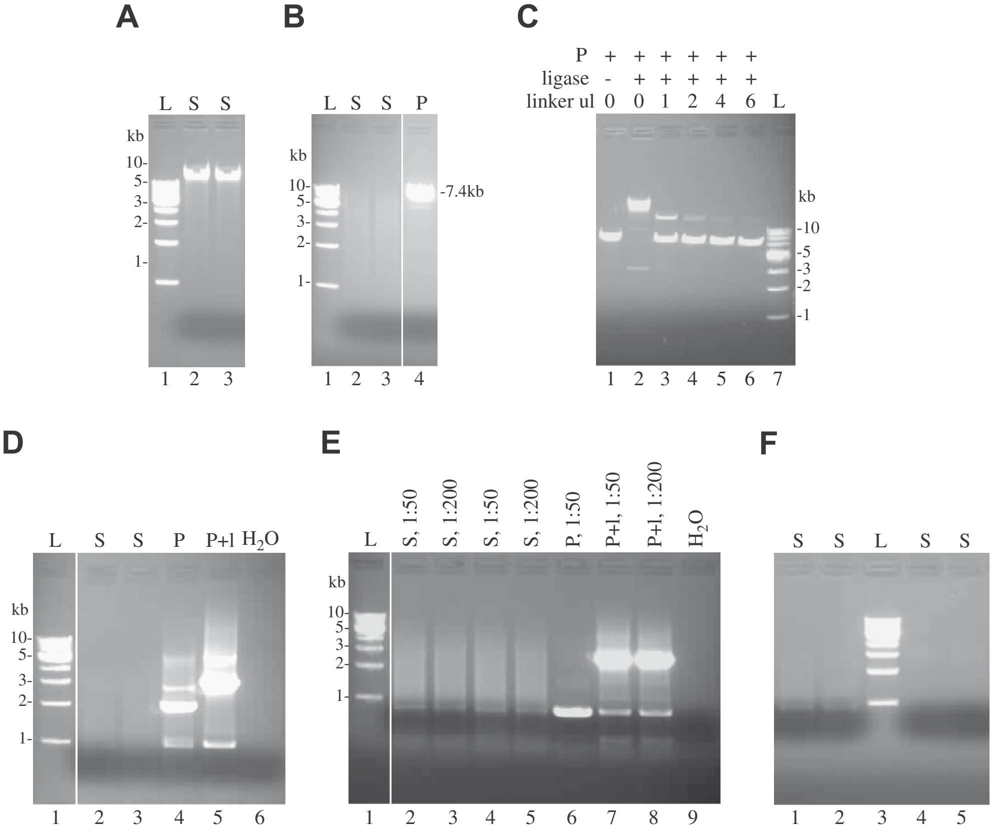

Fig. 2. Analysis of DNA products by native agarose gel electrophoresis and ethidium bromide staining. (A) Verification of the presence of high molecular weight genomic DNA upon extraction of HIV-infected cells. (B) Efficiency of enzymatic digestion with AvrII+NheI+SpeI, with linearization of the plasmid control (at 7.4 kb). (C) Titration of linker (in µl) in the ligation procedure with the plasmid as control, in presence and in absence of the ligase. Lanes 1 and 4 are used as controls for subsequent PCRs. (D) First PCR with one primer annealing in the LTR and one primer annealing in the linker. The plasmid control should give a band at 3.3 kb. (E) Nested PCR with 1:50 and 1:200 dilutions of the first PCR. (F) Purification of the samples to be subsequently cloned into TOPO TA vector (lanes 1 and 2) or TOPO XL vector (lanes 4 and 5). L, 1 kb DNA ladder; S, sample; P, plasmid SINcPPTeGFP as control; l, linker.

50 μl 10 mM Tris pH 8.0. Split the plasmid control digestion onto two columns, and elute each column with 50 μl 10 mM Tris pH 8.0.

3.2.4. Linker preparation

Prepare the linker by annealing two adapter oligonucleotides.

6. The mNAS and HincII adapters are added to the final concentrations of 10 μM each to buffer containing 10 mM Tris pH 8.0, 0.1 mM EDTA. Incubate for 5 min at 90 °C in a PCR machine, then cool 1 °C every 3 min until the temperature reaches ⩽20 °C. Alternatively, switch off the PCR machine, and let the tubes cool down to room temperature over 2–4 h. Store the annealed oligonucleotides at 4 °C until use (but 1 week maximum).

mNAS adapter	5′-[Phosp]CTAGGCAGCCCG[AmC7-Q]
HincII adapter	5′-GTAATACGACTCACTATAGGGCACGCGTG GTCGACGGCCCGGGCTGC

Phosp, 5′ phosphate modification; AmC7-Q, 3′ amino-modifier; underlined, HincII site.

3.2.5. Linker ligation

Ligase and reaction buffer are from New England Biolabs.

7. For genomic DNA, mix:

2 μl 10× ligase buffer
1 μl T4 DNA ligase (400 U/μl)
7 μl adapter mix
10 μl digested genomic DNA in H_2O
(20 μl total)

If the DNA was purified using a spin column, modify this step as follows:

4 μl 10× ligase buffer
2 μl T4 DNA ligase (400 U/μl)
1 μl H_2O
7 μl adapter mix
26 μl digested genomic DNA
(40 μl total)

8. For plasmid DNA, prepare:

2 μl 10× ligase buffer
1 μl T4 DNA ligase (400 U/μl)
0 or 2 μl adapter mix
1 μl control plasmid linearized with AvrII (2 μg)
→ top up to 20 μl with H_2O

For plasmid digests purified by spin column:

2 μl 10× ligase buffer
1 μl T4 DNA ligase (400 U/μl)
0 or 2 μl adapter mix
10 μl control plasmid linearized with AvrII (2 μg)
→ top up to 20 μl with H_2O

Incubate the ligation mixtures overnight at room temperature. Add 80 μl of low TE buffer (10 mM Tris pH 8.0–0.1 mM EDTA) to the genomic DNA mixture (60 μl if using the alternate purification) to attain a final DNA concentration of 10–15 ng/μl, and 47 μl low TE buffer for the control plasmid ligation (30 ng/μl final DNA concentration). Check 10 μl on an agarose gel to ensure proper ligation occurred (Fig. 2C).

3.2.6. Nested PCR amplification of linker ligation products

Use two PCR rounds to amplify host–HIV sequences from the ligation reaction mixtures.

9. The first round, PCRI, utilizes HIV LTR-specific primer SB-76 (5′-GAGGGATCTCTAGTTACCAGAGTCACA) and linker-specific primer ASB-9 (5′-GACTCACTATAGGGCACGCGT); use the Advantage 2 PCR Polymerase Mix from Clontech (catalog number 639201). In separate reactions, analyze DNA prepared from: control (mock-infected) cells, HIV-infected cells, and control plasmid digest (processed with versus without the linker in the previous ligation step). It is also important to include a no-DNA (H_2O-only) control reaction. Mix:

0.5 μl primer ASB-9 (15 μM) → 300 nM final
0.5 μl primer SB-76 (15 μM) → 300 nM final
2.5 μl 10× 2PCR buffer → 1× final
0.5 μl 50× dNTP stock (10 mM each) → 0.2 mM each final
0.5 μl Advantage 2PCR Polymerase mix
5 μl DNA sample (or H_2O)
→ top up to 25 μl with H_2O

PCR cycling conditions :

1× 94 °C for 2 s
7× 94 °C, 2 s; 72 °C, 3 min
37× 94 °C, 2 s; 67 °C, 3 min
1× 67 °C, 4 min
Hold 4 °C
Check 5 μl on an agarose gel (Fig. 2D).

10. PCRI samples are diluted at the following levels for analysis in the nested PCRII step: mock-infected cells, 1:50 (2 μl PCRI product + 98 μl H_2O) and 1:200 (2 μl PCRI product + 398 μl H_2O); HIV-infected cells, 1:50 and 1:200; digested control plasmid (processed without the linker), 1:50; digested control plasmid (processed with the linker), 1:50 and 1:200. Primer ASB-1 (5′-AGCCAGAGAGCTCCC AGGCTCAGATC) is specific for the HIV LTR, whereas ASB-16 (5′-GT CGACGGCCCGGGCTGCCTA) anneals to the linker. Mix:

0.5 μl primer ASB-1 (15 μM) → 300 nM final
0.5 μl primer ASB-16 (15 μM) → 300 nM final
2.5 μl 10× 2PCR buffer → 1× final
0.5 μl 50× dNTP stock (10 mM each) → 0.2 mM each final
0.5 μl Advantage 2PCR Polymerase mix
1 μl DNA sample (dilution of PCRI product, or H_2O)
→ top up to 25 μl with H_2O

PCR cycling conditions:

1× 94 °C for 2 s
7× 94 °C, 2 s; 72 °C, 3 min
37× 94 °C, 2 s; 67 °C, 3 min
1× 67 °C, 20 min
Hold 4 °C
Check 5 μl on an agarose gel (Fig. 2D).

3.2.7. Gel purification of PCR products

PCRII products are purified following agarose gel electrophoresis using the S.N.A.P. UV-Free Gel Purification Kit (Invitrogen catalog number K2000).

11. Prepare a 0.8% agarose gel in TAE (40 mM Tris base, 20 mM acetate, 1 mM EDTA) buffer (50 ml minigel). Add 40 μl Crystal Violet Solution (2 mg/ml). Add 4 μl 6× Crystal Violet loading dye to 20 μl of PCRII product. Analyze a mixture of 1 μl 6× Crystal Violet loading

dye, 2 μl 6× gel loading dye (Promega catalog number G190A, Madison, WI), and 3 μl H_2O as a no-DNA marker to allow the visualization of bromophenol blue and xylene cyanol dye migration positions. Load the gel (leave one or two empty lanes between the "marker" and the sample). Use different gels for samples originating from different cell lines to minimize possible cross-contamination between samples. Run about 10 min at 90 V; the crystal violet will migrate towards the negative pole. Isolate a gel piece extending from the bromophenol blue dye to the loading well of the gel, and cut it in two halves: The upper half, containing relatively large 3–6 kb fragments, will be used for cloning with the TOPO XL cloning kit, and the lower half, containing smaller 0.5–3 kb fragments, will be used for cloning with the TOPO TA cloning kit. Weigh the gel pieces, and purify the DNA using the S.N.A.P. kit according to the manufacturer's instructions. Elute the DNA in 40 μl low TE buffer. Check 10 μl of the purified products on a 1% agarose gel (Fig. 2F).

3.2.8. TOPO cloning and bacterial transformation

We use the TOPO TA cloning kit for sequencing (catalog number K4575) and TOPO XL PCR Cloning kit with OneShot TOP10 electrocompetent *Escherichia coli* (catalog number K4700) from Invitrogen.

12. XL cloning. Mix 4 μl of the purified PCRII XL (3–6 kb DNA fragments) product with 1 μl TOPO XL vector DNA. Incubate at room temperature for 5 min. Add 1 μl of 6× TOPO Stop solution. Quick spin, and place the tube on ice.

13. TA cloning. Mix 4 μl of the purified PCRII TA (0.5–3 kb DNA fragments) product, 1 μl of a 1:4 dilution of Salt Solution, and 1 μl TOPO TA vector. Incubate for 30 min at room temperature. Quick spin, and place the tube on ice.

14. TOP10 electrocompetent cell transformation. Pipet 2 μl of each TOPO reaction into one 50 μl vial of TOP10 bacteria. Do not mix by pipetting, but flick the tube gently. Place the tube back on ice for 5 min. Transfer the mixture into prechilled 2 mm electroporation cuvettes, and place back on ice. Tap the cuvette on the bench to ensure the bacteria are on the bottom of the cuvette, and to eliminate bubbles. Dry the cuvette walls with a kimwipe (this will prevent arcing problems). Electroporate at 2.5 kV (preset bacterial program number 2 on BioRad GenePulser System). Very rapidly add 450 μl (for XL) or 250 μl (for TA) SOC medium, mix by pipetting up and down, and transfer into a polypropylene round bottom tube (catalog number 2059 Falcon, Becton–Dickinson, NJ). Incubate the bacteria at 37 °C for 1 h at 225 rpm. Distribute the XL reaction onto four LBagar-50 μg/ml kanamycin plates: two plates each with 50 μl, and the other two with 200 μl. Also distribute the TA reaction onto 4 LBagar-50 μg/ml kanamycin plates: two plates each with 30 μl, and the remaining two with 120 μl. Incubate the plates overnight (~17 h) at 37 °C.

3.2.9. Colony picking and sequencing

15. Transfer colonies to individual wells of a 96-well plate containing 100 μl LB-50 μg/ml kanamycin, in duplicate (the second plate is prepared as a backup). Incubate the plates overnight at 37 °C. To the backup plate, add 34 μl LB-60% glycerol to each well, and freeze the plate at −80 °C. Submit the other plate to a core facility or company for plasmid extraction and sequencing using M13-reverse (5′-CAGGAAACAGCTATGAC) and M13-forward (5′-TGTAAAACGACGGCCAGT) primers.

4. Processing sequences

Integration site sequences are received from a sequencing center and housed in a MySql database. Sequences are initially trimmed to remove the viral LTR sequence and linker sequence if present. Each sequence is queried against the host cell genome using a local BLAT server. Integration site coordinates (chromosome, base position, and orientation) are calculated based on the best alignment hit for each sequence and entered into the database. Local annotation is then downloaded and added to the integration site record.

The collection of sites is sorted to recover high quality matches. Typically, sequences are required to have (1) a perfect match to the vector terminus, (2) a 98% match to genomic DNA, (3) a unique best hit to the host genome, and (4) a match to host cell DNA beginning within 3 bp of the end of the vector DNA. Trimmed sequences passing these quality controls are stored in separate databases.

A fraction of sequences are found to have high quality matches to multiple sequences in the genome under study. Such multiple hits can be ignored if small in number. In the statistical analysis, the multiple hits can be counted as $1/n$ of an integration site at each of the n equally high scoring genomic locations.

5. Comments on the statistical analysis

For many statistical comparisons of integration sites, it is natural to compare integration frequency near genomic features to the expectation from random integration. However, there is a danger in using random sites—the experimental integration sites were isolated after cleavage of genomic DNA with restriction enzymes, which was recently shown to introduce a recovery bias [85].

To account for restriction site bias in recovery, the random controls are "matched" to have the same potential bias. Random sites are matched to each experimental integration site so that they are randomly distributed, but constrained to lie the same distance from a restriction enzyme recognition site as the experimental site [23,70]. Comparison of experimental sites to the "matched random controls (MRCs)" washes out any potential restriction bias.

Note that restriction bias can have another consequence, which is that specific sites may be difficult to isolate if the distribution of nearby restriction cleavage sites is unfavorable. This is a major issue in essentially all the gene therapy literature on integration sites. However, this is generally not critical in surveys aimed at characterizing genomic features positively or negatively associated with integration frequency.

A question that often arises is "How many integration sites do I need to sequence?" To call a trend in the data, one needs to have a significant difference between the experimental integration sites and the matched random controls. That is, one needs a p value <0.05, which corresponds to making an erroneous claim for an effect when there is no effect (Type I error) less than one time in 20.

Use of p-values, however conflates the strength of the effect with the sample size. To obtain a significant difference with a small sample, a quite strong difference is required. To obtain significance with a large sample, quite modest differences may often be sufficient.

For example, imagine that you are characterizing integration frequency in transcription units for a newly discovered retrovirus. For the retrovirus, 40% of integration sites are in transcription units, while for the matched random control, only 30% are in transcription units. Whether or not this achieves significance depends on the sample size. If you have 100 integration sites and 100 matched random controls, analysis by Fisher's exact test (two tailed) yields $p = 0.1819$, a difference that does not achieve significance. However, if you sequence 1000 integration sites, and prepare 1000 matched random controls, then Fisher's exact test yields $p < 0.0001$ for the same difference in proportions. In practice, if possible, it can be helpful to carry out a pilot experiment to obtain an initial indication of the strength of trend of interest, and then scale the full study appropriately.

The section below provides an overview of some of the more advanced aspects of statistical analysis of integration sites, written for readers with statistical training. A comprehensive description can be found in [70].

The expected number of integration events, Y, given genomic location, L, (i.e. chromosome, position, strand) is represented by a log-linear model

$$E(Y|L=l) \propto \exp\left(\sum_i x_{il}\beta_i\right)$$

where x_{il} is the value of genomic feature 'i' at location 'l' and β_i is an unknown constant governing the effect of that feature. Features include 0/1 indicators for whether 'l' is in a gene, exon, or other kind of region of interest, quantitative measures such as position weight matrix scores for the local DNA sequence or counts of genes in neighborhoods of 'l', and terms for interactions among such features. Model fitting and inference is based on the conditional logit model as implemented by the clogit function of the R survival library [86] using a 'nested case-control' approach (as reviewed by [87]) when control genomic locations are matched to integration sites based on distances to restriction sites for the enzyme(s) used to recover the integration event. When random controls are employed, logistic regression as implemented in the glm R function [88] is used. Logistic regression is also used when two different types of integration events (from integration complexes 'a' and 'b', say) are compared in a common framework; in terms of the earlier model for expected integration events, the coefficients in the logistic regression, $logit(\Pr(IC=a|x_1,x_2,\cdots)) = \alpha + \sum_i x_{ij}\theta_i$, turn out to be $\theta_i = \beta_{ia} - \beta_{ib}$, allowing their interpretation as the difference in effect of x_i on integration of complex 'a' versus 'b'.

In screening quantitative variables, transformation of regressors to a uniform distribution on $[-1,1]$ diminishes the leverage of extreme data points due to highly skewed variables, which makes for easier comparisons of strength of effect of different variables. Also, interpretation of the fitted coefficient is convenient; it gives the effect of increasing the variable by 50 percentile points. Parametric spline functions accommodate non-linear effects; biological considerations (e.g. where transcription starts) sometimes dictates placement of knots.

Some features like gene density and GC density must be tallied in a window (or windows) of specified width(s). The best choice for this (these) window(s) is unknown *a priori*. We approach this problem by fitting a model that includes basis vectors for a variety of choices and use shrinkage methods to avoid overfitting. This method was especially helpful in revealing that HIV integration sites prefer 50 kb wide GC poor 'valleys' surrounded by GC rich regions several megabases in width in the Wang et al., 2007 study [69].

Generally, control of overfitting is achieved by use of penalized likelihood methods [89] or by Bayesian model averaging as implemented in bic.surv [90]. K-fold crossvalidation [91] provides honest estimates of prediction error and is also used for tuning ridge parameters of penalized likelihoods.

The method used to control Type I error due to multiple testing depends on the particular hypothesis/hypotheses and setup. They include Holm's [92] p-value adjustment when closely related hypotheses are studied (e.g. a dozen different measures of gene density are screened to see if gene density affects integration), omnibus general linear hypothesis tests to compare the profiles of effects of two different integration complexes (which implicitly control Type I error rates by constructing a single overarching test), and estimation of false discovery rates [93] when screening many candidates for useful leads. (It often happens that statistical significance is a secondary concern, because very high levels of significance may be achieved for many features, whose influence ranges from biologically trivial to important.)

These model based methods are supplemented by machine learning techniques [94], especially the randomForest algorithm, which has great flexibility in terms of detecting effects due to complicated combinations of features, resists being driven by a few observations of high leverage, measures the importance of each feature, and provides honest estimates of prediction error automatically.

6. Concluding remarks

The study of integration target site selection has taken off in recent years with the availability of complete gene sequences and new high-throughput methods. In all likelihood ongoing methods development will provide a wealth of new research opportunities, including, for example, the new deep sequencing techniques [95,96].

Acknowledgments

We are grateful to members of the Bushman laboratory for help and suggestions. This work was supported by NIH Grants AI52845 and AI66290 to F.D.B., the University of Pennsylvania Center for AIDS Research, and the Penn Genomic Frontiers Institute. A.C. was supported in part by a fellowship from the Swiss National Science Foundation.

References

[1] J.M. Coffin, S.H. Hughes, H.E. Varmus, Retroviruses, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1997.
[2] F.D. Bushman, Lateral DNA Transfer: Mechanisms and Consequences, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2001.
[3] E. Asante-Appiah, A.M. Skalka, Adv. Virus Res. 52 (1999) 351–369.
[4] D.P. Grandgenett, Proc. Natl. Acad. Sci. USA 102 (2005) 5903–5904.
[5] M.D. Miller, C.M. Farnet, F.D. Bushman, J. Virol. 71 (1997) 5382–5390.
[6] P.H. Patel, B.D. Preston, Proc. Natl. Acad. Sci. USA 91 (1994) 549–553.
[7] V. Ellison, P.O. Brown, Proc. Natl. Acad. Sci. USA 91 (1994) 7316–7320.
[8] M. Li, M. Mizuuchi, T.R. Burke Jr., R. Craigie, EMBO J. 25 (2006) 1295–1304.
[9] F.D. Bushman, T. Fujiwara, R. Craigie, Science 249 (1990) 1555–1558.
[10] F.D. Bushman, R. Craigie, J. Virol. 64 (1990) 5645–5648.
[11] R. Craigie, T. Fujiwara, F. Bushman, Cell 62 (1990) 829–837.
[12] P.A. Sherman, J.A. Fyfe, Proc. Natl. Acad. Sci. USA 87 (1990) 5119–5123.
[13] F.D. Bushman, R. Craigie, Proc. Nat. Acad. Sci. USA 88 (1991) 1339–1343.
[14] R.A. Katz, G. Merkel, J. Kulkosky, J. Leis, A.M. Skalka, Cell 63 (1990) 87–95.
[15] M. Katzman, R.A. Katz, A.M. Skalka, J. Leis, J. Virol. 63 (1989) 5319–5327.
[16] K. Yoder, F.D. Bushman, J. Virol. 74 (2000) 11191–11200.
[17] E. Brin, J. Yi, A.M. Skalka, J. Leis, J. Biol. Chem. 275 (2000) 39287–39295.
[18] E. Lander, Nature 409 (2001) 860–921.
[19] J.C. Venter, Science 291 (2001) 1304–1351.
[20] S. Katayama, Science 309 (2005) 1564–1566.
[21] P. Carninci, Nat. Genet. 38 (2006) 626–635.
[22] A.R. Schroder, P. Shinn, H. Chen, C. Berry, J.R. Ecker, F. Bushman, Cell 110 (2002) 521–529.
[23] R.S. Mitchell, B.F. Beitzel, A.R. Schroder, P. Shinn, H. Chen, C.C. Berry, J.R. Ecker, F.D. Bushman, PLoS Biol. 2 (2004) E234.
[24] M. Lewinski, D. Bisgrove, P. Shinn, H. Chen, E. Verdin, C.C. Berry, J.R. Ecker, F.D. Bushman, J. Virol. 79 (2005) 6610–6619.
[25] S.D. Barr, J. Leipzig, P. Shinn, J.R. Ecker, F.D. Bushman, J. Virol. 79 (2005) 12035–12044.
[26] A. Ciuffi, M. Llano, E. Poeschla, C. Hoffmann, J. Leipzig, P. Shinn, J.R. Ecker, F. Bushman, Nat. Med. 11 (2005) 1287–1289.
[27] A. Ciuffi, R.S. Mitchell, C. Hoffmann, J. Leipzig, P. Shinn, J.R. Ecker, F.D. Bushman, Mol. Ther. 13 (2006) 366–373.
[28] S.D. Barr, A. Ciuffi, J. Leipzig, P. Shinn, J.R. Ecker, F.D. Bushman, Mol. Ther. 14 (2006) 218–225.
[29] M.K. Lewinski, M. Yamashita, M. Emerman, A. Ciuffi, H. Marshall, G. Crawford, F. Collins, P. Shinn, J. Leipzig, S. Hannenhalli, C.C. Berry, J.R. Ecker, F.D. Bushman, PLoS Pathog. 2 (2006) e60.
[30] A. Engelman, Proc. Natl. Acad. Sci. USA 102 (2005) 1275–1276.
[31] F. Bushman, M. Lewinski, A. Ciuffi, S. Barr, J. Leipzig, S. Hannenhalli, C. Hoffmann, Nat. Rev. Microbiol. 3 (2005) 848–858.
[32] A. Ciuffi, F.D. Bushman, Retroviral DNA integration: HIV and the role of LEDGF/p75. Trends Genet. 22 (2006) 388–395.
[33] X. Wu, Y. Li, B. Crise, S.M. Burgess, Science 300 (2003) 1749–1751.
[34] A. Narezkina, K.D. Taganov, S. Litwin, R. Stoyanova, J. Hayashi, C. Seeger, A.M. Skalka, R.A. Katz, J. Virol. 78 (2004) 11656–11663.
[35] S. Sandmeyer, Proc. Natl. Acad. Sci. USA 100 (2003) 5586–5588.

[36] Y. Zhu, J. Dai, P.G. Fuerst, D.F. Voytas, Proc. Natl. Acad. Sci. USA 100 (2003) 5891–5895.
[37] J.D. Boeke, S.E. Devine, Cell 93 (1998) 1087–1089.
[38] F.D. Bushman, Science 267 (1995) 1443–1444.
[39] H. Goulaouic, S.A. Chow, J. Virol. 70 (1996) 37–46.
[40] R.A. Katz, G. Merkel, A.M. Skalka, Virology 217 (1996) 178–190.
[41] F. Bushman, M.D. Miller, J. Virol. 71 (1997) 458–464.
[42] F.D. Bushman, Mol. Ther. 6 (2002) 570–571.
[43] W. Tan, Z. Dong, T.A. Wilkinson, C.F. Barbas, S.A. Chow, J. Virol. 80 (2006) 1939–1948.
[44] F.D. Bushman, Proc. Natl. Acad. Sci. USA 91 (1994) 9233–9237.
[45] M.S. Lee, R. Craigie, Proc. Natl. Acad. Sci. USA 91 (1994) 9823–9827.
[46] C. Farnet, F.D. Bushman, Cell 88 (1997) 1–20.
[47] S. Violot, S.S. Hong, D. Rakotobe, C. Petit, B. Gay, K. Moreau, G. Billaud, S. Priet, J. Sire, O. Schwartz, J.F. Mouscadet, P. Boulanger, J. Virol. 77 (2003) 12507–12522.
[48] A. Cereseto, L. Manganaro, M.I. Gutierrez, M. Terreni, A. Fittipaldi, M. Lusic, A. Marcello, M. Giacca, EMBO J. 24 (2005) 3070–3081.
[49] G.V. Kalpana, S. Marmon, W. Wang, G.R. Crabtree, S.P. Goff, Science 266 (1994) 2002–2006.
[50] E. Yung, M. Sorin, A. Pal, E. Craig, A. Morozov, O. Delattre, J.C. Kappes, D. Ott, G. Kalpana, Nat. Med. 7 (2001) 920–926.
[51] P. Cherepanov, G. Maertens, P. Proost, B. Devreese, J. Van Beeumen, Y. Engelborghs, E. De Clercq, Z. Debyser, J. Biol. Chem. 278 (2003) 372–381.
[52] F. Turlure, E. Devroe, P.A. Silver, A. Engelman, Front. Biosci. 9 (2004) 3187–3208.
[53] M. Llano, M. Vanegas, O. Fregoso, D. Saenz, S. Chung, M. Peretz, E.M. Poeschla, J. Virol. 78 (2004) 9524–9537.
[54] M. Llano, S. Delgado, M. Vanegas, E.M. Poeschla, J. Biol. Chem. 279 (2004) 55570–55577.
[55] M. Vanegas, M. Llano, S. Delgado, D. Thompson, M. Peretz, E.M. Poeschla, J. Cell Sci. 118 (2005) 1733–1743.
[56] M. Llano, M. Vanegas, N. Hutchins, D. Thompson, S. Delgado, E.M. Poeschla, J. Mol. Biol. 360 (2006) 760–773.
[57] M. Llano, D.T. Saenz, A. Meehan, P. Wongthida, M. Peretz, W.H. Walker, W. Teo, E.M. Poeschla, Science 314 (2006) 461–464.
[58] H. Marshall, K. Ronen, C. Berry, M. Llano, H. Sutherland, D. Saenz, W. Bickmore, E. Poeschla, F. Bushman, PLoS One 2 (2007) e1340.
[59] A. Ciuffi, T. Diamond, Y. Hwang, H. Marshall, F.D. Bushman, Hum. Gene Ther. 17 (2006) 960–967.
[60] M.C. Shun, N.K. Raghavendra, N. Vandegraaff, J.E. Daigle, S. Hughes, P. Kellam, P. Cherepanov, A. Engelman, Genes Dev. 21 (2007) 1767–1778.
[61] A. Panet, H. Cedar, Cell 11 (1977) 933–940.
[62] H. Rohdewohld, H. Weiher, W. Reik, R. Jaenisch, M. Breindl, J. Virol. 61 (1987) 336.
[63] S. Vijaya, D.L. Steffan, H.L. Robinson, J. Virol. 60 (1986) 683–692.
[64] P.M. Pryciak, H.E. Varmus, Cell 69 (1992) 769–780.
[65] P.M. Pryciak, A. Sil, H.E. Varmus, EMBO J. 11 (1992) 291–303.
[66] P. Pryciak, H. Muller, H.E. Varmus, Proc. Natl. Acad. Sci. USA 89 (1992) 9237–9241.
[67] D. Pruss, R. Reeves, F.D. Bushman, A.P. Wolffe, J. Biol. Chem. 269 (1994) 25031–25041.
[68] D. Pruss, F.D. Bushman, A.P. Wolffe, Proc. Natl. Acad. Sci. USA 91 (1994) 5913–5917.
[69] G.P. Wang, A. Ciuffi, J. Leipzig, C.C. Berry, F.D. Bushman, Genome Res. 17 (2007) 1186–1194.
[70] C. Berry, S. Hannenhalli, J. Leipzig, F.D. Bushman, PLoS Comput. Biol. 2 (2006) e157.
[71] S. Carteau, C. Hoffmann, F.D. Bushman, J. Virol. 72 (1998) 4005–4014.
[72] T. Roe, T.C. Reynolds, G. Yu, P.O. Brown, EMBO J. 12 (1993) 2099–2108.
[73] M. Yamashita, M. Emerman, J. Virol. 78 (2004) 5670–5678.
[74] M. Yamashita, M. Emerman, PLoS Pathog. 1 (2005) e18.
[75] M. Cavazzana-Calvo, S. Hacein-Bey, G. de Saint Basile, F. Gross, E. Yvon, P. Nusbaum, F. Selz, C. Hue, S. Certain, J.L. Casanova, P. Bousso, F.L. Deist, A. Fischer, Science 288 (2000) 669–672.
[76] S. Hacein-Bey-Abina, F. Le Deist, F. Carlier, C. Bouneaud, C. Hue, J.P. De Villartay, A.J. Thrasher, N. Wulffraat, R. Sorensen, S. Dupuis-Girod, A. Fischer, E.G. Davies, W. Kuis, L. Leiva, M. Cavazzana-Calvo, N. Engl. J. Med. 346 (2002) 1185–1193.
[77] S. Hacein-Bey-Abina, C. Von Kalle, M. Schmidt, M.P. McCormack, N. Wulffraat, P. Leboulch, A. Lim, C.S. Osborne, R. Pawliuk, E. Morillon, R. Sorensen, A. Forster, P. Fraser, J.I. Cohen, G. de Saint Basile, I. Alexander, U. Wintergerst, T. Frebourg, A. Aurias, D. Stoppa-Lyonnet, S. Romana, I. Radford-Weiss, F. Gross, F. Valensi, E. Delabesse, E. Macintyre, F. Sigaux, J. Soulier, L.E. Leiva, M. Wissler, C. Prinz, T.H. Rabbitts, F. Le Deist, A. Fischer, M. Cavazzana-Calvo, Science 302 (2003) 415–419.
[78] S. Hacein-Bey-Abina, C. von Kalle, M. Schmidt, F. Le Deist, N. Wulffraat, E. McIntyre, I. Radford, J.L. Villeval, C.C. Fraser, M. Cavazzana-Calvo, A. Fischer, N. Engl. J. Med. 348 (2003) 255–256.
[79] A.J. Thrasher, H.B. Gaspar, C. Baum, U. Modlich, A. Schambach, F. Candotti, M. Otsu, B. Sorrentino, L. Scobie, E. Cameron, K. Blyth, J. Neil, S.H. Abina, M. Cavazzana-Calvo, A. Fischer, Nature 443 (2006) E5–E6 (discussion E6–7).
[80] S. Hacein-Bey-Abina, A. Garrigue, G.P. Wang, J. Soulier, A. Lim, E. Morillon, E. Clappier, L. Caccavelli, E. Delabesse, K. Beldjord, V. Asnafi, E. MacIntyre, L. Dal Cortivo, I. Radford, N. Brousse, F. Sigaux, D. Moshous, J. Hauer, A. Borkhardt, B.H. Belohradsky, U. Wintergerst, M.C. Velez, L. Leiva, R. Sorensen, N. Wulffraat, S. Blanche, F.D. Bushman, A. Fischer, M. Cavazzana-Calvo, J. Clin. Invest. 118 (2008) 3132–3142.
[81] F.D. Bushman, J. Clin. Invest. 117 (2007) 2083–2086.
[82] M. De Palma, E. Montini, F.R. Santoni de Sio, F. Benedicenti, A. Gentile, E. Medico, L. Naldini, Blood 105 (2005) 2307–2315.
[83] L. Naldini, U. Blomer, P. Gallay, D. Ory, R. Mulligan, F.H. Gage, I.M. Verma, D. Trono, Science 272 (1996) 263–267.
[84] R. Zufferey, T. Dull, R. Mandel, A. Bukovsky, D. Quiroz, L. Naldini, D. Trono, J. Virol. 72 (1998) 9873–9880.
[85] G.P. Wang, A. Garrigue, A. Ciuffi, K. Ronen, J. Leipzig, C. Berry, C. Lagresle-Peyrou, F. Benjelloun, S. Hacein-Bey-Abina, A. Fischer, M. Cavazzana-Calvo, F.D. Bushman, Nucleic Acids Res. 36 (2008) e49.
[86] T. Therneau, T. Lumley, Survival: survival analysis, including penalised likelihood. R package version 2.03, (2006).
[87] N.E. Breslow, J. Am. Stat. Soc. 91 (1996) 14–28.
[88] R Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing (2006).
[89] R. Gray, JASA 87 (1992) 942–951.
[90] A. Raftery, J. Hoeting, C. Volinsky, I. Painter, BMA: Bayesian Model Averaging. R package version 3.03 (2006).
[91] M. Stone, J. R. Stat. Soc., Ser. B—Methodol. 36 (1974) 111–147.
[92] S. Holm, SJS 6 (1979) 65–70.
[93] B. Efron, R. Tibshirani, Genet. Epidemiol. 23 (2002) 70–86.
[94] A. Liaw, M. Wiener, R News 2 (2002) 18–22.
[95] M. Margulies, M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen, S.B. Dewell, L. Du, J.M. Fierro, X.V. Gomes, B.C. Godwin, W. He, S. Helgesen, C.H. Ho, G.P. Irzyk, S.C. Jando, M.L. Alenquer, T.P. Jarvie, K.B. Jirage, J.B. Kim, J.R. Knight, J.R. Lanza, J.H. Leamon, S.M. Lefkowitz, M. Lei, J. Li, K.L. Lohman, H. Lu, V.B. Makhijani, K.E. McDade, M.P. McKenna, E.W. Myers, E. Nickerson, J.R. Nobile, R. Plant, B.P. Puc, M.T. Ronan, G.T. Roth, G.J. Sarkis, J.F. Simons, J.W. Simpson, M. Srinivasan, K.R. Tartaro, A. Tomasz, K.A. Vogt, G.A. Volkmer, S.H. Wang, Y. Wang, M.P. Weiner, P. Yu, R.F. Begley, J.M. Rothberg, Nature 437 (2005) 376–380.
[96] F.D. Bushman, C. Hoffmann, K. Ronen, N. Malani, N. Minkah, H.M. Rose, P. Tebas, G.P. Wang, AIDS 22 (2008) 1411–1415.