Video Article

Amplification, Next-generation Sequencing, and Genomic DNA Mapping of Retroviral Integration Sites

Erik Serrao[1], Peter Cherepanov[2], Alan N. Engelman[1]

[1]Department of Cancer Immunology and AIDS, Dana-Farber Cancer Institute

[2]Chromatin Structure and Mobile DNA, The Francis Crick Institute

Correspondence to: Alan N. Engelman at alan_engelman@dfci.harvard.edu

URL: http://www.jove.com/video/53840
DOI: doi:10.3791/53840

Keywords: Virology, Issue 109, Retrovirus, HIV-1, integration, integrase, integration sites, next-generation sequencing

Date Published: 3/22/2016

Citation: Serrao, E., Cherepanov, P., Engelman, A.N. Amplification, Next-generation Sequencing, and Genomic DNA Mapping of Retroviral Integration Sites. *J. Vis. Exp.* (109), e53840, doi:10.3791/53840 (2016).

Abstract

Retroviruses exhibit signature integration preferences on both the local and global scales. Here, we present a detailed protocol for (1) generation of diverse libraries of retroviral integration sites using ligation-mediated PCR (LM-PCR) amplification and next-generation sequencing (NGS), (2) mapping the genomic location of each virus-host junction using BEDTools, and (3) analyzing the data for statistical relevance. Genomic DNA extracted from infected cells is fragmented by digestion with restriction enzymes or by sonication. After suitable DNA end-repair, double-stranded linkers are ligated onto the DNA ends, and semi-nested PCR is conducted using primers complementary to both the long terminal repeat (LTR) end of the virus and the ligated linker DNA. The PCR primers carry sequences required for DNA clustering during NGS, negating the requirement for separate adapter ligation. Quality control (QC) is conducted to assess DNA fragment size distribution and adapter DNA incorporation prior to NGS. Sequence output files are filtered for LTR-containing reads, and the sequences defining the LTR and the linker are cropped away. Trimmed host cell sequences are mapped to a reference genome using BLAT and are filtered for minimally 97% identity to a unique point in the reference genome. Unique integration sites are scrutinized for adjacent nucleotide (nt) sequence and distribution relative to various genomic features. Using this protocol, integration site libraries of high complexity can be constructed from genomic DNA in three days. The entire protocol that encompasses exogenous viral infection of susceptible tissue culture cells to integration site analysis can therefore be conducted in approximately one to two weeks. Recent applications of this technology pertain to longitudinal analysis of integration sites from HIV-infected patients.

Video Link

The video component of this article can be found at http://www.jove.com/video/53840/

Introduction

Integration of viral DNA (vDNA) into the host cell genome is an essential step in the retroviral life cycle. Integration is accomplished by the viral enzyme integrase (IN), which carries out two distinct catalytic processes that lead to the establishment of the stably inserted provirus [1]. IN subunits engage the ends of the linear vDNA that is generated through reverse transcription, forming the higher-order intasome with vDNA ends held together by an IN multimer [2-4]. IN cleaves the 3' ends of the vDNA downstream from invariant 5'-CA-3' sequences in a process known as 3'-processing, leaving recessed 3' ends with reactive hydroxyl groups at each vDNA terminus [5-8]. The intasome is subsequently imported into the nucleus as part of a large assembly of host and viral proteins known as the preintegration complex (PIC) [9-11]. After encountering cellular target DNA (tDNA), IN uses the vDNA 3'-hydroxyl groups to cleave the tDNA top and bottom strands in a staggered fashion and simultaneously joins the vDNA to tDNA 5' phosphate groups through the process of strand transfer [12,13].

Retroviruses exhibit integration site preferences on the local and global scales. Locally, consensus integration sites consist of weakly conserved palindromic tDNA sequences that span from approximately five to ten bp upstream and downstream from the vDNA insertion sites [14,15]. Globally, retroviruses target specific chromatin annotations [16]. There are seven different retroviral genera - alpha through epsilon, lenti, and spuma. The lentiviruses, which include HIV-1, favor integration within the bodies of actively transcribed genes [17], while the gammaretroviruses preferentially integrate into transcriptional start sites (TSSs) and active enhancer regions [18-20]. In sharp contrast, spumavirus is strongly biased towards heterochromatic regions, such as gene-poor lamina-associated domains [21]. Local tDNA base preferences are in large part dictated by specific networks of nucleoprotein contacts between IN and tDNA [13,22,23]. For the lentiviruses and gammaretroviruses, integration relative to genomic annotations is in large part governed by interactions between IN and cognate cellular factors [24-27]. Altering the specifics of the IN-tDNA interaction network [13,22,23,28] and disrupting or re-engineering IN-host factor interactions [25-27,29-32] are proven strategies to retarget integration on the local and global levels, respectively.

The power of DNA sequencing procedures used to catalogue retroviral integration sites has increased immensely over the past decades. Integration sites were recovered in pioneering work using laborious purification and manual cloning techniques to yield just a handful of unique sites per study [33,34]. The combination of LM-PCR amplification of LTR-host DNA junctions with the ability to map individual integration sites to human and mouse draft genomes transformed the field, with the number of sites recovered from exogenous tissue culture cell infections

increasing to several hundred to thousands [17,18]. The more recent combination of LM-PCR with NGS methodology has sent library depth skyrocketing. Specifically, pyrosequencing yielded on the order of tens of thousands of unique integration sites [30,35-38], while libraries sequenced through the use of DNA clustering can yield millions of unique sequences [19-21,39]. Here we describe an optimized LM-PCR protocol for amplifying and sequencing retroviral integration sites using DNA clustering NGS. The method incorporates required adapter sequences into the PCR primers and hence directly into the amplified DNA molecules, thereby precluding the requirement for an additional adapter ligation step prior to sequencing [40]. The bioinformatic analysis pipeline, from the parsing of raw sequencing data for LTR-host DNA junctions to the mapping of unique integration sites to pertinent genomic features, is also generally described. In accordance with the precedence established from prior methodological protocols in this field [36,38,41-43], custom scripts can be developed to aid the completion of specific steps in the bioinformatics pipeline. The utility and sensitivity of the protocol is illustrated with representative data by amplifying, sequencing, and mapping HIV-1 integration sites from tissue culture cells infected at the approximate multiplicity of infection (MOI) of 1.0, as well as a titration series of this DNA diluted through uninfected cellular DNA in 5-fold steps to a maximum dilution of 1:15,625 to yield the approximate equivalent MOI of 6.4 x 10^{-5}.

Protocol

1. Generate Virus Stocks

Note: A flow chart of the wet bench aspect of this protocol is depicted in **Figure 1**. The details of viral stock production and subsequent infection of tissue culture cells will generally apply to different types of retroviruses. For some experiments, the target cell may not express the endogenous viral receptor(s), and in such cases the construction of pseudotyped retroviral particles harboring heterologous viral envelope glycoprotein, *e.g.* the G glycoprotein from vesicular stomatitis virus (VSV-G), will be required for infection [44,45].

Note: Precaution should be taken when working with HIV-1. Though specific guidelines will vary from institution to institution, all virus-based work should be conducted in a dedicated, operator restricted biological safety cabinet (typically referred to as a tissue culture hood). Proper personal protective equipment that includes face protection, shoe covers, a double glove layer, and a full-body coverall suit should be worn at all times. All liquid waste resulting from virus-related experiments should be inactivated with bleach (10% final concentration), and all waste including solids should be autoclaved prior to disposal.

1. One day prior to transfection, plate 3.3 x 10^6 HEK293T cells in 10 ml of Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% (v/v) fetal bovine serum and 1% (v/v) penicillin/streptomycin (10,000 U/ml stock) in each of five 100 mm dishes.
   Note: Supplemented-DMEM is referred to as DMEM-FPS from this point on.
2. On the subsequent day, transfect the cells with 10 µg of plasmid carrying full-length retroviral molecular clones or 9 µg of envelope-deleted single-round vectors with 1 µg of a VSV-G expression construct using commercially available transfection reagents or calcium phosphate.
   1. Incubate the cells at 37 °C in a humidified cell culture incubator with 5% CO_2 (this condition hereafter referred to as "tissue culture incubator"). After approximately 48 hr, harvest the virus-containing cell media using a volumetric pipette and pass it through a 0.45 µm filter by gravity flow.
   2. Concentrate the virus by ultracentrifugation at 200,000 x g for 1 hr at 4 °C. Resuspend the virus pellet in 500 µl DMEM-FPS containing 20 U DNase, and incubate for 1 hr at 37 °C.
      Note: The DNase step helps to reduce the recovery of unwanted plasmid sequences by eliminating the brunt of plasmid DNA that persists from the transfection procedure.
3. Determine p24 concentration [46] using an HIV-1 p24 antigen capture kit as per manufacturer's instructions.
   Note: Virus concentration can also be determined by reverse transcriptase activity assay [47,48]. Alternatively, the level of functional virus can be determined by measuring MOI. This is most readily done using fluorescence-activated cell sorting with viruses that express fluorescent reporter genes such as enhanced green fluorescent protein. MOI determination may be particularly useful when working with primary cells that may not support the same level of infection as optimized cell lines.

2. Infect Cells with Virus

1. Plate 3.0 x 10^5 HEK293T cells per well in a 6-well plate in 2.5 ml DMEM-FPS and incubate overnight in a tissue culture incubator.
   Note: The number of unique integration sites recovered with this protocol is directly proportional to the number of cells and quantity of active virus used in the infection.
2. Infect cells with a final viral p24 concentration of 500 ng/ml in a final volume of 500 µl fresh DMEM-FPS for 2 hr in a tissue culture incubator, then add 2 ml DMEM-FPS pre-warmed to 37 °C per well and continue incubation.
3. At 48 hr post-infection, remove the media and wash the cells with 2 ml phosphate-buffered saline (PBS). Add 0.5 ml trypsin-EDTA pre-warmed to 37 °C, and after a few sec visually inspect the wells for cell dislodgement.
4. Add 2 ml pre-warmed DMEM-FPS and resuspend the cells by gentle up/down pipetting with a volumetric pipette ~10 times. Transfer the solution to a 75 cm^2 tissue culture flask containing 18 ml pre-warmed DMEM-FPS, and incubate the cells in a tissue culture incubator.
5. After minimally five days from the start of the infection, collect the cells by removing the media, wash with 5 ml PBS, add 2 ml pre-warmed trypsin-EDTA, and resuspend with 5 ml pre-warmed DMEM-FPS by pipetting. Centrifuge the solution for 5 min at room temperature at 2,500 x g, and discard the supernatant.
   Note: Although integration under these conditions plateaus at about 48 hr post-infection [49,50], the additional 3 days of culture are required to sufficiently dilute the concentration of unintegrated DNA molecules that result from cell-based DNA recombination or viral-mediated autointegration.
6. Extract genomic DNA from the cell pellet using a commercially available kit (*e.g.*, see [51]). Elute the DNA from the supplied ion exchange column with 200 µl of 10 mM Tris-HCl, pH 8.5.
   Note: An aliquot of cells should be apportioned at 48 hr post-infection (Step 2.3) for an infectivity assay to ensure proper virus infection prior to NGS.

3. Fragment Genomic DNA by Sonication or by Restriction Enzyme Digest

Note: Sonication fragments genomic DNA in a virtually sequence-independent manner and is thus the preferable mode of fragmentation when sequencing samples with a low expected recovery rate (*e.g.*, infected patient cells or infections initiated at relatively low MOI). Furthermore, sonication allows one to distinguish PCR duplicates of a particular integration site sequence from unique integrations at the same site, which is critical to distinguish the clonal expansion of provirus-containing cells in infected patients (see Step 11 below) [39,52-54].
Note: The DNA should be cleaved immediately downstream from the upstream LTR to diminish amplification of internal viral sequences during LM-PCR. The restriction enzyme BglII that lies 43 bp downstream from the upstream U5 sequence and that is incompatible for subsequent ligation with MseI-generated DNA ends works well with many HIV-1 strains (**Figure 1B**). When preparing DNA by sonication, the internal-cleaving restriction enzyme should be applied after linker ligation (see **Figure 1C**-**E** and Step 4.3 below).

1. For sonication, mix 10 µg of genomic DNA in nuclease-free water to a final volume of 120 µl. Sonicate using parameters for an average break size of 500 bp (two rounds of the following parameters: duty cycle: 5%; intensity: 3; cycles per burst: 200; time: 80 sec).
2. Purify sonicated DNA using a PCR purification kit. Repair the DNA ends using a DNA end-repair kit and purify the DNA using a PCR purification kit. A-tail the DNA using Klenow exo$^-$ enzyme and purify the A-tailed DNA using a PCR purification kit. Refer to [51,52] for additional details of kit usage.
3. For restriction endonuclease digestion, cut 10 µg of genomic DNA overnight at 37 °C in a volume of 100 µl with buffer supplied by the manufacturer and a cocktail of enzymes (100 U each) that generate 5'-TA overhangs, as well as an incompatible enzyme such as BglII that cleaves downstream from the upstream viral LTR. Purify the DNA the next day using a PCR purification kit.
   Note: None of the restriction enzymes should cut within the terminal ~30 bp of the viral DNA end that is amplified by the LM-PCR protocol. This protocol specifically amplifies the U5 end of HIV-1 DNA.

4. Anneal Linker Oligonucleotides and Ligate to Fragmented Genomic DNA

Note: Prepare an asymmetric linker containing an overhang that is compatible with the above DNA fragments (see **Table 1** for the sequences of oligonucleotides utilized in this protocol). The linker to be used with sonicated DNA must contain a compatible T-3' overhang, while the linker for MseI-digested DNA must contain a compatible 5'-TA overhang (**Figure 1**). The short linker strand must additionally contain a non-extendable chemical modification, such as 3'-amine, to constrain the subsequent amplification reactions toward the DNA of interest.
Note: When preparing multiple different integration site libraries in parallel and/or when multiplexing unique samples on the same sequencing run, it is recommended to use unique linkers for each sample to limit the potential for sample cross-contamination during PCR. This additionally implies the use of unique linker primers for each sample during semi-nested PCR (described below). Unique linker strands and linker primers may be designed by scrambling the linker oligonucleotide sequences listed in **Table 1** while maintaining similar overall %GC content and applicable overhang positions.

1. Anneal the short and long linker strands in 35 µl of 10 mM Tris-HCl, pH 8.0-0.1 mM EDTA (final concentration of 10 µM of each oligonucleotide) by heating to 90 °C and slowly cooling to room temperature in steps of 1 °C per min.
2. Prepare at least four parallel ligation reactions per genomic DNA sample, which contain 1.5 µM ligated linker, 1 µg fragmented DNA, and 800 U T4 DNA ligase in 50 µl. Ligate overnight at 12 °C. Purify the next day with a PCR purification kit.
3. For samples prepared by sonication, digest the purified ligation reaction with 100 U of a restriction enzyme that cleaves downstream from the upstream LTR (*e.g.*, BglII for HIV-1) under the manufacturer's recommended conditions overnight. Purify the DNA using a PCR purification kit.

5. Amplify Viral LTR-Host Genomic DNA Junctions by Semi-nested PCR

Note: To ensure for optimal library diversity, at least 4-8 parallel PCRs, depending on the DNA concentration of the recovered ligation reaction, should be prepared for each sample for both PCR rounds. DNA template concentration should be quantified by spectrophotometry. In this protocol the first and second rounds of PCR employ nested LTR-specific primers, but the same linker-specific primer is used for both rounds (**Table 1**). The second round LTR-specific primer and the linker-specific primer encode adapter sequences for DNA clustering as well as sequencing primer-binding sites. The nested LTR-specific primer also encodes a 6 nt index sequence, which can be varied among different primers for multiplexing libraries within the same sequencing run.

1. Prepare first round PCRs containing the ingredients per tube as listed in **Table 2**.
   Note: The linker-specific primer harbors 22 nt of complementarity to the linker, a melting temperature of 53 °C, a GC-content of 45%, and its 3' end is located 15-16 bp upstream from the 3' termini of the different linker long strands (**Table 1**). The first round 27 nt LTR primer has a melting temperature of 59 °C, a GC-content of 48%, and its 3' end is located 34 bp upstream from the HIV-1 U5 terminus. The region of the second round 26 nt LTR primer that is complementary to the HIV-1 LTR has a melting temperature of 60 °C, a GC-content of 50%, and its 3' end is located 18 bp upstream from the viral U5 terminus. It is recommended that oligonucleotide melting temperature and GC-content should mimic these parameters if users design PCR primers with altered sequences (including for use with other retroviruses) [21].
2. Run first PCR round under the following thermocycler parameters: One cycle: 94 °C for 2 min; 30 cycles: 94 °C for 15 sec, 55 °C for 30 sec, 68 °C for 45 sec; one cycle: 68 °C for 10 min.
3. Pool reactions and purify using a PCR purification kit. Prepare second round PCRs containing the ingredients per tube as per **Table 3**. Run the second round of PCR using the thermocycler parameters described in Step 5.2. Pool the reactions and purify the DNA using a commercial PCR purification kit following the manufacturer's instructions.
   Note: A variety of recommended index sequences compatible with DNA clustering NGS are available [71].

6. Perform QC and NGS (Typically Completed by a Sequencing Facility)

1. (QC assay #1) Confirm Step 5.3 library DNA concentration using a fluorometer [55]. Briefly, prepare standards and experimental samples in a final volume of 200 µl nuclease-free water. Vortex tubes for 2-3 sec, incubate at room temperature for 2 min, and then read the samples in the fluorometer.
   Note: Samples should contain a minimum concentration of 2 nM library DNA in a minimum volume of 15 µl.
2. (QC assay #2) Confirm DNA fragment size distribution using a tape-based assay [56].
   Note: An ideal distribution is a relatively broad DNA peak centering around 500 bp in length. If a significant amount of material is larger than 1 kb, then it is recommended to incorporate a size-selection procedure to eliminate longer DNA species, which will impede bridge amplification during clustering. By contrast, if a significant peak is apparent around 100 to 200 bp, a primer dimer may have formed during PCR. In this case the procedure should be optimized to minimize the formation of primer dimers.
3. (QC assay #3) Confirm proper incorporation of adapters into DNA library by quantitative PCR [57].
4. Perform NGS following the manufacturer's application literature. Utilize a spike-in of 10% (w/w) ΦX174 DNA, which will optimize real time quality metrics by providing balanced base composition to the sequencing run.
   Note: Integration site sequencing experiments are typically subjected to single end 150 bp (SE150) or paired-end 150 bp (PE150) sequencing. PE150 is particularly useful to capture the linker attachment point on each DNA molecule (*e.g.*, when scrutinizing integration sites for evidence of host cell clonal expansion).

7. Use a Customized Python or PERL Script to Parse Sequencing Data for LTR-containing Sequences, Crop away LTR and Linker Sequences, and Map to Reference Genome with BLAT

1. Scan FASTA files for LTR-containing sequence reads, crop LTR and linker sequences away from host genomic DNA sequence, and export these sequences to a new FASTA file. Map cropped reads to both a reference genome (*e.g.* human genome versions hg19 or GRCh38) and the viral genome using BLAT [58], with output integration site coordinates exported to a separate .txt file, using the following settings: stepSize = 6, minIdentity = 97, and maxIntron = 0
2. Parse the BLAT output .txt file, remove autointegrations (*i.e.* evidence that the LTR end has integrated into an internal region of the viral DNA genome) and other sequences mapping to the HIV-1 genome, and create a separate output .txt file in which all duplicate integration sites have been condensed into single, unique coordinate hits.

8. Create .bed Files Containing 15-Nt Intervals Surrounding Integrations, Convert These to FASTA Files, and Construct Sequence Logos to Display Base Preferences Surrounding Integration Sites

1. Create .bed files that list an interval of bases for each integration site. At least 15 bases (5 upstream and 10 downstream) are suggested for sequence logo generation. Generate a FASTA file from these .bed files by using the *fastaFromBed* function from BEDTools [59] and this command:
   fastaFromBed -fi /directory/to/reference/genome/ -name -s -bed 15_base_pair_file.bed -fo output_file.fasta
   Note: The invariant viral 5'-CA-3' dinucleotide is joined to host DNA during integration, and verifying the junction of the LTR terminus to cellular DNA is an important initial filter to identify bona fide integration sites. We additionally compile sequence logos from this host DNA sequence population to verify the experimental results. As retroviruses display signature base preferences surrounding their integration sites [14,15], the sequence logos serve to validate that the mapped genomic sites arose through IN-mediated integration as compared to other recombination mechanisms such as non-homologous DNA end joining [60,61].
2. Use WebLogo 3 (http://weblogo.threeplusone.com/create.cgi) to create sequence logos from the FASTA files. Click 'Choose File' to upload FASTA file, and use the following settings: Output format, PDF (vector); Logo size, large; First position number, -5; Logo range, -5 to 5; Y-axis scale, 0.1, Y-axis tic spacing, 0.5, Color scheme, classic (NA).

9. Create Central Base Pair .bed Files, Check for Sample Cross-Contamination, and Map the Distribution of Unique Integration Sites Relative to Pertinent Genomic Features

1. Since retroviral integration occurs in a staggered fashion across the tDNA strands, adjust the precise coordinates of integration sites to reflect the central bp of the target site duplication for correct mapping of genomic distribution relative to genomic features.
   1. Therefore, for 5 bp duplicating viruses like HIV-1, create a .bed file with the central bp offset from the integration site by two bases downstream for integrations mapping to the plus strand, and two bases upstream for integrations mapping to the minus strand.
2. To check for sample cross-contamination, calculate the number of integration sites common among the different libraries by using the BEDTools *intersect* function to intersect central bp .bed files for two different samples and by following this command:
   bedtools intersect -a central_basepair_1.bed -b central_basepair_2.bed -f 1.00 -r -s > overlap1v2.txt
3. Count the number of lines within the output overlap1v2.txt file in order to quantify the exact number of sites common among the two libraries by using the following command:
   wc -l overlap1v2.txt
4. Download the RefSeq annotation .bed file for the version of reference genome that was used for integration site mapping from the UCSC Genome Annotation Database (*e.g.* http://hgdownload.cse.ucsc.edu/goldenPath/hg38/database) [62].
   1. Calculate the number of integration sites falling within RefSeq genes by using the BEDTools *intersect* function to intersect the central base pair .bed file that was generated for the sample with the RefSeq .bed file following this command:

bedtools intersect -a central_basepair_1.bed -b RefSeq_hg38.bed -u > RefSeq_sample1.bed

5. Count the number of lines within the output RefSeq_sample1.bed file in order to quantify the exact number of sites falling in RefSeq genes by using the following command:
   wc -l RefSeq_sample1.bed
6. Repeat steps 9.3 and 9.4 for mapping integration sites to any other annotation of interest for which an interval .bed file is available. Download the most current CpG island annotation .bed file for the reference genome of interest from the UCSC Genome Annotation Database as directed in Step 9.4.
   1. Calculate the number of integration sites falling within a certain distance (illustrated in this example is a 5 kb window) of CpG islands by using the BEDTools *window* function and following this command:
      bedtools window –w 2500 central_basepair_1.bed -b CpG_hg38.bed -u > CpG_sample1.bed
7. Count the number of lines within the output CpG_sample1.bed file in order to quantify the exact number of sites falling within 2.5 kb upstream or downstream of CpG islands by using the following command:
   wc -l CpG_sample1.bed
8. Repeat steps 9.6 and 9.7 for mapping integration sites nearby TSSs. Generate an alternate version of the RefSeq.bed file, where genomic coordinates mapping to more than one gene have been adjusted to reflect only a single gene present at that position. This prevents overestimation of gene density surrounding integration sites. Calculate the gene density in the 1 Mb region surrounding each integration site by using the BEDTools *window* function and following this command:
   bedtools window -w 500000 central_basepair_1.bed -b RefSeq_hg38_NonRedundant.bed -u > GeneDensity_sample1.bed
9. Calculate the average gene density for all integrations in the dataset by following this command:
   awk '(sum+=$7) END (print "Average = ",sum/NR)' GeneDensity_sample1.bed

10. Statistically Compare Integration Site Distributions among Samples Using Two-tailed Fisher's Exact Test and Two-tailed Wilcoxon Rank Sum Test in *R*

Note: Use Fisher's exact test for comparing the proportion of integration sites within RefSeq genes or within a window of CpG islands or TSSs, but use the Wilcoxon rank sum test for comparing the distribution in gene density surrounding the integration sites. The *R* program is available at http://www.r-project.org/.
Two-tailed Fisher's exact test:

1. Using the numbers calculated as instructed in Steps 9.4 and 9.7, create matrices for each comparison in *R* of observed occurrences (integrations within an annotation or within a window surrounding an annotation) versus remaining sites by following this command:
   (annotation_of_interest <- matrix(c(SampleA#in, SampleA#remaining, SampleB#in, SampleB#remaining),nrow=2,dimnames=list(c('Center', 'Remainder'),c('SampleA', 'SampleB'))))
2. Calculate the *P* value for the comparison by two-tailed Fisher's exact test with the following command:
   fisher.test(annotation_of_interest, alternative = 'two.sided')$p.value
   Two-tailed Wilcoxon rank sum test:
3. Create a tab-delimited .txt file in which each column contains the sample name in the top cell, followed below by the gene density values for all integration sites in that library (obtained from the .bed file generated in Step 9.9). Import this tab-delimited .txt file into *R* using the following command and navigating to the correct file directory:
   FILENAME <- as.data.frame(read.delim(file.choose(), header=T,check.names=FALSE, fill=TRUE,sep='\t'))
4. Calculate the *P* value for the comparison by two-tailed Wilcoxon rank sum test with the following command:
   wilcox.test(FILENAME$SampleA,FILENAME$SampleB,alternative ='two.sided', paired = F, exact = T)$p.value
   Note: *P* values can be calculated only down to a certain (extremely low) limit in *R*, after which zero will be returned by the program. For massively different samples that yield a *P* = 0 in *R*, estimate the *P* value as $<2.2 \times 10^{-308}$.

11. Examine Raw Sequencing Data for Evidence of Clonal Expansion of Cells Containing Integrated Viral DNA

Note: A small potential exists for more than one integration at the exact same nt in the reference genome. Alternatively, a single integration event may become redundantly present in sequencing data due to the use of PCR during library preparation and/or by cell duplication prior to DNA preparation. Recent analyses of genomic DNA from HIV-infected patients have distinguished these possibilities by identifying unique sonication shear points/linker attachment points (which can only arise prior to PCR) within DNA sequences containing identical integration sites [52-54]. There is currently a debate as to whether proviruses harbored within clonally expanded cells contribute to the latent viral reservoir, and thus it is of particular interest to characterize their level of expansion when studying integration sites in human patients.

1. Similar to the procedure listed in Step 8.1, generate .bed files listing an interval of bases extending, in this case, 25 nt downstream from each unique integration site (upstream bases are unnecessary here). Generate a FASTA file from these .bed files (as instructed in Step 8.1) by using the *fastaFromBed* function from BEDTools and following this command:
   fastaFromBed -fi /directory/to/reference/genome/ -name -s -bed 25_base_pair_file.bed -fo output_file.fasta
   Note: To improve the specificity of each search it is recommended to extract at least 25 nt downstream from each integration site for clonal expansion analyses.
2. Preferably using a customized script, search the raw sequence data FASTA file for all strings containing an exact match to the 25 nt downstream from each unique integration site, and deposit these sequences into a new file. Trim LTR and linker sequences from the raw strings. Merge PE sequence reads by converting reads to the reverse complement, trimming LTR and linker sequences, and then assigning read2 strings to their read1 pair if the strings share at least 20 overlapping nt.

3. Scan the linker attachment points of each integration site block. Classify each integration as "clonally expanded" if linker attachment points are ≥3 bp apart.
   Note: A protocol for clonal expansion analysis without merging sequence reads has been described [52].
   Note: Fragmentation of the genome at the exact same location by sonication leads to an underestimation of the extent of clonal expansion, and methods to correct the resulting experimental bias have been described [63,64].

Representative Results

Table 4 lists the results of a representative experiment to illustrate the sensitivity of NGS for recovering integration sites from a culture of infected cells. Uninfected cellular DNA was utilized to serially dilute genomic DNA from an infection in which every cell on average contained one integration [40]. Dilutions were prepared in steps of five to a maximal dilution of 1:15,625. Genomic DNA in the titration series was then fragmented by sonication or by digestion with restriction endonucleases MseI and BglII, followed by LM-PCR. The numbers of unique integration sites, as well as the number of sites mapping proximal to selected genomic annotations, were calculated according to the above protocol. Data analysis revealed dozens of unique integration sites (1-2% of the amount recovered from neat genomic DNA) recovered from libraries prepared from cells where in theory only one in 15,625 was infected.

When analyzing integration site datasets, it is critical to compare the data to a matched set of random genomic sites, which is called a matched random control or MRC. As the representative results sheared genomic DNA by restriction enzyme digestion or by sonication, two different MRC datasets were constructed. MRC_{enz} contained 50,000 unique genomic sites generated by randomly selecting sites from hg19 in proximity to the sites of MseI and BglII restriction enzyme digestion, whereas MRC_{random} harbored 10,000 sites generated without normalization for distance from set genomic markers. Only the sites that can be mapped back to a unique genomic location should be used in MRC datasets. As sonication shears genomic DNA essentially free from sequence bias, MRC_{random} may be viewed as more applicable to datasets produced by fragmentation of DNA by sonication. An alternative style of control integration site dataset can be generated *in vitro* by reacting recombinant IN protein, intasome nucleoprotein complex [21], or PICs extracted from acutely infected cells [17] with deproteinized genomic DNA, and then following the LM-PCR and NGS protocols [21].

P values for comparison of the distribution of integration sites recovered by sonication versus restriction digest (comparison is between the neat samples), as well as for comparison to the MRC_{enz} and MRC_{random}, are displayed in **Figure 2**. The distribution of integration sites recovered following sonication was similar to those recovered by restriction enzyme digest for all annotations examined, with the greatest variance evident in terms of proximity to CpG islands. As expected [18,65] both datasets differed significantly from the MRCs in terms of integrations within RefSeq genes and gene density surrounding the average integration site, while both datasets were similar to the MRCs in terms of distribution relative to CpG islands and TSSs. Since relatively few HIV-1 integration sites map within 2.5 kb of a CpG island or TSS, increasing the total number of sites recovered is likely to decrease the variability that can arise between datasets (**Table 4** and **Figure 2**). Sequence logos to confirm the authenticity of the integration site data are shown in **Figure 3**. The consensus HIV-1 integration site [14,22] (-3)TDG\(G/V)TWA(C/B)CHA(+7) (written using International Union of Biochemistry base codes; the backslash indicates the position of vDNA plus-strand joining, and the underline indicates the 5-bp sequence duplicated following HIV-1 integration and DNA repair) is apparent for libraries prepared by both fragmentation techniques, although the degree of certainty decreases with increasing dilution of infected cell DNA. The random sites aligned from the MRC dataset by contrast failed to generate appreciable levels of base preferences.

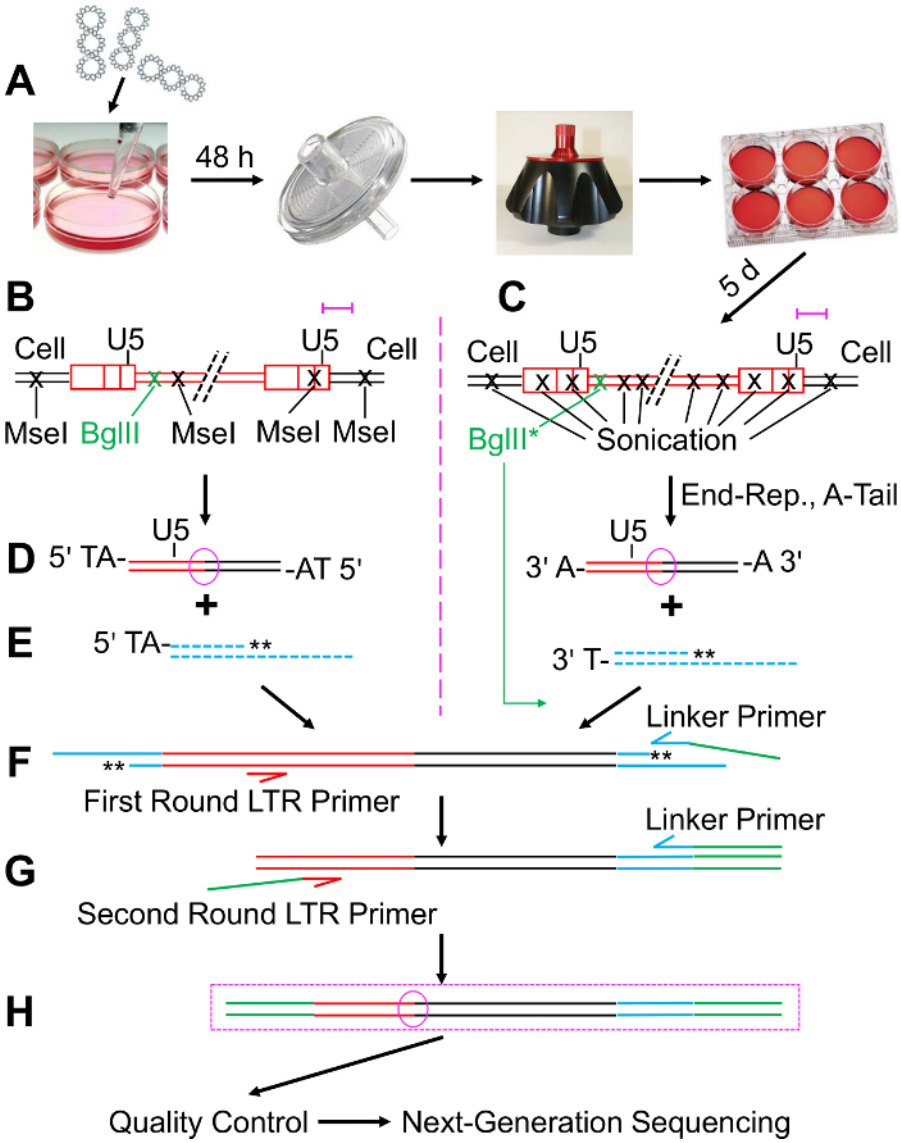

Figure 1: Flow Chart Illustration of Integration Site Library Preparations. (**A**) Generate virus stocks by transfecting HEK293T cells, harvesting and filtering supernatant 48 hr later, concentrating by ultracentrifugation, and infecting target cells with appropriate concentration of virus. At least five days after infection, extract genomic DNA. Refer to Sections 1 and 2 of main text for additional experimental details. (**B** and **C**) Fragment purified genomic DNA by digestion with restriction enzymes or by sonication. The restriction enzyme cocktail should include an enzyme (*e.g.* BglII) that cleaves downstream from the upstream viral LTR to counter-select for LM-PCR amplification of internal vDNA sequences. Green asterisk and branched arrow in (**C**) denote that BglII should be applied after linker ligation. Red highlights viral sequence, while black highlights host cellular sequence. Implied DNA break points (not to scale) are marked by "X." HIV-1 contains numerous MseI and BglII sites; only those relevant to the protocol are shown. The brackets above the maps denote the U5-cellular DNA regions preferentially amplified by LM-PCR. (**D**) Purify fragmented DNA (then end-repair and A-tail in the case of sonication) and ligate to (**E**) compatible asymmetric linker molecules (colored blue). Magenta circles in (**D**) indicate the integration site that will be amplified. Asterisks at the 3' ends of the linker short strands denote amino blocking modifications. (**F**) Conduct first round of semi-nested PCR using first round LTR primer (red) and linker primer (blue). In this PCR round, the linker primer encodes for DNA clustering and NGS primer binding sequences (grouped as a green appendage to the blue linker primer), while the LTR primer lacks such sequences. (**G**) Purify first round PCR product and conduct second round of semi-nested PCR. In this round of PCR, use the same linker primer as in the first round (blue + green appendage), together with the second round LTR primer (red) that carries DNA clustering and NGS primer binding sequences as well as a barcode for multiplexing (grouped as a green appendage to the red LTR primer). (H) Purify second round PCR product as the final integration site library (boxed in magenta, with integration site marked by magenta circle). Submit aliquot to sequencing facility for QC and NGS. Please click here to view a larger version of this figure.

Within RefSeq Gene[a]

	MRC_{random}[d]	Sonication	Digest
MRC_{enz}[c]	3.96E-10	5.93E-188	<2.2E-308
MRC_{random}		3.65E-193	2.40E-310
		Sonication	***0.16***

± 2.5 kB of a CpG Island[a]

	MRC_{random}	Sonication	Digest
MRC_{enz}	1.64E-06	0.02	0.01
MRC_{random}		***0.72***	8.03E-08
		Sonication	3.73E-04

± 2.5 kB of a TSS[a]

	MRC_{random}	Sonication	Digest
MRC_{enz}	***0.36***	***0.43***	2.00E-06
MRC_{random}		***0.22***	4.51E-06
		Sonication	0.03

Avg. Genes/Mb[b]

	MRC_{random}	Sonication	Digest
MRC_{enz}	3.11E-13	3.90E-198	<2.2E-308
MRC_{random}		8.21E-188	2.23E-298
		Sonication	***0.05***

Figure 2: *P* Values for Comparison of Integration Sites Amplified Following DNA Fragmentation by Sonication or by Restriction Enzyme Digestion versus Respective MRCs. Numbers of integration sites within RefSeq genes and nearby CpG islands and TSSs, as well as regional gene density profiles, are listed in **Table 4**. *P* values ≥0.05 are highlighted in bold and italic text. [a]*P* values calculated by Fisher's exact test. [b]*P* values calculated by Wilcoxon rank sum test. [c]MRC_{enz}: matched random control; a set of 50,000 unique integration sites was produced by randomly selecting positions in proximity to MseI/BglII restriction sites in hg build 19. [d]MRC_{random}: matched random control containing 10,000 unique integration sites produced by randomly selecting positions in hg19 without normalization to restriction site proximity. Please click here to view a larger version of this figure.

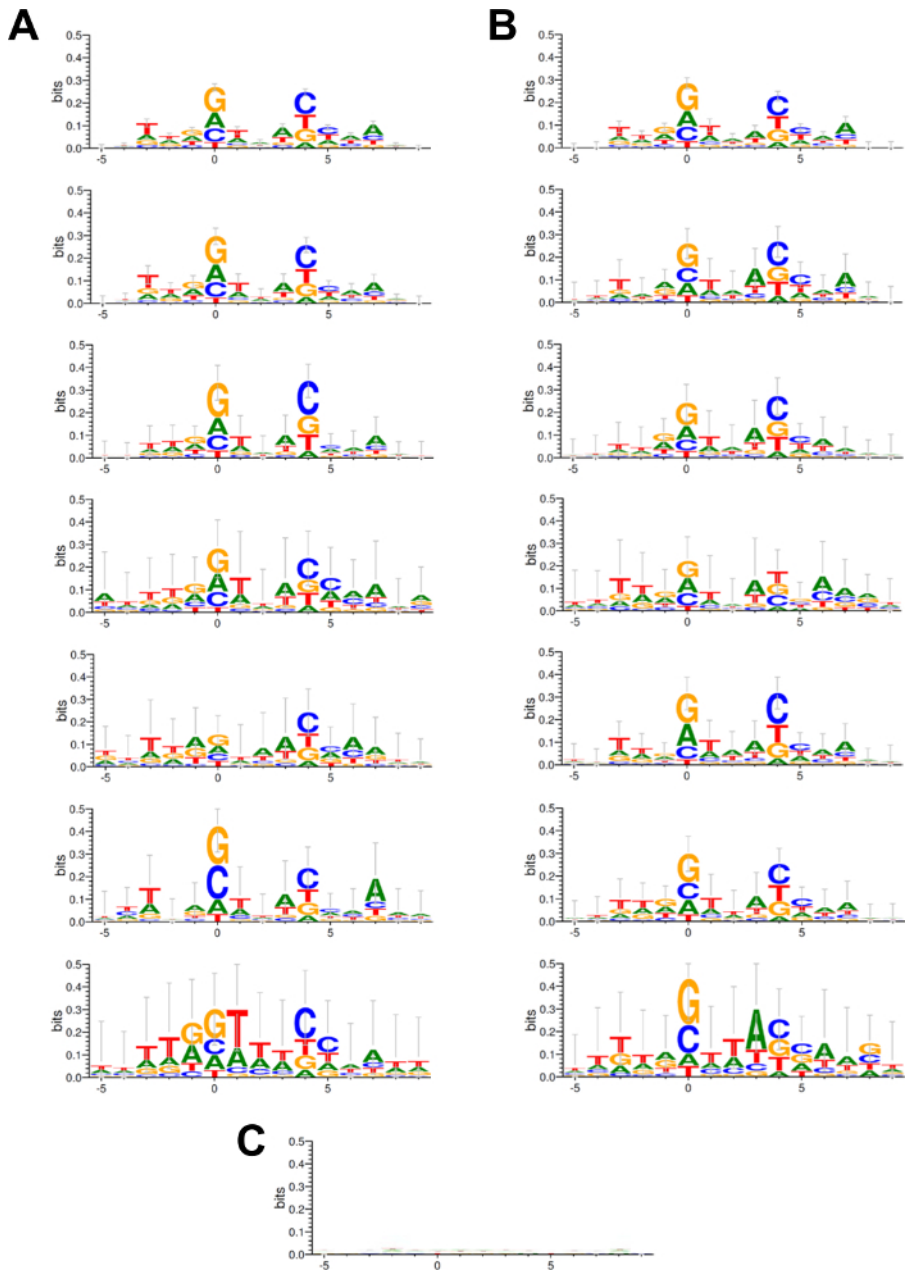

Figure 3: Sequence Logos Depicting HIV-1 Base Preferences from Representative Experiment Libraries. Integration sites from libraries prepared by (**A**) digestion with restriction enzymes or (**B**) sonication were aligned using WebLogo software. Each dilution in the titration series is depicted, from neat DNA at the top of the figure to the maximum dilution of 1:15,625 at the bottom. (**C**) Sequence logo for the MRC of 50,000 unique genomic sites. Error bars essentially represent the standard deviation in base incorporation at any particular position. More specifically, the total height of each error bar is equivalent to twice the small sample correction [66], which controls for underestimation of entropy present in relatively small datasets. The x-axis represents host cell genomic DNA nt positions relative to the site of integration at point zero. Please click here to view a larger version of this figure.

Sequence (5'-3')	Usage
GTCCCTTAAGCGGAG-NH_2 [a]	Sonication Linker Short Strand
GTAATACGACTCACTATAGGGCCTCCGCTTAAGGGACT	Sonication Linker Long Strand
TAGTCCCTTAAGCGGAG-NH_2 [a]	Restriction Digest Linker Short Strand
GTAATACGACTCACTATAGGGCCTCCGCTTAAGGGAC	Restriction Digest Linker Long Strand
CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTG AACCGCTCTTCCGATCTGTAATACGACTCACTATAGGGC	Linker-Specific Primer
TGTGACTCTGGTAACTAGAGATCCCTC	First Round LTR Primer
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGAC GCTCTTCCGATCTCGATGTGAGATCCCTCAGACCCTTTTAGTCAG	Second Round LTR Primer

Table 1: Oligonucleotide Sequences for Linker Construction and PCR Amplification. Linker-specific and second round LTR primers encode DNA clustering adapter sequences, which are color-coded as follows: black, bases complementary to the linker or to the HIV-1 LTR; red, unique index or barcode; green, sequencing primer binding sites; blue, adapter sequences for DNA clustering. Single-end (SE) sequencing reactions will utilize the sequencing primer that anneals to the second round LTR primer read1 (green) sequence, while paired-end (PE) reactions will use both (read1 and read2) sequencing primers. [a]Linker short strands contain 3' amino blocking modification. Please click here to view a larger version of this table.

Reagent	To Add per Reaction
First Round LTR primer (15 µM):	2.5 µl
Linker-specific primer (15 µM):	0.5 µl
10x PCR buffer:	2.5 µl
dNTPs (2.5 mM each)	0.5 µl
DNA polymerase mix:	0.5 µl
Ligation reaction:	100 ng
Nuclease-free water:	up to 25 µl

Table 2: Recipe for First Round PCR. The amount of each specified reagent to be added to each individual PCR tube is indicated.

Reagent	To Add per Reaction
Second Round LTR primer (15 µM):	2.5 µl
Linker-specific primer (15 µM):	0.5 µl
10x PCR buffer:	2.5 µl
dNTPs (2.5 mM each)	0.5 µl
DNA polymerase mix:	0.5 µl
First round PCR:	100 ng
Nuclease-free water:	up to 25 µl

Table 3: Second Round PCR Recipe. The amount of each reagent to be added to each PCR tube is indicated.

Library	#Unique Sites	%RefSeq[a]	%CpG +/- 2.5 kb[b]	%TSS +/- 2.5 kb[c]	Avg. Gene Density +/- 500 kb[d]
Sonication, neat	3,169	71.2	5.1	3.7	15.8
Sonication, 1:5	366	75.1	2.7	3	16.3
Sonication, 1:25	254	74	7.1	5.1	16.7
Sonication, 1:125	430	69.8	6.9	6	14.6
Sonication, 1:625	314	65.6	5.6	6.7	13.5
Sonication, 1:3,125	116	73.6	3.5	2.5	13.1
Sonication, 1:15,625	72	62.5	0	1.4	14.7
Digest, neat	7,428	69.8	3.6	2.9	15.2
Digest, 1:5	1,460	71.4	4.4	3.4	14.9
Digest, 1:25	394	68.8	4.3	3.3	15.8
Digest, 1:125	172	71	0	3	14
Digest, 1:625	134	73.9	3.7	3.7	14.1
Digest, 1:3,125	100	83.1	6.4	5.2	19.1
Digest, 1:15,625	73	74	4.1	1.4	9.7
MRC_{enz}[e]	50,000	44.7	4.2	4	8.7
MRC_{random}[f]	10,000	41.3	5.3	4.2	8.6

Table 4: Genomic Distribution of Integration Sites from Representative Titration Series. The percentage of total integration sites that fall within [a]RefSeq genes, [b]within 2.5 kb of CpG islands, and [c]within 2.5 kb of TSSs. [d]The gene density within 1 Mb surrounding the average integration site. [e]MRC_{enz}: matched random control; a set of 50,000 unique integration sites was produced by randomly selecting positions in proximity to MseI/BglII restriction sites in hg19. [f]MRC_{random}: matched random control containing 10,000 unique integration sites produced by randomly selecting positions in hg19 without normalization to fixed positions.

Discussion

A protocol for the analysis of retroviral integration sites, from the initial virus infection step through mapping of genomic distribution patterns, is described. This protocol is applicable to any retrovirus and any infectable cell type. Furthermore, the assay pipeline is quite sensitive, with the potential to recover a satisfactory number of unique integration sites from serial dilutions of genomic DNA equivalent to that of an infection initiated with an MOI of 6.4 x 10^{-5}. This sensitivity makes the protocol especially useful when applied to samples from infected patients that may contain a low viral load, where only a small fraction of cells will harbor an integrated provirus. Consistent with prior methodology papers in this field [36,38,41-43], multiple steps in the bioinformatics portion of this protocol will benefit from the development of customized scripts for processing large files of sequence data. While BLAT [58] is the mapping utility described in this protocol, users may find Bowtie [67] (http://bowtie-bio.sourceforge.net/index.shtml) to be a suitable alternative.

An alternative bioinformatics pipeline was recently reported for determination of Moloney murine leukemia virus (MoMLV) integration sites [19]. That pipeline is useful in that it was developed into standalone software that is publicly available, and is quite powerful in that it was originally used to map hundreds of thousands of unique MoMLV integration sites. However, the available software was originally designed to specifically re-analyze the reported MoMLV dataset, and so reprogramming would be necessary to customize the pipeline to alternate experimental designs (the functionality of the tool was recently expanded to include adeno-associated virus and Tol2 and Ac/Ds transposon vectors [68]). Furthermore, that protocol described the generation of the preliminary integration site .bed file, but did not lay out specific steps necessary to map sites to pertinent genomic annotations. Readers may find the "Vector Integration Site Analysis" server [69], which was released during the review of the current manuscript, useful to analyze the NGS sequences generated using the protocol described here.

Certain points should be emphasized when using any protocol to analyze retroviral integration site datasets. When preparing multiple libraries in tandem, a significant potential exists for sample cross-contamination. Even a very small level of sample crosstalk can obscure results to the level of rendering a NGS run unusable. Therefore, all wet-bench work should be completed in a sterilized, dedicated laminar flow hood or PCR workstation. A set of pipettes and reagents such as nuclease-free water should be dedicated solely to integration site amplification. The use of unique linkers for each library preparation can limit the potential for cross-amplification and also allow for identification of crossover reads within each library in the raw FASTA files.

It is important to consider the pros and cons of using sonication versus restriction endonuclease digestion to fragment genomic DNA. On the one hand, sonication provides a relatively random distribution of shear points, but the subsequently required DNA repair and A-tailing steps consistently reduce the yield of linker ligation products as compared to ligations performed with restriction enzyme-generated sticky ends. On the other hand, restriction enzyme digestion provides a less-disbursed population of shear points, which will invariably introduce some bias in the recovered data. Utilizing a restriction endonuclease to discard upstream LTR sequences will in both cases (**Figure 1**) result in the loss of a small fraction of integration sites that lie upstream of that site in the genome. Any data bias that may result can be addressed by omitting the enzymatic

digestion from the protocol during library preparation and filtering out the multitude of resulting upstream LTR sequences from the sequencing data.

Though the current protocol is quite sensitive and capable of generating millions of unique integration sites [21,40], only about one-third of all available integrations might be expected to be amplified in a given experiment even with the best of library preparations (ref. [70] and unpublished observations). This can cause complications when analyzing samples from low MOI infections or patients that harbor low viral load. This limitation can be overcome in part by repeatedly sequencing the same library preparation and/or sequencing multiple libraries derived from the same DNA sample in parallel. Future increases in assay sensitivity will accordingly be very beneficial to the furthering translational applications of retroviral integration site sequencing.

Disclosures

The authors have nothing to disclose.

Acknowledgements

We are grateful to our colleagues Stephen Hughes and Henry Levin for advice that was critical to establish the NGS protocol for retroviral integration site sequencing in the Engelman lab. This work was supported by US National Institutes of Health grants AI039394 and AI052014 (to A.N.E.) and AI060354 (Harvard University Center for AIDS Research).

References

1. Craigie, R., & Bushman, F. D. HIV DNA integration. *Cold Spring Harb. Perspect. Med.* **2**, a006890 (2012).
2. Li, M., Mizuuchi, M., Burke, T. R. J., & Craigie, R. Retroviral DNA integration: reaction pathway and critical intermediates. *EMBO J.* **25**, 1295-1304 (2006).
3. Hare, S., Gupta, S. S., Valkov, E., Engelman, A., & Cherepanov, P. Retroviral intasome assembly and inhibition of DNA strand transfer. *Nature*. **464**, 232-236 (2010).
4. Hare, S., Maertens, G. N., & Cherepanov, P. 3'-processing and strand transfer catalysed by retroviral integrase in crystallo. *EMBO J.* **31**, 3020-3028 (2012).
5. Fujiwara, T., & Mizuuchi, K. Retroviral DNA integration: structure of an integration intermediate. *Cell*. **54**, 497-504 (1988).
6. Roth, M. J., Schwartzberg, P. L., & Goff, S. P. Structure of the termini of DNA intermediates in the integration of retroviral DNA: dependence on IN function and terminal DNA sequence. *Cell*. **58**, 47-54 (1989).
7. Brown, P. O., Bowerman, B., Varmus, H. E., & Bishop, J. M. Retroviral integration: structure of the initial covalent product and its precursor, and a role for the viral IN protein. *Proc. Natl. Acad. Sci. USA*. **86**, 2525-2529 (1989).
8. Pauza, C. D. Two bases are deleted from the termini of HIV-1 linear DNA during integrative recombination. *Virology*. **179**, 886-889 (1990).
9. Bowerman, B., Brown, P. O., Bishop, J. M., & Varmus, H. E. A nucleoprotein complex mediates the integration of retroviral DNA. *Genes Dev.* **3**, 469-478 (1989).
10. Bukrinsky, M. I. *et al.* Active nuclear import of human immunodeficiency virus type 1 preintegration complexes. *Proc. Natl. Acad. Sci. USA* **89**, 6580-6584 (1992).
11. Miller, M. D., Farnet, C. M., & Bushman, F. D. Human immunodeficiency virus type 1 preintegration complexes: studies of organization and composition. *J. Virol.* **71**, 5382-5390 (1997).
12. Engelman, A., Mizuuchi, K., & Craigie, R. HIV-1 DNA integration: mechanism of viral DNA cleavage and DNA strand transfer. *Cell*. **67**, 1211-1221 (1991).
13. Maertens, G. N., Hare, S., & Cherepanov, P. The mechanism of retroviral integration from X-ray structures of its key intermediates. *Nature*. **468**, 326-329 (2010).
14. 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).
15. Wu, X., Li, Y., Crise, B., Burgess, S. M., & Munroe, D. J. Weak palindromic consensus sequences are a common feature found at the integration target sites of many retroviruses. *J. Virol.* **79**, 5211-5214 (2005).
16. Kvaratskhelia, M., Sharma, A., Larue, R. C., Serrao, E., & Engelman, A. Molecular mechanisms of retroviral integration site selection. *Nucleic Acids Res.* **42**, 10209-10225 (2014).
17. Schroder, A. R. *et al.* HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110**, 521-529 (2002).
18. 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).
19. LaFave, M. C. *et al.* MLV integration site selection is driven by strong enhancers and active promoters. *Nucleic Acids Res.* **42**, 4257-4269 (2014).
20. De Ravin, S. S. *et al.* Enhancers are major targets for murine leukemia virus vector integration. *J. Virol.* **88**, 4504-4513 (2014).
21. Maskell, D. P. *et al.* Structural basis for retroviral integration into nucleosomes. *Nature* **523**, 366-369 (2015).
22. Serrao, E. *et al.* Integrase residues that determine nucleotide preferences at sites of HIV-1 integration: implications for the mechanism of target DNA binding. *Nucleic Acids Res.* **42**, 5164-5176 (2014).
23. Aiyer, S. *et al.* Structural and sequencing analysis of local target DNA recognition by MLV integrase. *Nucleic Acids Res.* **43**, 5647-5663 (2015).
24. Ciuffi, A. *et al.* A role for LEDGF/p75 in targeting HIV DNA integration. *Nat. Med.* **11**, 1287-1289 (2005).
25. Sharma, A. *et al.* BET proteins promote efficient murine leukemia virus integration at transcription start sites. *Proc. Natl. Acad. Sci. USA* **110**, 12036-12041 (2013).
26. Gupta, S. S. *et al.* Bromo- and extraterminal domain chromatin regulators serve as cofactors for murine leukemia virus integration. *J. Virol.* **87**, 12721-12736 (2013).

27. De Rijck, J. *et al.* The BET family of proteins targets moloney murine leukemia virus integration near transcription start sites. *Cell Rep.* **5**, 886-894 (2013).
28. Demeulemeester, J. *et al.* HIV-1 integrase variants retarget viral integration and are associated with disease progression in a chronic infection cohort. *Cell Host Microbe* **16**, 651-662 (2014).
29. Meehan, A. M. *et al.* LEDGF/p75 proteins with alternative chromatin tethers are functional HIV-1 cofactors. *PLoS Pathog.* **5**, e1000522 (2009).
30. Ferris, A. L. *et al.* Lens epithelium-derived growth factor fusion proteins redirect HIV-1 DNA integration. *Proc. Natl. Acad. Sci. USA* **107**, 3135-3140 (2010).
31. Gijsbers, R. *et al.* LEDGF hybrids efficiently retarget lentiviral integration into heterochromatin. *Mol. Ther.* **18**, 552-560 (2010).
32. Aiyer, S. *et al.* Altering murine leukemia virus integration through disruption of the integrase and BET protein family interaction. *Nucleic Acids Res.* **42**, 5917-5928 (2014).
33. Jahner, D., & Jaenisch, R. Integration of Moloney leukaemia virus into the germ line of mice: correlation between site of integration and virus activation. *Nature*. **287**, 456-458 (1980).
34. Stevens, S. W., & Griffith, J. D. Human immunodeficiency virus type 1 may preferentially integrate into chromatin occupied by L1Hs repetitive elements. *Proc. Natl. Acad. Sci. USA*. **91**, 5557-5561 (1994).
35. Wang, G. P., Ciuffi, A., Leipzig, J., Berry, C. C., & Bushman, F. D. HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Res.* **17**, 1186-1194 (2007).
36. Wang, G. P. *et al.* DNA bar coding and pyrosequencing to analyze adverse events in therapeutic gene transfer. *Nucleic Acids Res.* **36**, e49 (2008).
37. Roth, S. L., Malani, N., & Bushman, F. D. Gammaretroviral integration into nucleosomal target DNA in vivo. *J. Virol.* **85**, 7393-7401 (2011).
38. Ciuffi, A., & Barr, S. D. Identification of HIV integration sites in infected host genomic DNA. *Methods*. **53**, 39-46 (2011).
39. Gillet, N. A. *et al.* The host genomic environment of the provirus determines the abundance of HTLV-1-infected T-cell clones. *Blood* **117**, 3113-3122 (2011).
40. Matreyek, K. A. *et al.* Host and viral determinants for MxB restriction of HIV-1 infection. *Retrovirology* **11**, 90 (2014).
41. Ciuffi, A. *et al.* Methods for integration site distribution analyses in animal cell genomes. *Methods* **47**, 261-268 (2009).
42. Brady, T. *et al.* A method to sequence and quantify DNA integration for monitoring outcome in gene therapy. *Nucleic Acids Res.* **39**, e72 (2011).
43. Beard, B. C., Adair, J. E., Trobridge, G. D., & Kiem, H. P. High-throughput genomic mapping of vector integration sites in gene therapy studies. *Methods Mol. Biol.* **1185**, 321-344 (2014).
44. Page, K. A., Landau, N. R., & Littman, D. R. Construction and use of a human immunodeficiency virus vector for analysis of virus infectivity. *J. Virol.* **64**, 5270-5276 (1990).
45. Emi, N., Friedmann, T., & Yee, J. K. Pseudotype formation of murine leukemia virus with the G protein of vesicular stomatitis virus. *J. Virol.* **65**, 1202-1207 (1991).
46. Wehrly, K., & Chesebro, B. p24 antigen capture assay for quantification of human immunodeficiency virus using readily available inexpensive reagents. *Methods*. **12**, 288-293 (1997).
47. Goff, S., Traktman, P., & Baltimore, D. Isolation and properties of Moloney murine leukemia virus mutants: use of a rapid assay for release of virion reverse transcriptase. *J. Virol.* **38**, 239-248 (1981).
48. Willey, R. L. *et al.* In vitro mutagenesis identifies a region within the envelope gene of the human immunodeficiency virus that is critical for infectivity. *J. Virol.* **62**, 139-147 (1988).
49. Butler, S. L., Hansen, M. S., & Bushman, F. D. A quantitative assay for HIV DNA integration in vivo. *Nat. Med.* **7**, 631-634 (2001).
50. Brussel, A., & Sonigo, P. Analysis of early human immunodeficiency virus type 1 DNA synthesis by use of a new sensitive assay for quantifying integrated provirus. *J. Virol.* **77**, 10119-10124 (2003).
51. Serrao, E., Ballandras-Colas, A., Cherepanov, P., Maertens, G. N., & Engelman, A. N. Key determinants of target DNA recognition by retroviral intasomes. *Retrovirology*. **12**, 39 (2015).
52. Maldarelli, F. *et al.* HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science* **345**, 179-183 (2014).
53. Wagner, T. A. *et al.* HIV latency. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. *Science* **345**, 570-573 (2014).
54. Cohn, L. B. *et al.* HIV-1 integration landscape during latent and active infection. *Cell* **160**, 420-432 (2015).
55. Li, X., Ben-Dov, I. Z., Mauro, M., & Williams, Z. Lowering the quantification limit of the QubitTM RNA HS assay using RNA spike-in. *BMC Mol. Biol.* **16**, 9 (2015).
56. Padmanaban, A., & Walker, D. M. *Analysis of high molecular weight genomic DNA using the Agilent 2200 TapeStation and genomic DNA ScreenTape.* Publication number 5991-1797EN Agilent Technologies, Inc., Santa Clara, CA, (2013).
57. [no authors listed]. *Kapa library quantification technical guide version v1.14.* KapaBiosystems, Boston, MA, (2014).
58. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res.* **12**, 656-664 (2002).
59. Quinlan, A. R., & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. **26**, 841-842 (2010).
60. Gaur, M., & Leavitt, A. D. Mutations in the human immunodeficiency virus type 1 integrase D,D(35)E motif do not eliminate provirus formation. *J. Virol.* **72**, 4678-4685 (1998).
61. Varadarajan, J., McWilliams, M. J., & Hughes, S. H. Treatment with suboptimal doses of raltegravir leads to aberrant HIV-1 integrations. *Proc. Natl. Acad. Sci. USA*. **110**, 14747-14752 (2013).
62. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996-1006 (2002).
63. Berry, C. C. *et al.* Estimating abundances of retroviral insertion sites from DNA fragment length data. *Bioinformatics* **28**, 755-762 (2012).
64. Firouzi, S. *et al.* Development and validation of a new high-throughput method to investigate the clonality of HTLV-1-infected cells based on provirus integration sites. *Genomic Med.* **6**, 46 (2014).
65. Mitchell, R. S. *et al.* Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol.* **2**, E234 (2004).
66. Schneider, T. D., Stormo, G. D., Gold, L., & Ehrenfeucht, A. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* **188**, 415-431 (1986).
67. Langmead, B. Aligning sort sequencing reads with Bowtie. *Curr. Protoc. Bioinformatics*. **Chapter 11**, Unit 11.17 (2010).
68. LaFave, M. C., Varshney, G. K., & Burgess, S. M. GeIST: a pipeline for mapping integrated DNA elements. *Bioinformatics.* **31**, 3219-3221 (2015).

69. Hocum, J. D. *et al.* VISA - Vector Integration Site Analysis server: a web-based server to rapidly identify retroviral integration sites from next-generation sequencing. *BMC Bioinformatics* **16**, 212 (2015).
70. Gabriel, R. *et al.* Comprehensive genomic access to vector integration in clinical gene therapy. *Nat. Med.* **15**, 1431-1436 (2009).
71. *TruSeq Library Prep Pooling Guide. Guidelines for pooling TruSeq libraries for Illumina sequencing systems that require balanced index combinations.* Source: https://support.illumina.com/downloads/truseq-library-prep-pooling-guide-15042173.html (2015).

