Journal of Virological Methods 267 (2019) 59–65

Contents lists available at ScienceDirect

Journal of Virological Methods

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

Identifying integration sites of the HIV-1 genome with intact and aberrant ends through deep sequencing

Hirotaka Ode[a,*], Ayumi Kobayashi[a,b], Masakazu Matsuda[a], Atsuko Hachiya[a], Mayumi Imahashi[a], Yoshiyuki Yokomaku[a], Yasumasa Iwatani[a,b]

[a] *Clinical Research Center, National Hospital Organization Nagoya Medical Center, Nagoya, Aichi, 460-0001, Japan*

[b] *Program in Integrated Molecular Medicine, Nagoya University Graduate School of Medicine, Nagoya, Aichi, 466-8550, Japan*

ARTICLE INFO

Keywords:
HIV-1
Integration site
Clonal expansion
Deep sequencing
Bioinformatics

ABSTRACT

Paired-end deep sequencing is a powerful tool to investigate integration sites of the HIV-1 genome in infected cells. Integration sites of HIV-1 proviral DNA carrying intact LTR ends have been well documented. In contrast, integration sites of proviral DNA with aberrant ends, which emerge infrequently but can also induce replication-competent viruses, have not been extensively examined, in part, because of the lack of a suitable bioinformatics method for deep sequencing. Here, we report a novel bioinformatics protocol, named the VINSSRM, to search for integration sites of proviral DNA carrying intact and aberrant LTR ends using paired-end deep sequencing data. The protocol incorporates split-read mapping to assign viral and human genome parts within read sequences and overlapping paired-end read merging to construct long error-corrected sequences. The VINSSRM not only consistently detects integration sites similar to the conventional method but also provides information on additional integration sites, including those of proviral DNA with aberrant ends, which were mainly found in non-exonic regions of the human genome. Therefore, the VINSSRM may help us to understand HIV-1 integration, persistence of infected cells, and viral latency.

1. Introduction

Similar to other retroviruses, human immunodeficiency virus type 1 (HIV-1) integrates its proviral DNA (vDNA) into the host genome. The integration is critical for HIV-1 replication. HIV-1 integrase, which is encoded within the HIV-1 genome, interacts with the long terminal repeat (LTR) U5 and U3 ends of reverse-transcribed vDNA and then catalyzes joining of these ends with human chromosomes (Esposito and Craigie, 1998; Gao et al., 2001; Yoshinaga and Fujiwara, 1995; Yoshinaga et al., 1994). The authentic vDNA ends flanking the human genome sequences consist of dinucleotide sequence motifs of cytosine-adenine (CA_{OH}). In contrast, integrase-independent integration has also been reported (Ebina et al., 2012; Gaur and Leavitt, 1998). Although infrequent, this type of integration is enhanced by DNA damage and tends to result in extensions or deletions at the vDNA terminal sequences (Supplementary Fig. S1).

Integration sites of vDNA are not random but are integrated throughout the host genomes (Mitchell et al., 2004). HIV-1 preferentially integrates vDNA into intronic regions of actively transcribed genes, as well as regions within or nearby *Alu* elements throughout the human genome (Achuthan et al., 2018; Lusic and Siliciano, 2017; Pinzone and O'Doherty, 2018; Schroder et al., 2002; Sowd et al., 2016). Therefore, integration sites of vDNA are signatures of retroviral infections and have been applied as a marker to evaluate the persistence and clonal expansion of infected cells (Maldarelli et al., 2014; Wagner et al., 2014).

To date, to identify HIV-1 integration sites, linker-mediated PCR of DNA fragments around junctions between the human genome and vDNA, followed by paired-end deep sequencing of their amplified DNA have generally been performed (Cohn et al., 2015; Maldarelli et al., 2014; Satou et al., 2017). From each fragment, two paired read sequences are outputted. One read (read1) and its counterpart (read2) contain the LTR sequence (U5 or U3) and a linker sequence at the 5´-ends, respectively. Subsequent bioinformatics analysis of these read sequences provides information about the integration sites and clonality of infected cells (Supplementary Fig. S1C). However, the previously reported conventional bioinformatics method (CBM) focuses only on intact vDNA ends (Maldarelli et al., 2014). Therefore,

Abbreviations: HIV-1, human immunodeficiency virus type 1; vDNA, proviral DNA; CBM, conventional bioinformatics method; VINSSRM, Viral INtegration site Search protocol based on Split-Read mapping and Merging algorithms

[*] Corresponding author.

E-mail address: hirotaka.ode@nnh.go.jp (H. Ode).

https://doi.org/10.1016/j.jviromet.2019.03.004
Received 5 December 2018; Received in revised form 5 February 2019; Accepted 8 March 2019
Available online 08 March 2019

integration sites of vDNA with aberrant ends are excluded from the deep sequencing analyses, even though integrated vDNA carrying aberrant ends can induce replication-competent viruses (Ebina et al., 2012). In this study, to investigate the integration sites of vDNA with both intact and aberrant ends, we developed a novel bioinformatics protocol.

2. Materials and methods

2.1. *In vitro culture of HIV-1–infected CD4$^+$ T cells*

CD4$^+$ T cells were negatively selected from blood of a healthy donor with the CD4$^+$ Isolation Kit (Miltenyi Biotec K.K., Tokyo, Japan). The cells were resuspended at 0.4 × 10^6 cells/mL in RPMI-1640 media (100 U/mL penicillin, 100 μg/mL streptomycin, 10% fetal bovine serum (FBS), 100 mM sodium pyruvate, 10 mM 2-mercaptoethanol). Cells in 2 mL of media were activated with 60 U/mL interleukin-2 (IL-2) and 12.5 μL of Dynabeads T-Activator CD3/CD28 (Thermo Fisher Scientific, Waltham, MA, USA) for 72 h in a CO_2 incubator (37 °C, CO_2 5%). The activated cells were resuspended with fresh media at 0.4 × 10^6 cells/mL. After addition of HIV-1_{JRCSF} stocks (100 ng p24) (GenBank accession No. U45960) (Koyanagi et al., 1987) into 2 mL of media, cells were incubated for 72 h in a CO_2 incubator (37 °C, CO_2 5%). The viruses were obtained from transfection of 24 μg of its infectious clone into HEK293 T cells (15 mL, 5 × 10^5 cells/mL) with FuGENE HD Transfection Reagent (Promega, Madison, WI, USA) followed by incubation for 48 h in a CO_2 incubator.

2.2. *Deep sequencing*

DNA was isolated from the cultured CD4$^+$ T cells with the QIAamp DNA Blood Mini Kit (QIAGEN K.K., Tokyo, Japan). The library was prepared from 30 μg of the extracted DNA according to the protocol in a previous report with slight modifications (Maldarelli et al., 2014). Briefly, the extracted DNA was sheared into fragments of 300–500 bp using an M220 focused-ultrasonicator (Covaris, Woburn, MA, USA). The sheared DNA fragments were subjected to end-repair with the End-it DNA End Repair-Kit (Epicentre, Madison, WI, USA) and addition of a single dA to the 3´-end with the dA-Tailing Kit (New England BioLabs, Ipswich, MA, USA). A partially double-stranded linker with a one-nucleotide 3´-T overhang was ligated to the DNA fragments. The flanking regions of the junctions between the human genome and each viral genome end (5´- or 3´-end) were selectively amplified with linker-mediated PCR using one outer-primer specific for the HIV-1 LTR sequence and another outer-primer specific for a single-stranded portion of the linker. Then, nested PCR was performed with an inner-primer pair specific for the other LTR and linker sequences. To attach the index and adaptor sequences onto the amplified fragments, index PCR was performed with the Nextera XT Index Kit (Illumina K.K., Tokyo, Japan). Of note, two different specific adaptors were attached to the LTR and liker sides of the amplified fragments for read1 and read2, respectively.

Paired-end 150 × 2-bp or 250 × 2-bp runs were performed for the pooled indexed fragments for HIV-1 5´- and 3´-ends with Illumina MiSeq (Illumina K.K.). The primer and linker sequences used in this study are listed in Supplementary Table S1.

2.3. *The VINSSRM, viral integration site search protocol based on split-read mapping and merging algorithms*

The workflow to identify integration sites is summarized in Fig. 1. Paired-end reads, of which read1 and read2 contain the LTR and linker sequences at their 5´-ends, were used for integration site analyses. First, each set of overlapping paired-end read sequences was merged using a merging program, PEAR (Paired-End reAd mergeR) v.0.9.6 (Zhang et al., 2014), with an option of "-m 300". The merging process potentially yields longer and higher-quality reads (read3) than the original reads. Then, to ensure the sequence quality, the presence of the identical LTR primer sequences at the 5´-ends was checked for read1 and read3. Similarly, we confirmed the presence of the identical linker sequences at the 5´-ends within read2 and at the 3´-ends of read3. Read sequences with an average quality score of < 20 were omitted, which is the same as in the CBM (Maldarelli et al., 2014). The linker sequences at the 5´- and 3´-ends were trimmed within read2 and read3, respectively. The read2 and read3 without the linker sequences are defined as read2t and read3t. Next, to identify the human genome and viral genome parts within each read sequence, split-read mapping (Pirooznia et al., 2015; Tattini et al., 2015; Zhao et al., 2013) of paired-end reads (read1 and read2t) and single-end reads (read3t) onto the human genome (UCSC hg19) and multiple viral reference genome sequences was performed by the BWA-MEM algorithm in the BWA (Burrows-Wheeler Aligner) program v.0.7.15-r1140 (Li, 2013). The "-C" and "-Y" options were used to append comments in FASTQ files to SAM outputs and to force soft-clipping for mappings. We used HIV-1 subtype reference genome sequences deposited in the Los Alamos HIV sequence database (downloaded from https://www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html) as reference sequences for mapping. For mapping read sequences for analyses of HIV-1_{JRCSF}–infected cells, we also used the sequence of HIV-1_{JRCSF} (U45960) as a reference sequence. Similar to the CBM (Maldarelli et al., 2014), mapped read sequences that met the following criteria were selected: 1) The whole read sequence (read1, read2t, or read3t) was mapped onto either viral reference sequences or human genome sequences, except for junctions between the LTR parts and human genome parts and for the 3´-end portions of read2t. However, the LTR part and the human genome part within each read1 and read3t were not separated (≤ 3 bp). 2) Human genome parts in each read sequence (read1, read2t, and read3t) were mapped onto a single region of the human genome. 3) In the case of paired-end reads, read1 and read2t were mapped within 1 kbp of the human genome sequences. Integration sites were defined as the human genome positions corresponding to 5´-end nucleotides of the human genome parts in read1 or read3t (red arrow in Fig. 1). In this study, we further classified the integration sites into two classes: integration sites of vDNA with intact or with aberrant viral ends (Supplementary Fig. S1AB). Integration sites estimated from read sequences of which the LTR parts were of similar length (≤ 3 bp) to the expected LTR sequence were regarded as the sites of vDNA with intact viral ends, while other integration sites were considered as the sites of vDNA with aberrant viral ends. The genetic characteristics of the identified integration sites were examined with genomic information in the UCSC Genome Browser (https://genome.ucsc.edu/).

In addition, to evaluate the clonal expansion of infected cells, we defined breakpoints as being the human genome positions corresponding to 5´-end nucleotides of the human genome parts in read2t or the 3´-end nucleotides of those in read3t (black arrow in Fig. 1). When the same set of read sequences containing the same integration sites but distinct breakpoints were found, we deemed this to be clonal expansion of the infected cells with the integration. Of note, only breakpoints that were > 3 bp apart were counted as independent events to avoid any potential errors that might have arisen when determining the breakpoints (Maldarelli et al., 2014).

The processes, except for merging of overlapping paired-end read sequences and read mapping, were achieved with in-house PERL scripts (available at https://github.com/odehir/VINSSRM).

2.4. *The conventional bioinformatics method (CBM) for integration site analyses*

Integration sites were searched as in the previous report (Supplementary Fig. S1C) (Maldarelli et al., 2014). In brief, read sequences were filtered based on sequence identity with the expected intact LTR end and linker sequences. Then, sequences of linkers and LTR ends were trimmed and mapped to human genome hg19 with the

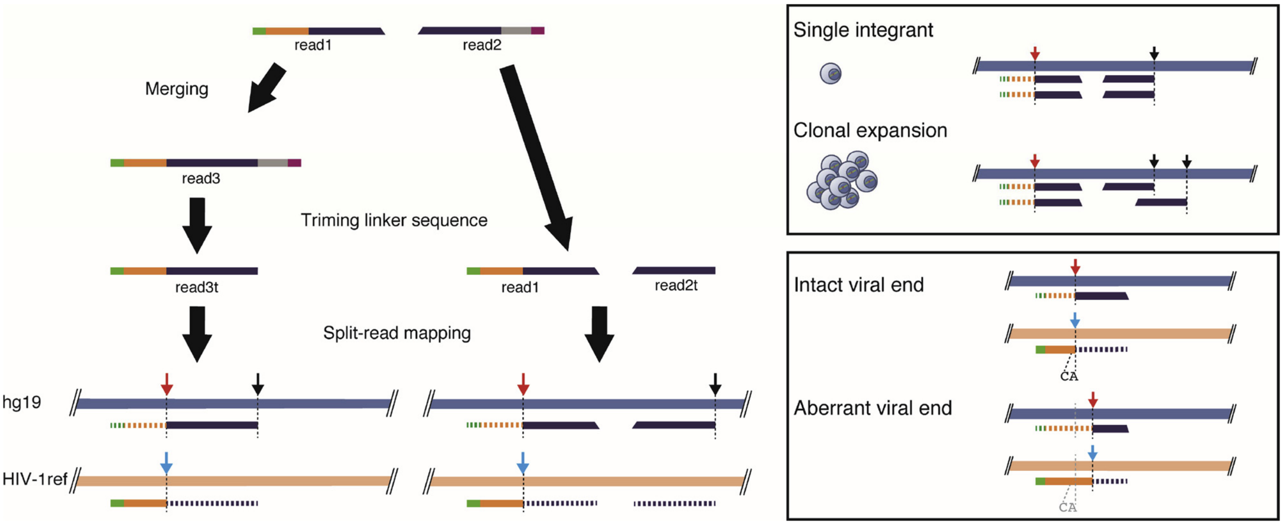

Fig. 1. The workflow of the integration site search in the VINSSRM. The paired-end read sequences to examine integration sites contain viral LTR portions (orange), human genome portions (navy), and linker portions (gray). The primer sequences for the LTR and linker are found in each 5´-end of the original paired-end read sequences (read1 and read2; light green and purple, respectively). Each set of overlapping paired-end read sequences was merged into one read (read3). Then, the merged read sequences and paired-end read sequences were subjected to trimming of linker sequences and subsequent split-read mapping onto the human genome (hg19; light blue) and HIV-1 reference sequences (HIV-1ref; beige). The mapped and unaligned regions are indicated in solid and dotted lines, respectively. The integration sites (red arrow) and breakpoints (black arrow) were assigned by nucleobase positions onto the human genome. The breakpoints were used to evaluate whether the integration sites are identified from clonally expanded cells (right-top panel). The LTR ends (cyan arrow) were assigned by nucleobase positions onto HIV-1 reference genomes. The nucleobase positions of the LTR ends are used to evaluate whether the viral end is intact or aberrant (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

BLAT program (BLAST-like alignment tool) (Kent, 2002). Read sequences that met the following criteria were further selected: 1) human genomic portions within read sequences had a > 95% match to hg19 sequences; 2) the human genome part of read1 was > 20 bp long, and an average quality score of bases within them exceeded 20; 3) the human genome portion started within 3 bp of the LTR junction in read1; 4) read1 and read2 were mapped within 1 kbp of the hg19 sequences.

2.5. Sequencing data for integration site analyses

To test the VINSSRM, we used publically available sequencing data for HIV-1 integration sites of eight samples (Supplementary Table S2). The reads were previously analyzed by the CBM for integration site detection (Maldarelli et al., 2014). Of note, sequence data for seven samples (P1–P7) were derived from Illumina MiSeq runs, while data for another sample (P8) were obtained from the MiSeq and HiSeq runs. In addition, we analyzed the sequence data for HIV-1_{JRCSF}–infected cells in this study, which are available as DNA Data Bank of Japan (DDBJ) accession IDs: DRR159945 and DRR159946.

3. Results

3.1. Design of a novel integration site search protocol

To date, deep sequencing data from linker-mediated PCR amplicons have been analyzed to search for integration sites using the CBM (Supplementary Fig. S1C) (Cohn et al., 2015; Maldarelli et al., 2014; Satou et al., 2017). However, this method is not applicable for the detection of integration sites of vDNA with aberrant LTR ends. In addition, high error rates of deep sequencing compared to Sanger sequencing may prevent integration site detection. Indeed, sequence analyses of publically available Illumina 150 × 2 paired-end reads to identify HIV-1 integration sites of seven HIV-1–infected patients' samples (P1–P7) (Supplementary Table S2) (Maldarelli et al., 2014) indicated that the putative LTR parts of read sequences may contain mutations (Supplementary Fig. S2).

To overcome these issues, we designed a novel integration site search protocol, the VINSSRM (Fig. 1 and the Materials and methods). To recognize the human genome and viral LTR end parts within read sequences, we adopted split-read mapping (Pirooznia et al., 2015; Tattini et al., 2015; Zhao et al., 2013) in the VINSSRM, instead of trimming the intact LTR-end sequences from read sequences as in the CBM (Supplementary Fig. S1C) (Maldarelli et al., 2014). The mapping method, which was originally developed for the detection of structural variations within the human genome, enables us to align the human genome part and the LTR part in a read sequence separately onto a human genome sequence and a viral reference sequence, respectively. In addition, to reduce mapping failures due to sequencing errors and mutations, we used multiple HIV-1 reference genome sequences from various subtypes as a part of the reference sequences for the mapping. As we have previously reported, a higher sequence identity between read sequences and a reference genome results in greater mapping coverage (Ode et al., 2015). The multiple HIV-1 reference sequences may be helpful when we analyze clinical samples since HIV-1 LTR sequences exhibit inter-host sequence diversity (Supplementary Fig. S3). Furthermore, in contrast to the CBM, the VINSSRM incorporates an analysis of the merged read sequences (read3) that were generated from the overlapping paired-end read sequences. The merged read sequences potentially have higher quality than the original paired-end read sequences due to error correction during the merging process. Therefore, after we trimmed the linker sequences from read3 (read3t), we also use read3t to detect the integration sites. Here, we selected the PEAR program (Zhang et al., 2014) as the merging program for the VINSSRM based on performance tests among four programs (Supplementary Fig. S4).

3.2. Comparison of performance between the VINSSRM and the CBM

Next, we tested the performance of the VINSSRM for the detection of integration sites flanking intact viral ends and breakpoints and compared the results with those obtained by the CBM (Maldarelli et al., 2014). We examined the sequencing data of the seven patients' samples (P1–P7) (Supplementary Table S2) and counted the number of

Table 1
The number of combination patterns of detected integration sites and breakpoints from the VINSSRM and the CBM.

	P1	P2	P3	P4	P5	P6	P7	Total
Both methods	273	68	36	45	257	204	108	991
Only VINSSRM	308	55	61	44	220	253	192	1133
Only CBM	14	2	0	2	10	11	8	47

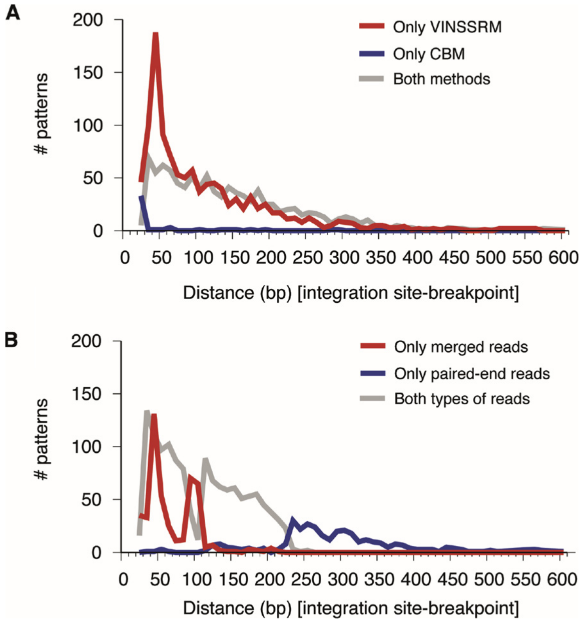

Fig. 2. The number of combination patterns of integration sites and breakpoints detected by the VINSSRM and the CBM. **A.** Performance comparison between the two methods. The numbers of patterns identified from *i)* both methods, *ii)* only the VINSSRM, and *iii)* only the CBM are shown with gray, red, and blue lines, respectively. **B.** The numbers of patterns from *i)* both types of merged and paired-end read sequences, *ii)* only merged read sequences, and *iii)* only paired-end read sequences in the VINSSRM are shown with gray, red, and blue lines, respectively (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

combination patterns of detected integration sites and breakpoints. Of note, for each integration site putatively in single cells or clonally expanded cells, one combination pattern or multiple patterns can be found (right-top panel of Fig. 1). For each sample, we could consistently detect patterns between the VINSSRM and the CBM but, in addition, the VINSSRM could detect unique extra patterns (Table 1). In total, the VINSSRM outputted 89–581 patterns for each sample, which is > 1.5-fold greater than that of the CBM. In particular, the VINSSRM outputted patterns that were assigned from short fragments (Fig. 2A). In contrast, we found only 0–14 unique patterns (47 sites in total) by the CBM. Almost of the sites (33/47 sites) were from fragments of human genome parts of 20–29 bp (Fig. 2A, blue line), which would be difficult to map onto the human genome. These results suggest that the VINSSRM has higher sensitivity to detect combination patterns of integration sites and breakpoints than the CBM.

Next, we examined how the merging process affected integration site detection (Fig. 2B). The sequencing data originated from 150 × 2 paired-end runs. Therefore, the merging process can generate less than ˜250-bp-long read sequences by excluding the LTR and linker sequence lengths. As expected, the sequence lengths of the human genome parts of the merged read sequences did not exceed 260 bp. In particular, detected patterns were accumulated from short read3t sequences of 35–49 bp from the human genome parts. Since the shorter fragments are preferentially amplified by PCR, more patterns were detected from the shorter read sequences. Besides, error corrections within the LTR primer and linker sequences during the merging process likely permitted trimming of the linker sequences from read sequences and subsequent detection of integration sites. In addition, combination patterns were uniquely detected from read3t with 90–109-bp-long human genome parts. The read3t was generated from paired-end reads carrying partial LTR and/or linker sequences at the 3´-end. When the original reads were analyzed, the partial sequences likely prevented proper mapping onto the human genome followed by detection of integration sites. Taken together, the merging process would contribute to the detection of integration sites from short read sequences.

We also examined whether the VINSSRM enhanced the detection of integration sites in single cells and/or clonally expanded cells for each sample (P1–P7). The VINSSRM promoted the detection of the sites in both single cells and clonally expanded cells (Fig. 3 and Supplementary Table S3A). Although some integration sites were considered to be from single cells only by the CBM (blue bars in Fig. 3), almost of them were evaluated as integration sites in clonally expanded cells by the VINSSRM. In total, > 1.5-fold higher numbers of integration sites were detected for each sample by the VINSSRM than the CBM. The results suggest that the VINSSRM is more sensitive to detect integration sites in clonally expanded cells compared to the CBM.

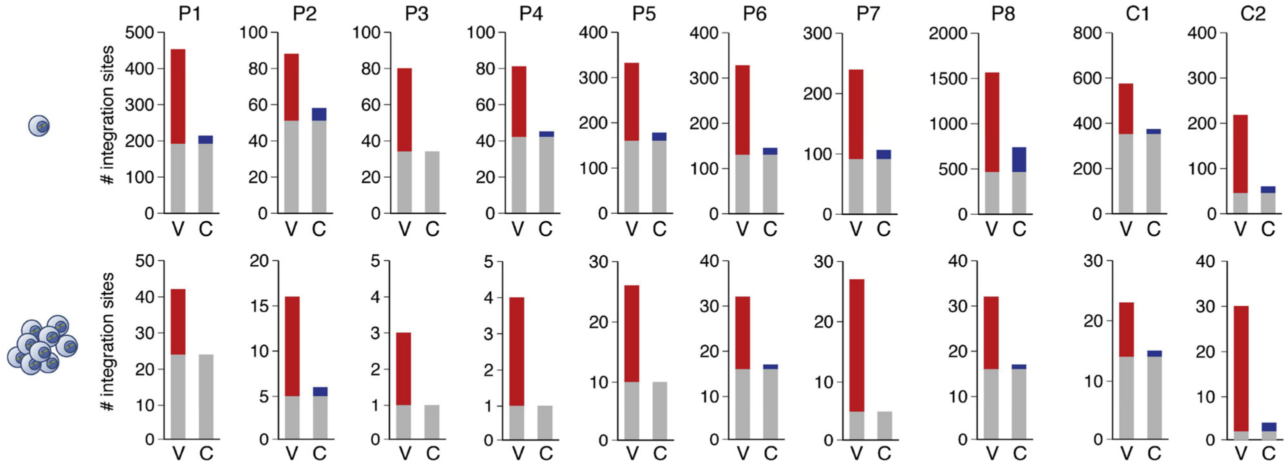

Fig. 3. The comparison of identified integration sites of vDNA carrying intact ends between the VINSSRM and the CBM. In each graph, the results of the VINSSRM and the CBM are shown with left and right bars (labeled “V” and “C”, respectively). Gray, red, and blue bars mean the number of integration sites detected by *i)* both methods, *ii)* only the VINSSRM, and *iii)* only by CBM. The top and bottom panels represent the results for integration sites in single cells and clonally expanded cells (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

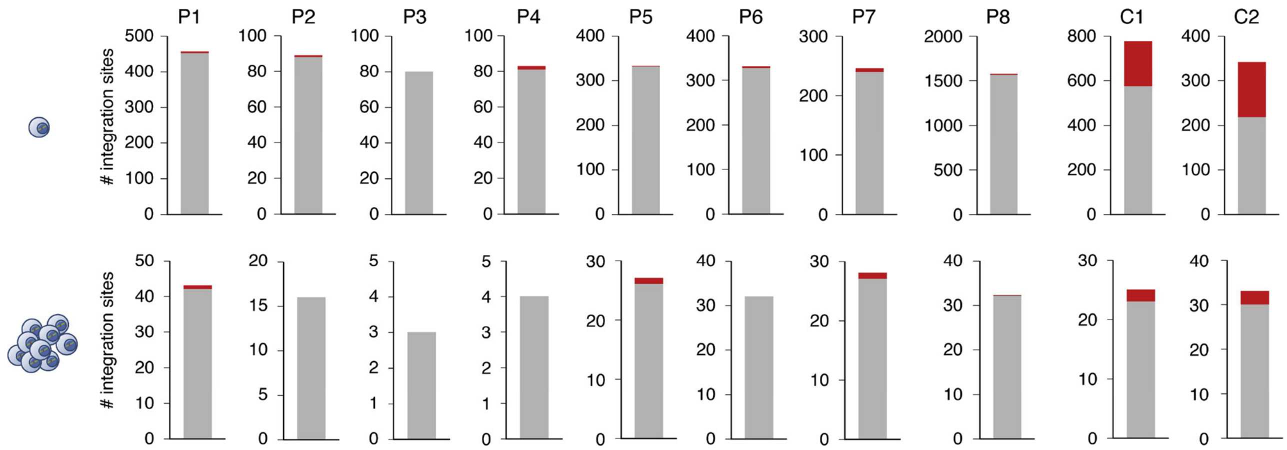

Fig. 4. The identified integration sites of vDNA carrying aberrant ends by the VINSSRM. Gray and red bars mean the number of integration sites of vDNA carrying intact viral ends and aberrant ends, respectively. The top and bottom panels represent the results for integration sites in single cells and clonally expanded cells (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

This high sensitivity was also observed when we analyzed sequencing data containing reads with a different length (100 × 2-bp run) for another sample (P8) (Fig. 3 and Supplementary Table S3A). We further explored the integration sites within another two sets of HIV-1-infected CD4$^+$ T cells cultured *in vitro* (C1 and C2). The Illumina MiSeq 150 × 2-bp run and 250 × 2-bp run were performed for C1 and C2, respectively. As observed in the integration site searches in patients' samples (P1–P8), the VINSSRM identified more integration sites for both C1 and C2 than the CBM (Fig. 3 and Supplementary Table S3A). Taken together, these results indicate that the VINSSRM is useful to detect integration sites, regardless of read lengths of sequencing runs.

3.3. Detection of integration sites of vDNA with aberrant LTR ends with the VINSSRM

Unlike the CBM (Maldarelli et al., 2014), the VINSSRM is designed theoretically to detect integration sites of vDNA with LTR terminal aberrant short extension or deletion sequences. When we searched for integration sites of vDNA with aberrant LTR ends for eight patients' samples (P1–P8), we identified 0–12 sites for each sample (Fig. 4 and Supplementary Table S3B). Among them, some sites were related to clonal expansion. Most of the aberrant viral ends identified in this study were 4–9-bp-long extensions (Supplementary Fig. S5). These vDNA sequences might result from the incomplete processing of 18-bp-long primer binding sites (PBSs) and 15-bp-long polypurine tracts (PPTs) by RNase H during vDNA synthesis and/or insufficient effects of DNA double-strand-break repair enzymes (Das and Berkhout, 2018; Sakurai et al., 2009). In contrast, we also found integration sites of vDNA with ≥ 20-bp-long extensions of viral ends (Supplementary Fig. S5). These vDNA sequences could be subjected to large deletions and only have the LTR sequences and the sequences of their flanking regions (Cohn et al., 2015; Hiener et al., 2017; Ho et al., 2013). More sites of vDNA with aberrant ends were found for *in vitro* cultures, C1 and C2. Furthermore, integration sites of vDNA with much longer extensions were found in the *in vitro* cultures. However, it remains unclear why a greater proportion of integration sites of vDNA with aberrant viral ends were found from CD4$^+$ T cells infected with HIV-1$_{\mathrm{JRCSF}}$ *in vitro*, in contrast to integration sites from patients' samples (P1–P8) and *in vitro* experiments in the previous reports based on Sanger sequencing of cloned DNA fragments (Varadarajan et al., 2013, 2016). A possible reason for the greater proportion might be the short culture of HIV-1–infected cells. Because much greater amount of integrase-independent integration of vDNA with aberrant ends appears to be found in DNA-damaged cells (Ebina et al., 2012), cells carrying integrated vDNA with aberrant ends might tend to be short-lived. Hence, longer cell culturing *in vitro* might reduce the integration-site ratio of vDNA with aberrant ends. Further analyses are required to clarify this possibility.

We also examined base preferences around these integration sites (Demeulemeester et al., 2014; Holman and Coffin, 2005) for the *in vitro* culture sample C1. The results showed that the base preferences for integration sites differed between vDNA with intact ends and with aberrant ends (Supplementary Fig. S6), which might support that vDNA with aberrant ends were preferentially integrated in an integrase-independent manner (Ebina et al., 2012).

3.4. Enrichment of integration sites within Alu elements with the VINSSRM

The above results suggest that the VINSSRM can provide more information on the integration sites of vDNA with intact ends, as well as aberrant ends, than the CBM (Maldarelli et al., 2014). Next, we sought to determine which genomic regions were enriched with integration sites by using the VINSSRM. When we annotated the gene structures of the identified integration sites, we found that the integration sites were especially enriched within non-exonic and non-intronic regions compared to intronic or exonic regions (Fig. 5A, Supplementary Fig. S7A and Table S3). In addition, vDNA with aberrant ends were found within non-exonic regions. Furthermore, when we focused on the repetitive sequences within the human genome, the number of detected integration sites was increased at *Alu* elements compared with other repetitive domains (Fig. 5B, Supplementary Fig. S7B). It is well known that the *Alu* elements are preferred sites for HIV-1 integration. In contrast, the VINSSRM did not indicate enriched integration sites in the regions nearby the *Alu* elements, although the regions are also known to be a preferred target of HIV-1 integration (Fig. 5C, Supplementary Fig. S7C).

4. Discussion

In this study, we propose a novel sensitive bioinformatics protocol, named the VINSSRM, to detect integration sites of vDNA using paired-end deep sequencing data. Deep sequencing is a powerful tool to identify integration sites of vDNA, as well as to evaluate whether integrated DNA is present in single cells or in clonally expanded cells. It has been reported that clonally expanded cells are likely associated with persistence, proliferation, and/or latency of HIV-1–infected cells (Bui et al., 2017; Maldarelli et al., 2014; Simonetti et al., 2016; Wagner et al., 2014). Therefore, information on integration sites may be a marker to understand the fates of infected cells, although we could not evaluate whether vDNA is replication-competent or -defective by the deep sequencing.

One of the key points of the VINSSRM is the application of split-read

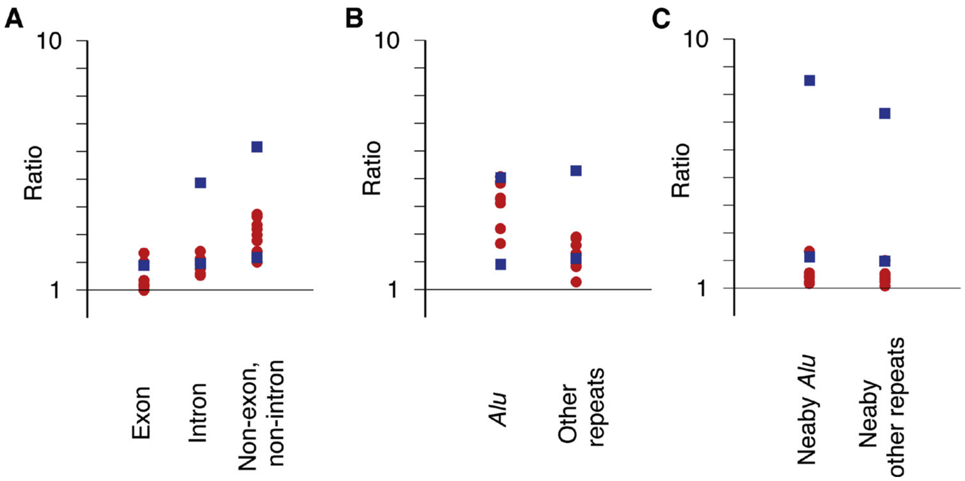

Fig. 5. Enrichment of detected integration sites by the VINSSRM. The ratio of the number of detected integration sites at specific genomic regions was calculated between the VINSSRM and the CBM. Red circles and blue squares represent the ratios for patients' samples (P1–P8) and for samples of *in vitro* infections (C1 and C2), respectively (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

mapping to recognize viral end portions within read sequences. The mapping method enables us to separately align two portions within a read sequence onto the human genome and the viral reference genome. Therefore, in contrast to the CBM (Maldarelli et al., 2014), this mapping procedure allows us to identify integration sites of vDNA with aberrant ends. In addition, although linker-mediated PCR might amplify human genome regions carrying LTR-end-like sequences (Supplementary Table S4), we can easily exclude these sequences from the integration site analyses. This mapping method would also be applicable to find chimeric sequences of human and viral genomes from a whole genome sequencing data of HIV-1 infected cells (Supplementary Table S5).

The second key point is the merging process of overlapping paired-end read sequences. As shown in Fig. 2, this merging process improved the detection of integration sites from short read sequences. Since smaller fragments are more easily amplified by PCR, deep sequencing produces many shorter read sequences. Hence, the merging process is beneficial for data analyses of deep sequencing that outputs short read sequences, such as the Illumina platforms.

Notably, to reduce the likelihood of false positives, we strictly confirmed the reliability of sequences in the reads and the mapping quality in the VINSSRM. The read sequences used for the integration site search fulfilled the following criteria. First, merged read sequences (read3t) were generated by the program *via* a statistical test to minimize the number of false-positive results (Zhang et al., 2014). Second, read sequences contain identical sequences to either the HIV-1 LTR primer sequence or the linker sequence at their ends. Third, after trimming of the linker sequences, the whole read sequences are mapped onto either the human genome or the viral reference sequences, except for short gaps (≤ 3 bp) between the human genome part and the viral genome part within the read sequences (read1 or read3t) and the 3´-end of read2t. Fourth, read sequences are mapped onto a unique region on the human genome.

In consequence, we could consistently identify integration sites that were common to both the VINSSRM and the CBM but we also found more integration sites by the VINSSRM than the CBM, despite the strict criteria. We detected many combination patterns of integration sites and breakpoints especially from read sequences carrying short human genome sequences (˜50 bp) by the VINSSRM (Fig. 2). In addition, the VINSSRM improved the detection of integration sites, particularly within *Alu* elements (Fig. 5 and Supplementary Fig. S7). Mapping of short read sequences is more difficult than that of long ones (Li and Durbin, 2010; Thankaswamy-Kosalai et al., 2017). Alignment of sequences onto the human genome repetitive regions is also challenging (Treangen and Salzberg, 2011), because the repetitive regions are found throughout the genome and *Alu* covers > 11% of the human genome (Batzer and Deininger, 2002; Wildschutte et al., 2015). Hence, the enhanced integration site detection might be partially due to improvements in the mapping programs (Li, 2013; Li and Durbin, 2010). Since the *Alu* elements are preferred targets of HIV-1 integration, the VINSSRM is beneficial for identifying integration sites of HIV-1 vDNA.

The VINSSRM also helps us to identify integration sites of vDNA with aberrant ends. These aberrant viral ends likely result from integrase-independent integrations (Ebina et al., 2012; Gaur and Leavitt, 1998). This type of integration is rare but can induce replication-competent viruses. Therefore, by application of the VINSSRM, we can detect uncharacterized integration sites in persistently or latently infected cells. In addition, aberrant viral ends are likely associated with resistance against integrase inhibitors (Das and Berkhout, 2018; Varadarajan et al., 2013, 2016). Hence, the VINSSRM may also be useful to understand the resistance mechanism against integrase inhibitors.

However, there are still limitations of the VINSSRM for integration site identification. First, since it is difficult to determine an integration site from a read sequence that can be mapped onto two or more positions of the human genome, the VINSSRM ignores the read sequences that result in ambiguous mapping. Therefore, we may still miss the detection of some integration sites. Second, we can detect integration sites of vDNA with limited types of aberrant ends by the VINSSRM because of the short-read character of Illumina sequencing. It is difficult to detect integration sites following insertion of a different human chromosome or unknown origin (Supplementary Fig. S1B) (Varadarajan et al., 2013). When the extra insertion is short, we may detect the insertion type of aberrant ends by split-read mapping of a read sequence and subsequent alignment of an unmapped portion within the read sequence. In fact, we could find seven reads from a sample (P8) that support the integration site at chr11:71,237,165 flanking the vDNA with intact ends but with 42-bp-long insertions from chr22:40,851,344. However, we cannot distinguish whether or not the insertion is derived from artificial PCR recombination. Further improvement of the bioinformatics analytical protocol or application of long-read deep sequencing is required to detect these insertion types of aberrant ends.

We have proposed a novel bioinformatics protocol to search HIV-1 integration sites from deep sequencing read sequences. The VINSRRM has been tested for integration site identification of HIV-1 but would also be suitable for other retroviruses, such as HTLV-1. The VINSSRM may help us to better understand retroviral integration, the escape mechanism from integrase inhibitors, and the persistence of infected cells.

Funding

This study was supported in part by a Grant-in-Aid for Scientific

Research (C) from the Japan Society for the Promotion of Science (JSPS) KAKENHI [JP17K08872 to H.O.] and the Program for an Integrated Database of Clinical and Genomic Information from the Japan Agency for Medical Research and Development (AMED) [JP18kk0205011h020 to Y.I.].

Acknowledgments

We thank the laboratory staff for kind help and assistance.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.jviromet.2019.03.004.

References

Achuthan, V., Perreira, J.M., Sowd, G.A., Puray-Chavez, M., McDougall, W.M., Paulucci-Holthauzen, A., Wu, X., Fadel, H.J., Poeschla, E.M., Multani, A.S., Hughes, S.H., Sarafianos, S.G., Brass, A.L., Engelman, A.N., 2018. Capsid-CPSF6 interaction licenses nuclear HIV-1 trafficking to sites of viral DNA integration. Cell Host Microb. 24, 392–404. https://doi.org/10.1016/j.chom.2018.08.002. e398.

Batzer, M.A., Deininger, P.L., 2002. Alu repeats and human genomic diversity. Nat. Rev. Genet. 3, 370–379. https://doi.org/10.1038/nrg798.

Bui, J.K., Halvas, E.K., Fyne, E., Sobolewski, M.D., Koontz, D., Shao, W., Luke, B., Hong, F.F., Kearney, M.F., Mellors, J.W., 2017. Ex vivo activation of CD4+ T-cells from donors on suppressive ART can lead to sustained production of infectious HIV-1 from a subset of infected cells. PLoS Pathog. 13, e1006230. https://doi.org/10.1371/journal.ppat.1006230.

Cohn, L.B., Silva, I.T., Oliveira, T.Y., Rosales, R.A., Parrish, E.H., Learn, G.H., Hahn, B.H., Czartoski, J.L., McElrath, M.J., Lehmann, C., Klein, F., Caskey, M., Walker, B.D., Siliciano, J.D., Siliciano, R.F., Jankovic, M., Nussenzweig, M.C., 2015. HIV-1 integration landscape during latent and active infection. Cell 160, 420–432. https://doi.org/10.1016/j.cell.2015.01.020.

Das, A.T., Berkhout, B., 2018. How polypurine tract changes in the HIV-1 RNA genome can cause resistance against the integrase inhibitor dolutegravir. mBio 9, e00006–00018. https://doi.org/10.1128/mBio.00006-18.

Demeulemeester, J., Vets, S., Schrijvers, R., Madlala, P., De Maeyer, M., De Rijck, J., Ndung'u, T., Debyser, Z., Gijsbers, R., 2014. HIV-1 integrase variants retarget viral integration and are associated with disease progression in a chronic infection cohort. Cell Host Microb. 16, 651–662. https://doi.org/10.1016/j.chom.2014.09.016.

Ebina, H., Kanemura, Y., Suzuki, Y., Urata, K., Misawa, N., Koyanagi, Y., 2012. Integrase-independent HIV-1 infection is augmented under conditions of DNA damage and produces a viral reservoir. Virology 427, 44–50. https://doi.org/10.1016/j.virol.2012.02.004.

Esposito, D., Craigie, R., 1998. Sequence specificity of viral end DNA binding by HIV-1 integrase reveals critical regions for protein-DNA interaction. EMBO J. 17, 5832–5843. https://doi.org/10.1093/emboj/17.19.5832.

Gao, K., Butler, S.L., Bushman, F., 2001. Human immunodeficiency virus type 1 integrase: arrangement of protein domains in active cDNA complexes. EMBO J. 20, 3565–3576. https://doi.org/10.1093/emboj/20.13.3565.

Gaur, M., Leavitt, A.D., 1998. Mutations in the human immunodeficiency virus type 1 integrase D,D(35)E motif do not eliminate provirus formation. J. Virol. 72, 4678–4685.

Hiener, B., Horsburgh, B.A., Eden, J.S., Barton, K., Schlub, T.E., Lee, E., von Stockenstrom, S., Odevall, L., Milush, J.M., Liegler, T., Sinclair, E., Hoh, R., Boritz, E.A., Douek, D., Fromentin, R., Chomont, N., Deeks, S.G., Hecht, F.M., Palmer, S., 2017. Identification of genetically intact HIV-1 proviruses in specific CD4(+) t cells from effectively treated participants. Cell Rep. 21, 813–822. https://doi.org/10.1016/j.celrep.2017.09.081.

Ho, Y.C., Shan, L., Hosmane, N.N., Wang, J., Laskey, S.B., Rosenbloom, D.I., Lai, J., Blankson, J.N., Siliciano, J.D., Siliciano, R.F., 2013. Replication-competent non-induced proviruses in the latent reservoir increase barrier to HIV-1 cure. Cell 155, 540–551. https://doi.org/10.1016/j.cell.2013.09.020.

Holman, A.G., Coffin, J.M., 2005. Symmetrical base preferences surrounding HIV-1, avian sarcoma/leukosis virus, and murine leukemia virus integration sites. Proc. Natl. Acad. Sci. U. S. A. 102, 6103–6107. https://doi.org/10.1073/pnas.0501646102.

Kent, W.J., 2002. BLAT–the BLAST-like alignment tool. Genome Res. 12, 656–664. https://doi.org/10.1101/gr.229202.

Koyanagi, Y., Miles, S., Mitsuyasu, R.T., Merrill, J.E., Vinters, H.V., Chen, I.S., 1987. Dual infection of the central nervous system by AIDS viruses with distinct cellular tropisms. Science 236, 819–822.

Li, H., 2013. Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM. https://arxiv.org/abs/1303.3997.

Li, H., Durbin, R., 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589–595. https://doi.org/10.1093/bioinformatics/btp698.

Lusic, M., Siliciano, R.F., 2017. Nuclear landscape of HIV-1 infection and integration. Nat. Rev. Microbiol. 15, 69–82. https://doi.org/10.1038/nrmicro.2016.162.

Maldarelli, F., Wu, X., Su, L., Simonetti, F.R., Shao, W., Hill, S., Spindler, J., Ferris, A.L., Mellors, J.W., Kearney, M.F., Coffin, J.M., Hughes, S.H., 2014. HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. Science 345, 179–183. https://doi.org/10.1126/science.1254194.

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

Ode, H., Matsuda, M., Matsuoka, K., Hachiya, A., Hattori, J., Kito, Y., Yokomaku, Y., Iwatani, Y., Sugiura, W., 2015. Quasispecies analyses of the HIV-1 near-full-length genome with illumina MiSeq. Front. Microbiol. 6, 1258. https://doi.org/10.3389/fmicb.2015.01258.

Pinzone, M.R., O'Doherty, U., 2018. Measuring integrated HIV DNA ex vivo and in vitro provides insights about how reservoirs are formed and maintained. Retrovirology 15, 22. https://doi.org/10.1186/s12977-018-0396-3.

Pirooznia, M., Goes, F.S., Zandi, P.P., 2015. Whole-genome CNV analysis: advances in computational approaches. Front. Genet. 6, 138. https://doi.org/10.3389/fgene.2015.00138.

Sakurai, Y., Komatsu, K., Agematsu, K., Matsuoka, M., 2009. DNA double strand break repair enzymes function at multiple steps in retroviral infection. Retrovirology 6, 114. https://doi.org/10.1186/1742-4690-6-114.

Satou, Y., Katsuya, H., Fukuda, A., Misawa, N., Ito, J., Uchiyama, Y., Miyazato, P., Islam, S., Fassati, A., Melamed, A., Bangham, C.R.M., Koyanagi, Y., Sato, K., 2017. Dynamics and mechanisms of clonal expansion of HIV-1-infected cells in a humanized mouse model. Sci. Rep. 7, 6913. https://doi.org/10.1038/s41598-017-07307-4.

Schroder, A.R., Shinn, P., Chen, H., Berry, C., Ecker, J.R., Bushman, F., 2002. HIV-1 integration in the human genome favors active genes and local hotspots. Cell 110, 521–529.

Simonetti, F.R., Sobolewski, M.D., Fyne, E., Shao, W., Spindler, J., Hattori, J., Anderson, E.M., Watters, S.A., Hill, S., Wu, X., Wells, D., Su, L., Luke, B.T., Halvas, E.K., Besson, G., Penrose, K.J., Yang, Z., Kwan, R.W., Van Waes, C., Uldrick, T., Citrin, D.E., Kovacs, J., Polis, M.A., Rehm, C.A., Gorelick, R., Piatak, M., Keele, B.F., Kearney, M.F., Coffin, J.M., Hughes, S.H., Mellors, J.W., Maldarelli, F., 2016. Clonally expanded CD4+ T cells can produce infectious HIV-1 in vivo. Proc. Natl. Acad. Sci. U. S. A. 113, 1883–1888. https://doi.org/10.1073/pnas.1522675113.

Sowd, G.A., Serrao, E., Wang, H., Wang, W., Fadel, H.J., Poeschla, E.M., Engelman, A.N., 2016. A critical role for alternative polyadenylation factor CPSF6 in targeting HIV-1 integration to transcriptionally active chromatin. Proc. Natl. Acad. Sci. U. S. A. 113, E1054–1063. https://doi.org/10.1073/pnas.1524213113.

Tattini, L., D'Aurizio, R., Magi, A., 2015. Detection of genomic structural variants from next-generation sequencing data. Front. Bioeng. Biotechnol. 3, 92. https://doi.org/10.3389/fbioe.2015.00092.

Thankaswamy-Kosalai, S., Sen, P., Nookaew, I., 2017. Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. Genomics 109, 186–191. https://doi.org/10.1016/j.ygeno.2017.03.001.

Treangen, T.J., Salzberg, S.L., 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. 13, 36–46. https://doi.org/10.1038/nrg3117.

Varadarajan, J., McWilliams, M.J., Hughes, S.H., 2013. Treatment with suboptimal doses of raltegravir leads to aberrant HIV-1 integrations. Proc. Natl. Acad. Sci. U. S. A. 110, 14747–14752. https://doi.org/10.1073/pnas.1305066110.

Varadarajan, J., McWilliams, M.J., Mott, B.T., Thomas, C.J., Smith, S.J., Hughes, S.H., 2016. Drug resistant integrase mutants cause aberrant HIV integrations. Retrovirology 13 (71). https://doi.org/10.1186/s12977-016-0305-6.

Wagner, T.A., McLaughlin, S., Garg, K., Cheung, C.Y., Larsen, B.B., Styrchak, S., Huang, H.C., Edlefsen, P.T., Mullins, J.I., Frenkel, L.M., 2014. HIV latency. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. Science 345, 570–573. https://doi.org/10.1126/science.1256304.

Wildschutte, J.H., Baron, A., Diroff, N.M., Kidd, J.M., 2015. Discovery and characterization of Alu repeat sequences via precise local read assembly. Nucleic Acids Res. 43, 10292–10307. https://doi.org/10.1093/nar/gkv1089.

Yoshinaga, T., Fujiwara, T., 1995. Different roles of bases within the integration signal sequence of human immunodeficiency virus type 1 in vitro. J. Virol. 69, 3233–3236.

Yoshinaga, T., Kimura-Ohtani, Y., Fujiwara, T., 1994. Detection and characterization of a functional complex of human immunodeficiency virus type 1 integrase and its DNA substrate by UV cross-linking. J. Virol. 68, 5690–5697.

Zhang, J., Kobert, K., Flouri, T., Stamatakis, A., 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30, 614–620. https://doi.org/10.1093/bioinformatics/btt593.

Zhao, M., Wang, Q., Wang, Q., Jia, P., Zhao, Z., 2013. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics 14 (Suppl. 11), S1. https://doi.org/10.1186/1471-2105-14-S11-S1.