Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Distinct Strengths and Limitations of Short- and Long-Read Sequencing for HTLV-1 Proviral Characterization

Sharmin Nahar Sithi
Hokkaido University
Natasha Jansz
The University of Melbourne
Samiul Alam Rajib
Hokkaido University
Ashley Hirons
The University of Melbourne
Naho Ninomiya
Kumamoto University
Sachiho Kawahara
Hokkaido University
Jiazhou Li
Kagoshima University
Shingo Nakahata
Kagoshima University
Shingo Usuki
Kumamoto University
Hiroo Katsuya
Saga University
Masahito Tokunaga
Imamura Hospital
Atae Utsunomiya
Imamura Hospital
Damian F.J. Purcell
The University of Melbourne
Yorifumi Satou
y-satou@kumamoto-u.ac.jp

Hokkaido University

Research Article

Keywords: HTLV-1, Retrovirus, HIV-1, Defective provirus, Short-read sequencing, Long-read sequencing, Proviral integration site, Clonal expansion, Sequencing bias

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11538000/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Distinct Strengths and Limitations of Short- and Long-Read Sequencing for HTLV-1 Proviral Characterization

Sharmin Nahar Sithi[1,2], Natasha Jansz[3], Samiul Alam Rajib[1,2], Ashley Hirons[3], Naho Ninomiya[2], Sachiho Kawahara[1], Jiazhou Li[4], Shingo Nakahata[4], Shingo Usuki[5], Hiroo Katsuya[6], Masahito Tokunaga[7], Atae Utsunomiya[7], Damian F.J. Purcell[3], Yorifumi Satou[1,2]*

[1]Department of Microbiology and Immunology, Faculty of Medicine, Hokkaido University, Sapporo, Japan

[2]Division of Genomics and Transcriptomics, Joint Research Center for Human Retrovirus Infection, Kumamoto University, Kumamoto, Japan

[3]Department of Microbiology and Immunology at the Peter Doherty Institute for Infection and Immunity, University of Melbourne, Victoria, Australia

[4]Division of HTLV-1/ATL Carcinogenesis and Therapeutics, Joint Research Center for Human Retrovirus Infection, Kagoshima University, Kagoshima, Japan

[5]Liaison Laboratory Research Promotion Center, Institute of Molecular Embryology and Genetics, Kumamoto University, Kumamoto, Japan

[6]Division of Haematology, Respiratory Medicine and Oncology, Department of Internal Medicine, Faculty of Medicine, Saga University, Saga, Japan

[7]Department of Hematology, Imamura General Hospital, Kagoshima, Japan

* Corresponding author: Yorifumi Satou

Kita 15, Nishi 7, Kita-ku, Sapporo, 060-8638, Japan

Tel: +81-11-716-2111

Email: y-satou@kumamoto-u.ac.jp

Abstract (276 words)

Background

Human T-cell leukemia virus type 1 (HTLV-1) persists predominantly through clonal proliferation of infected cells carrying integrated proviruses. Comprehensive characterization of proviral sequences, structural alterations, integration sites, and clonal abundance is essential for understanding HTLV-1 persistence and pathogenesis. Although short- and long-read sequencing technologies have advanced proviral analysis, their relative strengths and limitations, particularly in detecting defective proviruses, remain incompletely understood.

Results

We compared three sequencing approaches using clinical samples from six individuals infected with HTLV-1a: short-read DNA-capture sequencing, single-provirus amplification followed by nanopore long-read sequencing (SPA-ONT-seq), and PCR-free Oxford Nanopore Technologies (ONT) long-read whole-genome sequencing (WGS). DNA-capture sequencing enabled simultaneous characterization of proviral sequences, integration sites, and clonal abundance, and detected type 2 defective proviruses lacking the 5′ long terminal repeat (LTR). However, its sensitivity for detecting internal deletions was limited, particularly for those present at low abundance. In contrast, SPA-ONT-seq resolved the structures of individual proviruses, including internal deletions, rearrangements, and host-derived insertions. Among 106 amplified proviruses, 74 were full length, 26 contained internal deletions, five contained host-derived insertions, and one was rearranged. However, SPA-ONT-seq failed to detect type 2 defective proviruses because the amplification primers targeted the 5′ LTR. PCR-free ONT long-read WGS independently detected 5′-defective proviruses and provided integration-site information without amplification, although low viral read depth limited population-level analysis.

Conclusions

Our findings demonstrate that sequencing approaches introduce distinct, method-dependent biases into HTLV-1 proviral characterization. Short-read DNA-capture sequencing and SPA-ONT-seq provide complementary information on infected-cell clonality and individual proviral structures, while PCR-free ONT WGS can overcome amplification-associated biases at the expense of sequencing depth. Combining these approaches enables a more comprehensive

assessment of proviral populations and provides a framework for selecting appropriate strategies for integrated viral genome analysis.

Keywords: HTLV-1; Retrovirus; HIV-1; Defective provirus; Short-read sequencing; Long-read sequencing; Proviral integration site; Clonal expansion; Sequencing bias

Introduction

Human T-cell leukemia virus type 1 (HTLV-1) was the first human retrovirus identified and is the causative agent of adult T-cell leukemia/lymphoma (ATL) and HTLV-1-associated myelopathy/tropical spastic paraparesis (HAM/TSP). ATL was first characterized as a distinct clinical entity in Japan [1] , and HTLV-1 was subsequently identified as the causative retrovirus[2]. HAM was later recognized as a distinct neurological disease associated with HTLV-1 infection [3, 4]. An estimated 5–10 million people worldwide were reported to be infected with HTLV-1, although the actual number may be higher [5]. Several HTLV-1 subtypes have been identified, among which the cosmopolitan subtype HTLV-1a is the most widespread, whereas HTLV-1c is predominantly found in First Nations communities in Central Australia and other Oceania regions [6, 7]. Most infected individuals remain asymptomatic throughout their lives, while a small proportion develop ATL or inflammatory diseases after prolonged infection [8]. Understanding the mechanisms underlying viral persistence and disease development therefore remains an important challenge in HTLV-1 research.

Following reverse transcription, HTLV-1 integrates into the host genome as a provirus and persists predominantly through clonal proliferation of infected T cells. Because independently infected cells generally carry the provirus at different genomic integration sites, and each infected cell carries only a single provirus [9], the integration site serves as a molecular marker for identifying infected-cell clones. Early studies demonstrated that clonal expansion occurs even in asymptomatic HTLV-1 carriers and that infected-cell clones can persist for years [10, 11]. Subsequent advances in high-throughput integration-site analysis enabled comprehensive identification of infected-cell clones and quantitative estimation of their relative abundance [12]. In addition to the integration site, the proviral structure itself is important for understanding

HTLV-1 persistence and pathogenesis. Defective proviruses frequently contain internal deletions while retaining both long terminal repeats (LTRs), designated type 1, or lack the 5′ LTR, designated type 2 [13]. Loss of the 5′ LTR can suppress Tax expression while preserving the 3′ LTR and HBZ, potentially contributing to immune evasion and the survival of infected cells [14, 15]. Defective proviruses are frequently detected in ATL and are also present in asymptomatic carriers and patients with HAM/TSP [16]. Therefore, comprehensive characterization of proviral sequences, structural integrity, integration sites, and clonal abundance is essential for understanding the biology of HTLV-1-infected cells.

Advances in high-throughput sequencing technologies have substantially improved our ability to characterize proviruses integrated within the vast excess of host genomic DNA. Short-read sequencing, particularly when combined with targeted DNA capture, enables efficient recovery of viral sequences and virus-host junctions from infected-cell populations. This approach has been used to characterize HTLV-1 proviral sequences, integration sites, and clonal abundance simultaneously [16, 17]. However, short DNA fragments limit the reconstruction of complete individual proviruses and the detection of complex structural alterations [18]. In contrast, long-read sequencing technologies, including Oxford Nanopore Technologies (ONT) sequencing, can resolve large deletions, rearrangements, and insertions within individual proviral molecules [19]. Recent applications to HTLV-1 have demonstrated their utility for characterizing proviral structures and integration sites, including complex defective proviruses and virus–host chimeric sequences [7, 20]. Nevertheless, some long-read approaches require PCR amplification of viral DNA, which may introduce biases depending on the proviral structure and primer-binding sites. PCR-free ONT long-read whole-genome sequencing (WGS) can directly analyze native DNA molecules without targeted enrichment, but, because the provirus represents a minute fraction of the infected-cell genome, yields correspondingly few viral reads. Thus, the relative advantages and limitations of these approaches, particularly in detecting different classes of defective proviruses, require systematic evaluation using the same clinical samples.

In this study, we compared three sequencing approaches for HTLV-1 proviral characterization: short-read DNA-capture-seq, single-provirus amplification followed by nanopore sequencing

(SPA-ONT-seq), and PCR-free ONT long-read WGS. Using clinical samples from individuals infected with HTLV-1a, we evaluated their performance in characterizing proviral sequences, detecting defective and rearranged proviruses, identifying integration sites, and assessing infected-cell clonality. By directly comparing these approaches, we aimed to clarify their respective strengths and limitations and provide a practical basis for selecting appropriate sequencing strategies for HTLV-1 and other retroviral research.

Result

Experimental workflow and clinical sample information

To evaluate how the choice of sequencing approach affects HTLV-1 proviral analysis, we characterized the same clinical samples using three methods: short-read HTLV-1 DNA-capture-seq, single-provirus amplification followed by long-read Oxford Nanopore Technologies (ONT) sequencing (SPA-ONT-seq), and amplification-free long-read ONT whole-genome sequencing (WGS) (Figure 1). DNA-capture-seq was performed using 148 probes, each 120 bp in length, as described in our previous study [16]. For SPA-ONT-seq, nearly full-length HTLV-1 proviral DNA was amplified by limiting-dilution long-range PCR and subsequently subjected to ONT long read sequencing, following a previously reported approach applied to HTLV-1c [7]. To enable a robust comparison of these approaches, we selected individuals living with HTLV-1 who had relatively high proviral loads, comprising two asymptomatic carriers (AC-1 and AC-2) and four patients with indolent ATL (ATL-1 to ATL-4). Their clinical characteristics are summarized in Table 1.

HTLV-1 provirus characterization by the Illumina short-read sequence

Following DNA library preparation, we performed targeted DNA capture using 148 probes designed against HTLV-1a. Between 2.4 and 12.4 million reads were mapped per sample, of which 546 to 8,889 reads (0.01% to 0.29% of mapped reads) aligned to the HTLV-1 provirus (Table 2). Relative to the proviral fraction expected from each individual's proviral load, this

represents an enrichment of approximately 300- to 3,800-fold, comparable in magnitude to that reported in our previous study [16].

Read alignments across the provirus were visualized using the Integrative Genomics Viewer (IGV) and are shown in Figure 2A. Mean read depth over the region between the two LTRs nucleotides 756 to 8,507 of reference (GenBank accession no. AB513134) ranged from 16.5× to 266× across the six samples. In some samples, coverage was relatively uniform across the entire provirus, suggesting that full-length proviruses constituted the predominant population. In contrast, other samples showed uneven coverage patterns, consistent with the presence and clonal expansion of defective proviruses.

We next compared proviral sequences between and within individuals (Figure 2B). For each individual, a consensus proviral sequence was built from the capture-seq reads and compared with the reference (GenBank accession no. AB513134). Each consensus differed from the reference at 2 to 53 positions, and ATL-3 and ATL-4 shared most of their variant positions, indicating closely related proviruses. Variation within individuals was far lower: only three positions across the entire cohort carried two nucleotides at appreciable frequency. Each individual therefore carried an essentially homogeneous proviral population, and sequence differences arose mainly between individuals rather than within them.

The virus–host chimeric reads recovered by DNA-capture-seq also allowed proviral integration sites to be mapped across the human genome (Figure 2C). Identification of integration-site was therefore possible with this approach.

Establishment of SPA-ONT-seq method for HTLV-1a subtype

A previous study characterizing HTLV-1c proviruses employed single-provirus amplification (SPA) followed by Oxford Nanopore Technologies (ONT) long-read sequencing [7]. Because HTLV-1a and HTLV-1c differ at several genomic positions [21], including the primer-binding regions, we designed a new primer set specifically for SPA of HTLV-1a (Figure 3A). After quantifying the proviral load by dPCR, genomic DNA was diluted to achieve an input corresponding to one or fewer proviral copies per PCR reaction (Figure 3B). To ensure specific amplification of HTLV-1a, we tested annealing temperatures of 56, 59 and 62 °C on HTLV-1-negative Jurkat DNA and on Jurkat DNA spiked with DNA from the HTLV-1-infected ED cell

line (Supplementary figure 1). At 56 and 59 °C, multiple non-specific products were amplified from HTLV-1-negative DNA. At 62 °C these were no longer detected and the expected full-length product was obtained only from HTLV-1-positive template, and 62 °C was therefore used throughout.

We then performed HTLV-1a SPA using six clinical samples, with representative amplification results shown in Figure 3C, and the complete set of reactions in Supplementary figure 2. Fewer PCR products than expected were obtained from some samples, potentially reflecting the presence of defective proviruses or differences in amplification efficiency. Because both forward primers anneal within the 5′ LTR, a provirus that has lost this region also lacks the forward primer site and cannot be amplified. SPA-ONT-seq, therefore, samples only those proviruses that retain the 5′ LTR, and every proportion reported for this method below is conditional on that restriction. Individual SPA products were sequenced on ONT. For each provirus, a consensus sequence was reconstructed from reads carrying both nested primer sequences, so that only reads spanning the complete amplicon contributed (median 826 reads per provirus); averaging across these reads compensates for the higher per-base error rate of the ONT platform. We obtained 106 single-provirus sequences, 11 to 25 per individual. Of these, 74 (70%) were full length, 26 (25%) carried an internal deletion, 5 (5%) were chimeric with host-derived sequence and 1 (1%) was rearranged (Figure 3D, E; Supplementary figure 3). The proportion of full-length proviruses varied considerably between individuals, from 89% in ATL-4 (17 of 19) to 25% in ATL-3 (4 of 16), in which 10 of 16 proviruses shared a single 410 bp deletion in *env*.

Comparison of DNA-capture-seq and SPA-ONT-seq

The two methods differ first in how deeply each one samples the proviral population. DNA-capture-seq achieved a mean read depth of 16.5× to 266× across the inter-LTR region, whereas SPA-ONT-seq yielded 11 to 25 independent proviruses per sample, partly reflecting the more labor-intensive nature of the SPA-based workflow (Figure 4A). Given the relatively

limited intra-host sequence diversity of HTLV-1, however, this difference may not substantially affect proviral characterization in most cases.

They differ more fundamentally in what each can resolve within a single provirus. Short reads cannot generally resolve the complete structure of individual proviruses, although internal deletions can be detected when reads span the deletion junction. In our original analysis, internal deletions were not identified from DNA-capture-seq data. We therefore refined the analysis to use soft-clipped reads, in which part of the read aligns to the provirus and the remainder aligns to a distant part of the provirus. Multiple soft-clipped reads sharing the same breakpoint provide evidence of an internal deletion, whereas reads with heterogeneous breakpoints were considered potential library preparation artifacts and excluded from further analysis (Figure 4B). Applying this criterion to the capture-seq data identified type 1 defective proviruses in two ATL patients that the original analysis had missed entirely, at 56% in ATL-3 and 7% in ATL-2 (Supplementary figure 4A).

DNA-capture-seq identifies virus–host chimeric reads and can therefore detect proviruses that have lost the 5′ LTR, whereas SPA-ONT-seq resolves the internal structure of individual provirus but, as described above, cannot amplify such proviruses at all. To compare the two directly, we estimated the composition of the proviral population in each individual by both methods (Figure 4C). The two assays disagreed in a systematic manner. DNA-capture-seq detected type 2 proviruses in four of the six individuals but remained considerably less sensitive to type 1 deletions than SPA-ONT-seq. Of the eighteen distinct internal deletions observed by SPA-ONT-seq, capture sequencing recovered only the two that were carried by more than one provirus: the 410 bp deletion in *env* in ATL-3, present in 10 of 16 proviruses and independently placed at nucleotides 5,718 to 6,123 by capture sequencing at 56% of proviral load, and the 1,072 bp deletion in ATL-2, present in 2 of 11 proviruses and placed at nucleotides 6,865 to 7,934 at 7.3% of proviral load. In both cases the two methods placed the deletion endpoints within 5 bp of one another. DNA-capture-seq locates a breakpoint from the common soft-clip boundary of short reads, whereas SPA-ONT-seq locates it from a gap in the alignment of a long-read consensus, and short direct repeats of 1 to 2 bp at these junctions allow the deletion to be placed at slightly different positions. None of the sixteen deletions observed in only a single provirus was detected by capture sequencing. Conversely, the dominant clone in ATL-1

was represented by 114 capture reads at a single integration site, with no read crossing the 5′ LTR, identifying it as a type 2 provirus; this clone was absent from the SPA-ONT-seq library prepared from the same genomic DNA. Because the two methods have complementary detection biases, we combined the type 2 fraction estimated by DNA-capture-seq with the type 1 fraction obtained by SPA-ONT-seq to derive an approximate composition of the proviral population, shown as the third row of Figure 4C.

One particular advantage of SPA-ONT-seq is its ability to detect chimeric proviruses, in which host-derived sequence is inserted within the provirus that are difficult to resolve using short reads. Consistent with previous report [7], we identified five proviral sequences containing host-derived insertions among the SPA-ONT-seq products (Figure 4D). In every case the inserted sequence derived from pericentromeric satellite repeats.

Because SPA relies on PCR, such chimeric provirus could in principle arise from mispriming within the host-derived segment rather than from a genuine virus–host junction. To exclude this, we aligned all four SPA primers, on both strands, to the 50 bp of viral sequence flanking each of the ten junctions carried by the five chimeric proviruses (Supplementary figure 4B). The best alignment obtained at any flank was 9 of 18 mismatches to the inner forward primer, or 10 of 19 mismatches to the outer forward primer, with the 3′-terminal base unpaired in most cases. No plausible primer-binding sites were identified near these junctions, making mispriming at the examined junction sequences an unlikely explanation for the observed chimeric structures.

HTLV-1 provirus characterization by amplification-free long-read whole-genome sequencing

Both approaches described above enrich for the provirus before sequencing, one by hybridization capture and one by PCR. To establish what is gained and lost by removing enrichment altogether, we performed deep (> 30X) amplification-free ONT long-read whole-genome sequencing on all six individuals. We achieved read-length N50s of > 25 kb, enabling identification of intact proviral sequences spanning host-virus junctions. Because the provirus occupies roughly one part in 10^5 of the infected genomes, only 4 to 19 reads per sample aligned to HTLV-1, in line with PVLs of 14.5 to 58.6%, giving a mean depth of 1.1× to 5.3× (Figure 5A). Depth is therefore the limiting property of this approach, an order of magnitude below DNA-

capture-seq and comparable to the number of proviruses recovered by SPA-ONT-seq. The integrity of the input genomic DNA, measured as the DNA shearing index (DSI) is given in Supplementary Table 1.

The reads that were obtained were nevertheless informative in a way that neither enrichment method allows. In AC-1, all four viral reads began at nucleotide 6,566 of AB513134 and continued into the 3′ LTR. Each read was a single DNA molecule of 24 to 48 kb, of which roughly 2.5 kb was proviral and the remainder host genome, with 16 to 36 kb of contiguous host sequence. Similarly, in ATL-4, reads began at nucleotide 6,036 on molecules carrying up to 93 kb of flanking host sequence. In both, the flanking sequence assigned these proviruses to a single integration site, chr11:55,449,127 in AC-1 and chr4:183,962,108 in ATL-4, each of which DNA-capture-seq had independently recovered with 49 and 514 reads respectively. These are 5′-defective proviruses observed directly, from unamplified single molecules, in precisely the class that SPA-ONT-seq cannot amplify. In ATL-1, three independent molecules shared the same 5′ viral boundary at nucleotide 4,208, identifying them as a single clone carrying a 5′-truncated provirus (Figure 5A).

One 26 kb read from ATL-2 carried a provirus with an internal deletion of 1,072 bp beginning at nucleotide 6,861 that DNA-capture-seq placed at nucleotides 6,865 to 7,934 (1,069 bp) and that SPA-ONT-seq called as 1,072 bp. Three methods with entirely different chemistries therefore placed one breakpoint within a few base pairs of one another.

Whole-genome sequencing also recovered integration sites, but sparsely (Figure 5B). In AC-1, AC-2, ATL-2 and ATL-4 it identified the most abundant clone, representing 94%, 50%, 32% and 98% of the clonal abundance measured by DNA-capture-seq; in ATL-1 and ATL-3 it identified no clone that capture sequencing had also detected. Integration-site detection is thus far more sensitive by short-read capture sequencing, although sites lying within repetitive regions may be missed by short reads and resolved by long ones.

The sparseness of proviral sequencing in these data sets a clear limitation of this approach. In ATL-3, five WGS reads spanned the env region, but none carried the approximately 410 bp internal deletion detected by DNA-capture-seq at an estimated 56% of proviral load and by

SPA-ONT-seq in 10 of 16 proviruses. Thus, variants present at moderate abundance may still be missed when only a limited number of proviruses are sampled.

Discussion

Integration of retroviral DNA into the host genome occurs at different genomic locations in independently infected cells. Because the integration site is inherited by daughter cells during cell division, it serves as a molecular marker for identifying individual infected-cell clones. This principle has long been used to investigate the clonal proliferation of HTLV-1-infected cells [10, 11]. Early studies using Southern blotting and PCR-based approaches demonstrated the presence of defective HTLV-1 proviruses, including those with internal deletions or loss of the 5′ long terminal repeat (LTR) [13]. Subsequent advances in high-throughput integration-site analysis enabled comprehensive identification of infected-cell clones and quantitative estimation of their relative abundance, providing a detailed picture of the clonal architecture within infected individuals [12, 22]. More recently, DNA-capture sequencing has enabled simultaneous analysis of proviral sequences, integration sites, and clonal abundance [16, 17], while long-read sequencing has expanded our ability to characterize complex proviral structures at the single-provirus level [20]. However, how different sequencing approaches influence the resulting characterization of proviral populations remains insufficiently understood. Here, we compared short-read DNA-capture sequencing, single-provirus amplification followed by nanopore sequencing (SPA-ONT-seq), and PCR-free long-read whole-genome sequencing (WGS) using clinical samples from individuals infected with HTLV-1a. Our findings demonstrate that these approaches provide complementary information, with distinct technical biases that substantially influence the apparent composition of the proviral population.

One of the most notable differences was the detection efficiency of defective proviruses. DNA-capture-seq successfully identified type 2 defective proviruses lacking the 5′ LTR, whereas these proviruses were not recovered by SPA-ONT-seq. This discrepancy is explained by the location of the forward primers within the 5′ LTR, which prevents amplification of proviruses lacking this region. Such bias is particularly relevant to HTLV-1 infection, where 5′-defective

proviruses may undergo substantial clonal expansion [16]. Loss of the 5′ LTR suppresses plus-strand viral gene expression, including Tax, while potentially preserving HBZ expression from the 3′ LTR, thereby facilitating escape from Tax-specific cytotoxic T-cell responses [14, 15]. Consequently, LTR-dependent amplification may systematically underestimate biologically important infected-cell populations. A comparable, assay-dependent problem is well recognised in HIV-1 research, where the measured size of the latent reservoir depends on which regions of the provirus an assay interrogates, and the discrepancy between assays has itself become a subject of study [23, 24]. Conversely, SPA-ONT-seq was more effective in detecting internal deletions, including those present in relatively infrequent proviruses. In DNA-capture-seq, we improved deletion detection by examining partially aligned reads, termed soft-clipped reads, in which one portion aligns to a particular position in the reference genome while the remaining sequence can be aligned to a distant position. Multiple reads supporting the same junction provide evidence of an internal deletion. Nevertheless, this approach predominantly detected deletions represented by multiple proviruses, whereas SPA-ONT-seq resolved structural alterations within individual amplified proviruses. These findings highlight the difference between population-level detection and single-provirus structural characterization. An additional advantage of DNA-capture-seq is its applicability to the analysis of proviral chromatin and transcriptional regulation. Our previous ChIP-seq analysis of the HTLV-1 provirus highlighted the challenge of obtaining sufficient viral sequence reads from conventional sequencing libraries [25]. This challenge motivated the development of approaches combining viral DNA capture with chromatin-based sequencing assays. Such approaches subsequently contributed to the identification of an intragenic enhancer regulating HTLV-1 transcription [26] and a viral silencer involved in transcriptional latency [27]. These studies demonstrate that DNA-capture-based enrichment can facilitate not only proviral sequence characterization but also the investigation of transcriptional and epigenetic regulatory mechanisms within integrated viral genomes.

The two approaches also differed in their ability to identify integration sites and assess infected-cell clonality. DNA-capture-seq detects virus–host junction reads, allowing integration-site identification and estimation of clonal abundance based on the representation of individual

integration sites. In contrast, our SPA-ONT-seq approach amplifies proviral sequences without retaining the flanking host DNA and therefore cannot directly identify integration sites. Although identical proviral sequences may suggest clonal expansion, sequence identity alone cannot establish cellular clonality, particularly given that each infected cell carries a single provirus in natural infection [9] and that intra-host sequence diversity of HTLV-1 is relatively limited. DNA-capture-seq also achieved substantially greater sequencing depth, allowing broader sampling of the proviral population, whereas SPA-ONT-seq characterized fewer independent proviruses because of its labor-intensive workflow. However, sequencing depth and the number of independently characterized proviruses are not directly equivalent measures. Short reads also have limited ability to link distant sequence variations or structural alterations within individual proviruses, while long reads provide greater sequence continuity. These limitations are method-dependent rather than universal properties of the sequencing platforms, as alternative long-read approaches, such as PCIP-seq, can recover proviral sequences together with their integration sites [20].

PCR-free, unenriched long-read WGS further clarified the origins of these technical differences. By sequencing native DNA without targeted amplification, this approach identified 5′-defective proviruses that could not be recovered by SPA-ONT-seq, confirming that their absence was attributable to the amplification strategy rather than an intrinsic limitation of long-read sequencing. An internal deletion was also identified with closely corresponding breakpoints across all three methods, supporting the reliability of these structural observations. This may be particularly useful for HTLV-1, where proviral transcription is regulated by both structural alterations and epigenetic modifications [27, 28]. However, the low abundance of proviral DNA relative to host genomic DNA resulted in very limited viral read depth. Thus, although PCR-free WGS provides valuable information on individual native proviruses and their genomic context, substantially greater sequencing depth or targeted enrichment would be required for comprehensive population-level analysis.

Long-read sequencing also identified proviral structures containing host-derived DNA insertions, including sequences originating from pericentromeric satellite regions. Similar virus–

host chimeric proviruses have been reported in HTLV-1c infection [7]. However, their biological authenticity requires further investigation, as artificial chimeric molecules may arise through template switching during PCR amplification or other processes associated with library preparation. Although we found no plausible alternative primer-binding sites near the observed junctions, this does not exclude other sources of artifacts. Notably, these chimeric proviruses were not independently demonstrated to belong to expanded infected-cell clones, an observation consistent with the possibility that some represent amplification- or library preparation-associated artifacts rather than stable integrated proviruses. In HTLV-1c, such chimeric proviruses were validated with independent junction spanning PCR and sequencing [7]. However, this would not be possible with the repetitive sequences detected in these samples.

Comparison with previous HTLV-1c findings raises additional questions regarding the origin of proviral structural diversity. A recent study that developed SPA-ONT-seq approach reported a markedly lower proportion of full-length proviruses in HTLV-1c infection than observed in our HTLV-1a cohort [7]. This difference may reflect intrinsic biological differences between viral subtypes, distinct selective pressures, or differences in clinical backgrounds. However, methodological factors, particularly the quality and integrity of the input genomic DNA, may also influence the apparent frequencies of defective proviruses. DNA fragmentation, damage during sample storage or processing, and preferential amplification of particular proviral structures could alter the population recovered by long-range PCR. Moreover, both studies examined only proviruses that retained the primer-binding regions required for amplification. Therefore, whether the observed differences reflect genuine subtype-specific biology or differences in DNA quality and experimental conditions remains unresolved. Standardized comparisons using DNA samples of comparable integrity and complementary PCR-free approaches will be important for addressing this question.

The utility of viral DNA-capture-seq extends beyond HTLV-1 to other viruses that integrate their genomes into host DNA. This approach has been applied to characterize HIV-1 proviral sequences and integration sites [29], bovine leukemia virus (BLV) proviral structures and infected-cell clonality [30], human papillomavirus (HPV) genomic sequences in cancer tissues

[31], and hepatitis B virus (HBV) integration in hepatocellular carcinoma [32]. These applications demonstrate the versatility of target enrichment for characterizing integrated viral DNA across different virus families. In particular, bovine leukemia virus (BLV) and simian T-cell leukemia virus type 1 (STLV-1), like HTLV-1, persist in substantial part through clonal proliferation of infected cells. This has been directly demonstrated in BLV infection, in which the relative contribution of viral propagation shifts from de novo infection toward clonal expansion during the course of infection [33]. The approaches evaluated here should therefore be directly relevant to studies of proviral integrity, integration sites, and clonal architecture in these infections. By contrast, HIV and SIV can undergo extensive rounds of viral replication involving reverse transcription, generating considerable sequence diversity through mutation and recombination. This diversity introduces additional challenges for comprehensive proviral characterization, including sequence-dependent biases in primer binding and hybridization capture. Nevertheless, infected-cell clonal expansion also contributes importantly to HIV persistence, particularly during suppressive antiretroviral therapy [34, 35]. Thus, the methodological principles identified here remain applicable to HIV and SIV, although greater sequence diversity and more complex population structures may require additional assay optimization. Similar considerations apply to murine leukemia virus (MLV), depending on the experimental model and the relative contributions of ongoing infection and infected-cell proliferation.

Taken together, our findings demonstrate that the optimal strategy for retroviral proviral characterization depends on the specific research objective (Table 3). Short-read DNA-capture sequencing is particularly useful for integration-site mapping, quantitative assessment of infected-cell clonality, and detection of 5′-defective proviruses across infected-cell populations. In contrast, SPA-ONT-seq provides detailed characterization of individual proviral structures, including internal deletions, rearrangements, and complex insertions, but is limited by its dependence on conserved primer-binding regions. Combining these approaches can provide a more comprehensive view of the proviral population. Future development of PCR-free targeted long-read sequencing, including CRISPR/Cas9-based enrichment [36] and nanopore adaptive sampling [37], may further overcome these limitations by enabling simultaneous

analysis of proviral structure, integration sites, and epigenetic modifications without amplification-associated bias. Although our study was limited by a relatively small cohort, selection of samples with high proviral loads, and low viral read depth in PCR-free WGS, it provides a practical framework for selecting appropriate sequencing strategies. Importantly, the apparent structure and diversity of retroviral proviruses reflect not only underlying viral biology but also the technical characteristics of the methods used to investigate them. Understanding these methodological biases is essential for accurate interpretation of proviral persistence, clonal expansion, and retroviral evolution.

Methods

Ethics statement and clinical samples

All protocols involving human subjects were reviewed and approved by the Institutional Review Board of Kumamoto University (approval numbers 248 and 263). The study was conducted in accordance with the principles of the Declaration of Helsinki, and written informed consent was obtained from all participants. Peripheral blood mononuclear cells (PBMCs) were isolated from whole blood within 24 h of collection by density-gradient centrifugation using Ficoll-Paque (GE Healthcare Life Sciences) according to the manufacturer's instructions, and were cryopreserved until use.

High-molecular-weight DNA extraction

High-molecular-weight (HMW) genomic DNA was extracted from PBMCs using the Monarch® spin gDNA Extraction Kit (New England Biolabs) following the manufacturer’s instructions. Briefly, approximately 1×10^6 frozen PBMCs were resuspended in 100 µL PBS, and 1 µL Proteinase K and 3 µL RNase A were added to the suspension, followed by lysis, protein precipitation and gentle bead-based binding and elution to preserve fragment length. Throughout the procedure, harsh vortexing was avoided and wide-bore pipette tips were used during elution and handling to minimize mechanical shearing; eluted DNA was allowed to homogenize at room temperature before quantification. DNA concentration was measured with a Qubit dsDNA BR assay (Thermo Fisher Scientific) and purity was confirmed on a NanoDrop

spectrophotometer. Fragment-size distribution was assessed using Genomic DNA ScreenTape on an Agilent TapeStation system (Agilent Technologies).

Quantification of proviral load by digital PCR

Proviral load (PVL), expressed as the percentage of cells carrying HTLV-1 provirus, was determined by digital PCR on the QuantStudio Absolute Q Digital PCR System (Applied Biosystems). The copy number of the viral *tax* gene was quantified and normalized to the copy number of the single-copy host albumin (*ALB*) gene. Because *ALB* is present at two copies per diploid cell, PVL was calculated as: PVL (%) = [(copy number of *tax*) / (copy number of *ALB*) / 2] x 100.

The *tax* and *ALB* targets were amplified in a single-well duplex reaction. Each 10 µL reaction contained 2 µL of 5× master mix (Supermix), 0.9 µL each of the *tax* and *ALB* primer pairs (forward and reverse, 10 µM), 0.2 µL each of the FAM-labelled *tax* probe and HEX-labelled *ALB* probe, 3.8 µL of nuclease-free water and 2 µL of template DNA (5 ng/µL; 10 ng total). Primer and probe sequences (Supplementary Table 2) were as described previously [16]. Reactions were partitioned and amplified on the QuantStudio Absolute Q instrument under the following conditions: initial denaturation at 96 °C for 10 min, followed by 40 cycles of 96 °C for 5 s and 60 °C for 30 s. Absolute copy numbers were derived from the fraction of positive partitions using the Absolute Q Analysis Software. Each sample was assayed in duplicate together with a no-template control, and only reactions meeting the instrument's minimum valid-partition threshold were accepted. Reported PVL therefore represents the number of *tax* copies per 100 cells (assuming one integrated provirus per infected cell).

DNA shearing index

DNA integrity was assessed by digital PCR on the QuantStudio Absolute Q Digital PCR System (Applied Biosystems) using the human *RPP30* copy-reference and *RPP30*-Shear assays. Primer and probe sequences were adapted from published study with minor modification (Supplementary Table 2) [38]. Partitions positive for both probes represent unsheared templates; single-positive partitions represent fragments sheared between the two targets. The DNA shearing index was calculated as S̃/(S̃ + D), where D is the number of double-positive

partitions and $\bar{S}$ the mean of the two single-positive counts. Approximately 20,400 partitions were analyzed per sample.

Short-read DNA-capture-seq

Whole proviral genomes and their flanking host sequences were enriched by targeted DNA-capture sequencing, adapted from previous study with some minor modifications [16, 17]. Genomic DNA was sheared to an average fragment size of ~300 bp using a Bioruptor® Pico device (Diagenode), and short-read sequencing libraries were prepared from 1 µg of sheared DNA with the NEBNext® Ultra II DNA Library Prep Kit for Illumina (New England Biolabs) following the manufacturer's protocol. Fragments were end-repaired and dA-tailed, ligated to adaptors, size-selected for ~300–400 bp inserts with AxyPrep magnetic beads (Axygen), and amplified with a unique dual index per sample. Library size and concentration were checked on an Agilent TapeStation before pooling.

Up to four uniquely indexed libraries were pooled (500 ng total) and hybridized to a panel of HTLV-1-specific biotinylated capture probes (xGen® Lockdown® probes, Integrated DNA Technologies) at 65 °C in the presence of Cot-1 DNA and xGen® Universal Blockers to suppress non-specific and repetitive binding. Hybridized targets were captured on streptavidin-coated magnetic beads (Dynabeads™ M-280, Thermo Fisher Scientific), washed under stringent and room-temperature conditions to remove off-target DNA, and the enriched pool was amplified directly off the beads and purified. The final capture library was quantified on an Agilent TapeStation and sequenced on an Illumina NovaSeq system (2 × 150 bp paired-end).

Single-provirus amplification and long-read sequencing (SPA-ONT-seq)

To resolve the structure of individual proviruses, single-provirus amplification (SPA) coupled with long-read sequencing was performed following the HTLV-1c SPA-ONT-seq method [7], with modifications to the PCR reactions to accommodate HTLV-1 subtype-1a. SPA is a limiting-dilution nested PCR designed to amplify single copies of integrated provirus. Genomic DNA was serially diluted such that fewer than 30% of wells yielded an amplification product; by Poisson statistics this corresponds to a >80% probability that a given amplicon derives from a single integrated provirus. Subtype-1a-specific primers were designed for the HTLV-1 long terminal repeats (LTRs) and were staggered to amplify as much of the full-length provirus as

possible while avoiding non-specific LTR amplification; the amplicon was biased toward the 3′ LTR and away from the 5′ LTR spanning approximately nucleotides 685–8892 of the reference (GenBank accession no. AB513134). Primer specificity was confirmed by control PCR on HTLV-1-negative genomic DNA. Primer sequences are provided in Supplementary Table 2.

Both PCR rounds were carried out in 25 µL reactions using PrimeSTAR GXL DNA Polymerase (Takara Bio). The first-round reaction contained 1 µL PrimeSTAR GXL DNA Polymerase, 5 µL 5× PrimeSTAR GXL Buffer, 2 µL dNTP mixture, 1 µL outer primer mix (forward and reverse, 10 µM each), 1 µL limiting-dilution template DNA and 15 µL nuclease-free water, and was cycled as follows: initial denaturation at 98 °C for 2 min; 30 cycles of 98 °C for 10 s, 62 °C for 15 s and 68 °C for 90 s; and a final extension at 68 °C for 5 min. The first-round product was diluted 20-fold and 2 µL of the dilution served as template for the nested reaction, which was assembled identically but with the inner primer mix (10 µM) and 14 µL water and run for 35 cycles under the same conditions. Amplification products were resolved on an agarose gel, and provirus bands were visualized under blue light, excised and purified using the QIAquick Gel Extraction Kit (Qiagen).

Purified single-provirus amplicons were barcoded and prepared for sequencing using the ONT native barcoding workflow (Native Barcoding Kit, SQK-NBD114.96) and sequenced on the ONT MinION platform. Basecalling, demultiplexing and adaptor trimming were performed with Dorado using the super-high-accuracy model.

Long-read whole-genome sequencing

Long-read whole-genome sequencing was performed on the Oxford Nanopore Technologies (ONT) PromethION platform. Sequencing libraries were prepared from 500 ng of HMW genomic DNA using the Ligation Sequencing Kit V14 (SQK-LSK114, ONT) according to the manufacturer's protocol with minor modifications. DNA was first end-repaired and dA-tailed using the NEBNext FFPE DNA Repair Mix and the NEBNext Ultra II End-repair/dA-tailing module (New England Biolabs), and cleaned with AMPure XP beads (Beckman Coulter). Sequencing adapters were ligated using the Ligation Adapter (LA), NEBNext Quick Ligation Module and Salt-T4® DNA Ligase, followed by a Long Fragment Buffer (LFB) clean-up to retain long fragments. The final library was quantified on a Qubit fluorometer and loaded onto a

PromethION flow cell (R10.4.1) according to the ONT instructions. Basecalling, demultiplexing and adaptor trimming were performed with Dorado using the super-high-accuracy model[7]

Bioinformatics analysis

DNA-capture-seq. Provirus sequences and integration site analysis was performed following previous study with some modifications [16]. Adaptors were removed with cutadapt (v4.5) and read quality assessed with FastQC (v0.11.9). Reads were aligned with BWA-MEM (v0.7.17-r1188) to a composite reference consisting of GRCh38 with AB513134 added as an additional contig. PCR duplicates removed with Picard MarkDuplicates (v3.1.1); depth and breadth across the inter-LTR region were computed with SAMtools (v1.18). Internal deletions were called from soft-clipped reads sharing a common breakpoint.

SPA-ONT-seq. Reads were processed per excised gel band, each band representing one candidate single provirus. Reads were first size-filtered using a window derived from the mode of the read-length distribution of the band; where the range spanned by the most frequent lengths among the top 1,000 reads exceeded 200 bp, the window was set to the mode ± 200 bp. Length-filtered reads were then required to contain exact matches to both nested (inner) primers, each searched as its given sequence and as its reverse complement so that reads sequenced from either strand were retained. A consensus sequence was built for each band from all reads passing both filters with lamassemble (v1.7.2, built on LAST v1652), using an alignment matrix trained on the HTLV-1a reference with last-train. Consensus sequences were aligned to AB513134 with minimap2 (v2.24-r1122) and classified as full length, internally deleted, chimeric or rearranged, taking a reference gap of at least 100 bp as a deletion and an unaligned internal segment of at least 100 bp, identified by BLAT against the T2T-CHM13 (HS1) human assembly, as a chimeric insertion**.**

Whole-genome sequencing. Reads were converted to FASTQ retaining the MM and ML tags and aligned to GRCh38 with minimap2 (v2.24; -ax map-ont -Y -y), then sorted and indexed with SAMtools. HTLV-1 integration sites were called from the host alignment with TLDR (v1.2.2), using the HTLV-1 proviral sequence as the query element, with consensus extension of 10 kb and a flank size of 1 kb. The same reads were then realigned to the AB513134 proviral reference, again retaining the modification tags; reads mapping to the provirus were extracted and their read names used to recover the corresponding alignments from the host-aligned BAM,

so that each provirus could be examined together with its flanking host sequence. Viral read counts, mean depth, breadth, aligned-block lengths, virus–host junctions and internal deletions of at least 100 bp were summarized per sample, and host sequence clipped at the junctions was compared with the capture-seq integration sites to establish which clones each platform recovered.

Figures were produced in Python with Matplotlib (v3.10.9), and the values underlying every panel are provided as per-panel source-data files.

Data availability

Data supporting the findings reported in this study are available from the corresponding author upon request. We are preparing data deposition in the SRA, and the FASTQ files generated during this study will be made publicly available upon manuscript publication.

Code availability

The source code to reproduce the analysis will be uploaded to GitHub and made publicly available upon manuscript publication.

Correspondence and requests for materials should be addressed to Yorifumi Satou

Acknowledgements

We thank all individuals who generously donated clinical samples for this study, and the clinical staff involved in sample collection and processing. We thank Professor Takamasa Ueno, Joint Research Center for Human Retrovirus Infection, Kumamoto University, and Dr Hari Prasad Devkota, Graduate School of Pharmaceutical Sciences, Kumamoto University, for their scientific advice, helpful guidance and continuous support throughout this study.

Funding

This work was supported by the Japan Agency for Medical Research and Development (AMED) through the Multidisciplinary Research Program (26wm0325082) and the HIV/AIDS Research Program (JP25fk0410070 and JP25fk0410071); the Japan Science and Technology Agency

(JST) through the ASPIRE Program (JP29jf0126018) and the SPRING Program (JPMJSP2127); and the Ministry of Education, Culture, Sports, Science and Technology (MEXT)/Japan Society for the Promotion of Science (JSPS) through KAKENHI Grants-in-Aid for Scientific Research (B) (JP25K02693, JP23K27384, and JP24K02289), Scientific Research (C) (JP24K11620), Challenging Exploratory Research (JP25K22556); the National Health and Medical Research Council (NHMRC) of Australia (GNT2044048 and GNT1176574 to NJ); and the Centre for Pathogen Genomics (CPG-IHPR0230).

The funders had no role in the study design, data collection and analysis, interpretation of the results, or the decision to submit the manuscript for publication.

Author Contributions

S.N.S., N.J., S.A.R., A.H., N.N., S.K., J.L., S.U. performed the research; S.N.S., N.J., S.A.R. performed bioinformatics analysis; S.N.S., N.J., D.P., and Y.S. analyzed data; J.L., S.N., H.K., M.T., and A.U., contributed materials/reagents/analytic tools; S.N.S. and Y.S. wrote the original draft; all authors reviewed and revised the manuscript draft and approved the final version for submission.

Competing Interest Statement

Authors declare no competing interest.

Figures and legends

Table 1 | Clinical characteristics. Age is given in years at sampling. PVL, proviral load; WBC, white blood cell count; abnormal lymphocyte %, percentage of morphologically abnormal lymphocytes; sIL-2R, soluble interleukin-2 receptor. AC, asymptomatic carrier; ATL, adult T-cell leukaemia/lymphoma.

Table 2 | DNA-capture-seq sequencing metrics. Total mapped reads, reads mapping to the composite GRCh38 plus AB513134 reference after removal of PCR duplicates; viral reads, the subset mapping to AB513134; % of viral reads, viral reads as a percentage of total mapped reads; chimera reads, reads aligning partly to the provirus and partly to the human genome.

Table 3 | Comparison of the three sequencing approaches. For each approach the table gives the input material and the amount and quality of DNA required, the read depth achieved across the provirus, how densely the proviral population is sampled, the principal advantages and limitations observed in this study, and the approximate reagent cost and turnaround time. Depth and sampling values are the ranges measured across the six participants.

Figure 1 | Study design. Genomic DNA was extracted from PBMCs of six HTLV-1-infected individuals and analyzed in parallel by three approaches; Illumina short-read DNA-capture-seq, using HTLV-1 hybridization probes, Single-provirus amplification followed by Oxford Nanopore sequencing on MinION (SPA-ONT-seq), and Amplification-free long-read whole-genome sequencing on PromethION. The three outputs were then compared. The figure is created in BioRender.

Figure 2 | HTLV-1a provirus characterization by short-read DNA-capture-seq. (A) Read coverage across the provirus, aligned to the HTLV-1a reference (GenBank accession no. AB513134), visualized in IGV. The display range differs between tracks and is given in brackets at the top left of each. (B) Sequence differences from the reference. For each individual, a consensus proviral sequence was built from the capture-seq reads and compared with the reference (GenBank accession no. AB513134). Each black tick is one position that differs. (C) Proviral integration sites across the human genome (GRCh38), one ring per individual from AC-1 (outermost) to ATL-4 (innermost); bar height, log10(chimeric reads + 1).

Figure 3 | Establishment of SPA-ONT-seq for HTLV-1a. (A) Position of the HTLV-1a SPA primers on reference sequence (GenBank accession no. AB513134). The first-round (outer) product spans 8,267 bp, nucleotides 649 to 8,915; the nested second-round product spans 8,207 bp, nucleotides 686 to 8,892. Both forward primers anneal within the 5′ LTR. (B) Experimental workflow, from PBMC isolation and limiting dilution of genomic DNA to nested PCR, gel extraction, library preparation and sequencing on MinION. The figure is created in BioRender. (C) Representative amplification results for AC-1 and ATL-4. (D) Structure of each of the 106 single-provirus, one row per provirus, coloured by structural class and ordered by

672 class. Gaps within a provirus indicate regions absent from that provirus. (E) Proportion of each
673 structural class across all 106 proviruses.
674
675 **Figure 4 | Comparison of short-read DNA-capture-seq and long-read SPA-ONT-seq.** (A)
676 Sequencing depth: reads for DNA-capture-seq, and number of single proviruses for SPA-ONT-
677 seq. (B) Schematic of improved DNA-capture-seq analysis for the detection of internal
678 deletions. Soft-clipped reads whose breakpoints are heterogeneous arise from random ligation
679 during library preparation and were treated as artefacts, whereas multiple soft-clipped reads
680 sharing a single breakpoint were called as a true internal deletion. (C) Composition of the
681 proviral population in each individual estimated by DNA-capture-seq (top row), SPA-ONT-seq
682 (middle row) and the combined estimate (bottom row). (D) The five chimeric proviruses, for
683 each, the host-derived sequence inserted between the LTRs is shown with its length, its repeat
684 class and its coordinates in the T2T-CHM13 (HS1) assembly; the schematic at the top defines
685 the structure. Scale bar, 1 kb.
686
687 **Figure 5 | Amplification-free long-read whole-genome sequencing.** (A) Reads aligned to
688 the HTLV-1 provirus (GenBank accession no. AB513134). (B) Proportion of clones detected by
689 both DNA-capture-seq and WGS, and by DNA-capture-seq alone; the clone recovered by WGS
690 is labelled with the chromosome of its integration site.
691
692 **Supplementary figure 1 | Optimization of the PCR condition for HTLV-1a SPA.** Reactions
693 at 56°C, 59°C and 62°C using Jurkat DNA, Jurkat DNA spiked with ED DNA, and a no-template
694 control.
695
696 **Supplementary figure 2 | Complete set of SPA reactions.** (A) AC-1. (B) ATL-4.
697
698 **Supplementary figure 3 | Single-provirus structures of six individuals.** The 106 proviruses
699 shown separately for AC-1 (n = 12), AC-2 (n = 25), ATL-1 (n = 23), ATL-2 (n = 11), ATL-3 (n =
700 16) and ATL-4 (n = 19).
701

Supplementary figure 4 | (A) Composition of the proviral population measured by DNA-capture-seq under the original and the improved analysis. (B) Best alignment of the SPA primers to the 50 bp of viral sequence flanking each junction of the five chimeric proviruses. All four primers were tested on both strands and the best-scoring alignment is shown for each flank, with its mismatch count.

Supplementary Table 1 | DNA shearing index (DSI) for each sample.

Supplementary Table 2 | Oligonucleotides used in this study.

References

1. Uchiyama T, Yodoi J, Sagawa K, Takatsuki K, Uchino H. Adult T-cell leukemia: clinical and hematologic features of 16 cases. Blood. 1977;50:481–92.

2. Poiesz BJ, Ruscetti FW, Gazdar AF, Bunn PA, Minna JD, Gallo RC. Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient with cutaneous T-cell lymphoma. Proc Natl Acad Sci U S A. 1980;77:7415–9. https://doi.org/10.1073/pnas.77.12.7415.

3. Osame M, Usuku K, Izumo S, Ijichi N, Amitani H, Igata A, et al. HTLV-I associated myelopathy, a new clinical entity. Lancet. 1986;1:1031–2. https://doi.org/10.1016/s0140-6736(86)91298-5.

4. Gessain A, Barin F, Vernant JC, Gout O, Maurs L, Calender A, et al. Antibodies to human T-lymphotropic virus type-I in patients with tropical spastic paraparesis. Lancet. 1985;2:407–10. https://doi.org/10.1016/s0140-6736(85)92734-5.

5. Gessain A, Cassar O. Epidemiological Aspects and World Distribution of HTLV-1 Infection. Front Microbiol. 2012;3:388. https://doi.org/10.3389/fmicb.2012.00388.

6. Einsiedel L, Pham H, Talukder MR, Taylor K, Wilson K, Kaldor J, et al. Very high prevalence of infection with the human T cell leukaemia virus type 1c in remote Australian Aboriginal communities: Results of a large cross-sectional community survey. PLoS Negl Trop Dis. 2021;15:e0009915. https://doi.org/10.1371/journal.pntd.0009915.

7. Hirons A, Jansz N, Khoury G, Ellenberg P, Collins S, Cooney J, et al. Immunophenotype and proviral landscape of HTLV-1c infection and pulmonary disease. EBioMedicine. 2026;130:106403. https://doi.org/10.1016/j.ebiom.2026.106403.

8. Iwanaga M, Watanabe T, Utsunomiya A, Okayama A, Uchimaru K, Tsukasaki K, et al. Human T-cell leukemia virus type I (HTLV-1) proviral load and disease progression in asymptomatic HTLV-1 carriers: a nationwide prospective study in Japan. Blood. 2010;116:1211–9. https://doi.org/10.1182/blood-2009-12-257410.

9. Cook LB, Rowan AG, Melamed A, Taylor GP, Bangham CRM. HTLV-1-infected T cells contain a single integrated provirus in natural infection. Blood. 2012;120:3488–90. https://doi.org/10.1182/blood-2012-07-445593.

10. Wattel E, Vartanian JP, Pannetier C, Wain-Hobson S. Clonal expansion of human T-cell leukemia virus type I-infected cells in asymptomatic and symptomatic carriers without malignancy. J Virol. 1995;69:2863–8. https://doi.org/10.1128/JVI.69.5.2863-2868.1995.

11. Etoh K, Tamiya S, Yamaguchi K, Okayama A, Tsubouchi H, Ideta T, et al. Persistent clonal proliferation of human T-lymphotropic virus type I-infected cells in vivo. Cancer Res. 1997;57:4862–7.

12. Gillet NA, Malani N, Melamed A, Gormley N, Carter R, Bentley D, et al. The host genomic environment of the provirus determines the abundance of HTLV-1-infected T-cell clones. Blood. 2011;117:3113–22. https://doi.org/10.1182/blood-2010-10-312926.

13. Tamiya S, Matsuoka M, Etoh K, Watanabe T, Kamihira S, Yamaguchi K, et al. Two types of defective human T-lymphotropic virus type I provirus in adult T-cell leukemia. Blood. 1996;88:3065–73.

14. Satou Y, Yasunaga J, Yoshida M, Matsuoka M. HTLV-I basic leucine zipper factor gene mRNA supports proliferation of adult T cell leukemia cells. Proc Natl Acad Sci U S A. 2006;103:720–5. https://doi.org/10.1073/pnas.0507631103.

15. Miyazaki M, Yasunaga J-I, Taniguchi Y, Tamiya S, Nakahata T, Matsuoka M. Preferential selection of human T-cell leukemia virus type 1 provirus lacking the 5’ long terminal repeat during oncogenesis. J Virol. 2007;81:5714–23. https://doi.org/10.1128/JVI.02511-06.

16. Katsuya H, Islam S, Tan BJY, Ito J, Miyazato P, Matsuo M, et al. The Nature of the HTLV-1 Provirus in Naturally Infected Individuals Analyzed by the Viral DNA-Capture-Seq Approach. Cell Rep. 2019;29:724-735.e4. https://doi.org/10.1016/j.celrep.2019.09.016.

17. Miyazato P, Katsuya H, Fukuda A, Uchiyama Y, Matsuo M, Tokunaga M, et al. Application of targeted enrichment to next-generation sequencing of retroviruses integrated into the host human genome. Sci Rep. 2016;6:28324. https://doi.org/10.1038/srep28324.

18. Jansz N, Faulkner GJ. Viral genome sequencing methods: benefits and pitfalls of current approaches. Biochem Soc Trans. 2024;52:1431–47. https://doi.org/10.1042/BST20231322.

19. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. Nat Rev Genet. 2020;21:597–614. https://doi.org/10.1038/s41576-020-0236-x.

20. Artesi M, Hahaut V, Cole B, Lambrechts L, Ashrafi F, Marcais A, et al. PCIP-seq: simultaneous sequencing of integrated viral genomes and their insertion sites with long reads. Genome Biol. 2021;22:97. https://doi.org/10.1186/s13059-021-02307-0.

21. Hirons A, Yurick D, Jansz N, Ellenberg P, Franchini G, Einsiedel L, et al. High level of genomic divergence in orf-I p12 and hbz genes of HTLV-1 subtype-C in Central Australia. Retrovirology. 2024;21:14. https://doi.org/10.1186/s12977-024-00647-w.

22. Artesi M, Marcais A, Durkin K, Rosewick N, Hahaut V, Suarez F, et al. Monitoring molecular response in adult T-cell leukemia by high-throughput sequencing analysis of HTLV-1 clonality. Leukemia. 2017;31:2532–5. https://doi.org/10.1038/leu.2017.260.

23. Bruner KM, Wang Z, Simonetti FR, Bender AM, Kwon KJ, Sengupta S, et al. A quantitative approach for measuring the reservoir of latent HIV-1 proviruses. Nature. 2019;566:120–5. https://doi.org/10.1038/s41586-019-0898-8.

24. Reeves DB, Gaebler C, Oliveira TY, Peluso MJ, Schiffer JT, Cohn LB, et al. Impact of misclassified defective proviruses on HIV reservoir measurements. Nat Commun. 2023;14:4186. https://doi.org/10.1038/s41467-023-39837-z.

25. Satou Y, Miyazato P, Ishihara K, Yaguchi H, Melamed A, Miura M, et al. The retrovirus HTLV-1 inserts an ectopic CTCF-binding site into the human genome. Proc Natl Acad Sci U S A. 2016;113:3054–9. https://doi.org/10.1073/pnas.1423199113.

788 26. Matsuo M, Ueno T, Monde K, Sugata K, Tan BJY, Rahman A, et al. Identification and
789 characterization of a novel enhancer in the HTLV-1 proviral genome. Nat Commun.
790 2022;13:2405. https://doi.org/10.1038/s41467-022-30029-9.
791 27. Sugata K, Rahman A, Niimura K, Monde K, Ueno T, Rajib SA, et al. Intragenic viral silencer
792 element regulates HTLV-1 latency via RUNX complex recruitment. Nat Microbiol.
793 2025;10:1447–62. https://doi.org/10.1038/s41564-025-02006-7.
794 28. Koiwa T, Hamano-Usami A, Ishida T, Okayama A, Yamaguchi K, Kamihira S, et al. 5'-long
795 terminal repeat-selective CpG methylation of latent human T-cell leukemia virus type 1 provirus
796 in vitro and in vivo. J Virol. 2002;76:9389–97. https://doi.org/10.1128/jvi.76.18.9389-9397.2002.
797 29. Iwase SC, Miyazato P, Katsuya H, Islam S, Yang BTJ, Ito J, et al. HIV-1 DNA-capture-seq
798 is a useful tool for the comprehensive characterization of HIV-1 provirus. Sci Rep.
799 2019;9:12326. https://doi.org/10.1038/s41598-019-48681-5.
800 30. Ohnuki N, Kobayashi T, Matsuo M, Nishikaku K, Kusama K, Torii Y, et al. A target
801 enrichment high throughput sequencing system for characterization of BLV whole genome
802 sequence, integration sites, clonality and host SNP. Sci Rep. 2021;11:4521.
803 https://doi.org/10.1038/s41598-021-83909-3.
804 31. Morodomi S, Hirosue A, Rahman A, Nohata K, Matsuo M, Reda O, et al. Robust HPV-16
805 Detection Workflow for Formalin-Fixed Cancer Tissue and Its Application for Oral Squamous
806 Cell Carcinoma. Cancer Med. 2025;14:e70544. https://doi.org/10.1002/cam4.70544.
807 32. Saito T, Suzuki R, Rahman A, Mori K, Rajib SA, Kobayashi N, et al. Analysis of HBV
808 Genome Integration in Patients With HBV-Previously Infected NBNC-HCC. Hepatol Res.
809 2026;56:1101–13. https://doi.org/10.1111/hepr.70163.
810 33. Gillet NA, Gutierrez G, Rodriguez SM, de Brogniez A, Renotte N, Alvarez I, et al. Massive
811 depletion of bovine leukemia virus proviral clones located in genomic transcriptionally active
812 sites during primary infection. PLoS Pathog. 2013;9:e1003687.
813 https://doi.org/10.1371/journal.ppat.1003687.
814 34. Maldarelli F, Wu X, Su L, Simonetti FR, Shao W, Hill S, et al. HIV latency. Specific HIV
815 integration sites are linked to clonal expansion and persistence of infected cells. Science.
816 2014;345:179–83. https://doi.org/10.1126/science.1254194.

35. Wagner TA, McLaughlin S, Garg K, Cheung CYK, Larsen BB, Styrchak S, et al. HIV latency. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. Science. 2014;345:570–3. https://doi.org/10.1126/science.1256304.

36. Gilpatrick T, Lee I, Graham JE, Raimondeau E, Bowen R, Heron A, et al. Targeted nanopore sequencing with Cas9-guided adapter ligation. Nat Biotechnol. 2020;38:433–8. https://doi.org/10.1038/s41587-020-0407-5.

37. Payne A, Holmes N, Clarke T, Munro R, Debebe BJ, Loose M. Readfish enables targeted nanopore sequencing of gigabase-sized genomes. Nat Biotechnol. 2021;39:442–50. https://doi.org/10.1038/s41587-020-00746-x.

38. Kinloch NN, Ren Y, Conce Alberto WD, Dong W, Brumme CJ, Lee GQ, et al. HIV-1 diversity considerations in the application of the Intact Proviral DNA Assay (IPDA). Nat Commun. 2021;12:165. https://doi.org/10.1038/s41467-020-20442-3.

847 Table 1:

Sample ID	Age	Sex	Disease status	PVL%	WBC (/mm^3)	Abnormal lymphocyte %	Lymphocyte %	sIL-2R (U/ml)
AC-1	74	F	Carrier	17.5	5030	3	37	476
AC-2	59	F	Carrier	14.5	4370	0	61	761
ATL-1	72	F	Smoldering ATL	33.0	5820	15	30	434
ATL-2	72	M	Smoldering ATL	35.8	5680	5	37	404
ATL-3	70	M	Smoldering ATL	26.7	4360	4	15	515
ATL-4	72	F	Chronic ATL	58.6	9380	12	36	1077

848

849

850

851

852

853 Table 2:

Sample name	Total mapped reads	Viral reads	% of viral reads	Chimera reads
AC-1	12358498	886	0.01	223
AC-2	6593219	546	0.01	119
ATL-1	2546007	3293	0.13	782
ATL-2	4705652	8889	0.19	1917
ATL-3	4012377	1033	0.03	326
ATL-4	2365724	6818	0.29	2653

854

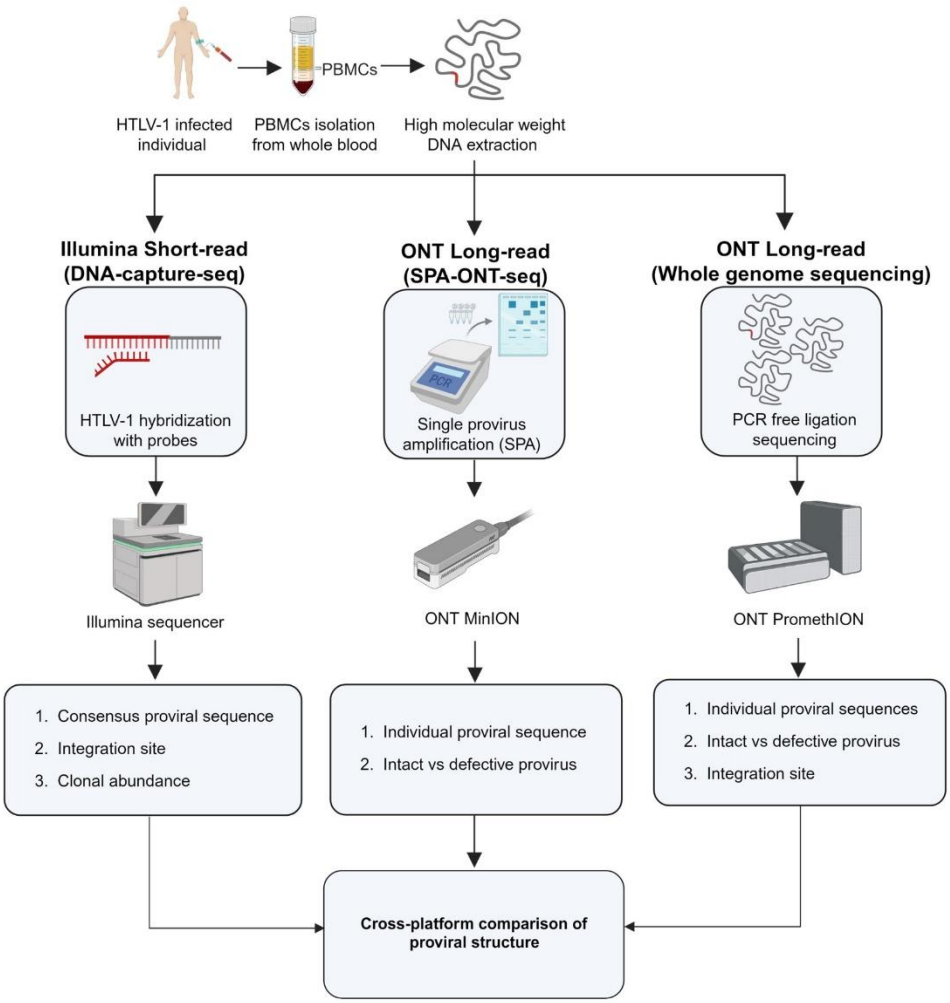
855

856 Table 3:

	DNA-capture-seq	SPA-ONT-seq	PCR-free long-read WGS
Input material	Bulk genomic DNA, hybridization capture	Single provirus amplified by limiting-dilution nested PCR	Native genomic DNA, no enrichment
Input DNA amount	1 µg genomic DNA	0.2–1.3 ng total across 96 limiting-dilution wells	500 ng
DNA quality	High molecular weight DNA is not required as DNA is sheared ~300bp	High molecular weight DNA extraction to preserve full-length provirus	High molecular weight DNA extraction to preserve fragment length
Read depth	16.5–266×	Not applicable	1.1–5.3×
Sampling of the proviral population	Population-level; 546–8,889 viral read fragments per sample, not resolvable into individual provirus	11–25 complete proviruses per sample	4–19 complete proviruses per sample
Advantages	1. Proviral sequence, integration sites and clonal abundance from one library. 2. Detection of type 2 defective proviruses. 3. No requirement for intact DNA. 4. Compatible with chromatin-based assays.	1. Complete structure of each provirus 2. Internal deletions resolved in individual proviruses 3. Host-sequence insertions and rearrangements detected 4. Sequence variation between proviruses 5. Generates no host genomic sequence, so it can be used where consent does not extend to host-genome sequencing, and the data can be shared without controlled access	1. No amplification or capture bias 2. Proviral sequence together with its integration site on the same molecule 3. Recovers 5′-deleted proviruses directly
Limitations	1. Structure largely inferred from coverage 2. Detects only internal deletions shared by two or more proviruses	4. Cannot amplify proviruses lacking the 5′ LTR (type 2 defective) 5. Does not retain flanking host DNA, so no integration site	1. Not quantitative at this depth 2. Samples the proviral population sparsely, so variants of moderate

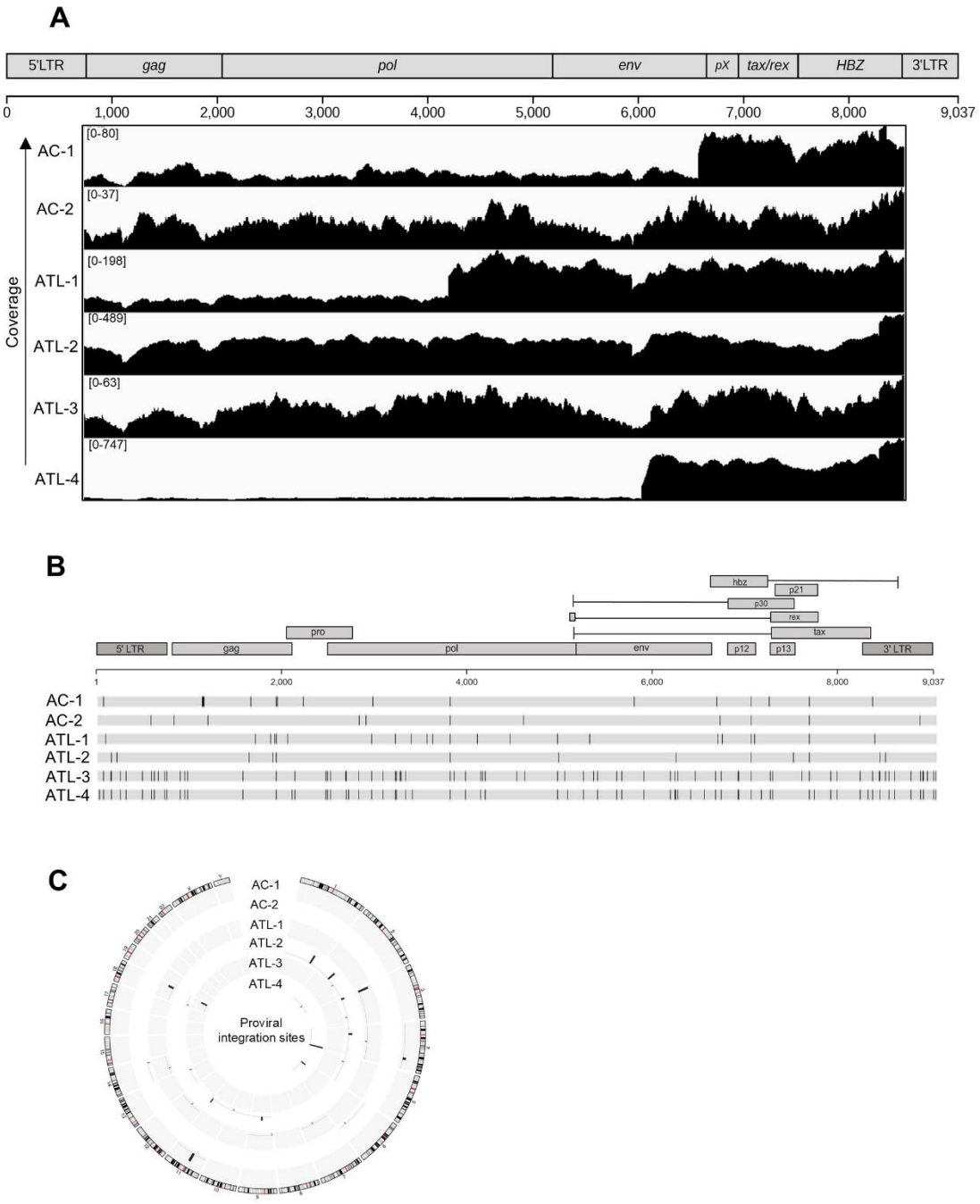
	3. Libraries are predominantly host sequence, so deposition may require controlled access	6. One PCR per molecule limits throughput	abundance may be missed 3. Requires high-molecular-weight DNA 4. Reads are predominantly host sequence, so deposition may require controlled access
Approximate reagent cost	~$10 per sample	~$11 per provirus	~$2000 per sample
Turnaround	5 days	Several weeks	3 days

Figure 1


PBMCs
HTLV-1 infected individual
PBMCs isolation from whole blood
High molecular weight DNA extraction
Illumina Short-read (DNA-capture-seq)
ONT Long-read (SPA-ONT-seq)
ONT Long-read (Whole genome sequencing)
HTLV-1 hybridization with probes
Single provirus amplification (SPA)
PCR free ligation sequencing
Illumina sequencer
ONT MinION
ONT PromethION
1. Consensus proviral sequence
2. Integration site
3. Clonal abundance
1. Individual proviral sequence
2. Intact vs defective provirus
1. Individual proviral sequences
2. Intact vs defective provirus
3. Integration site
Cross-platform comparison of proviral structure

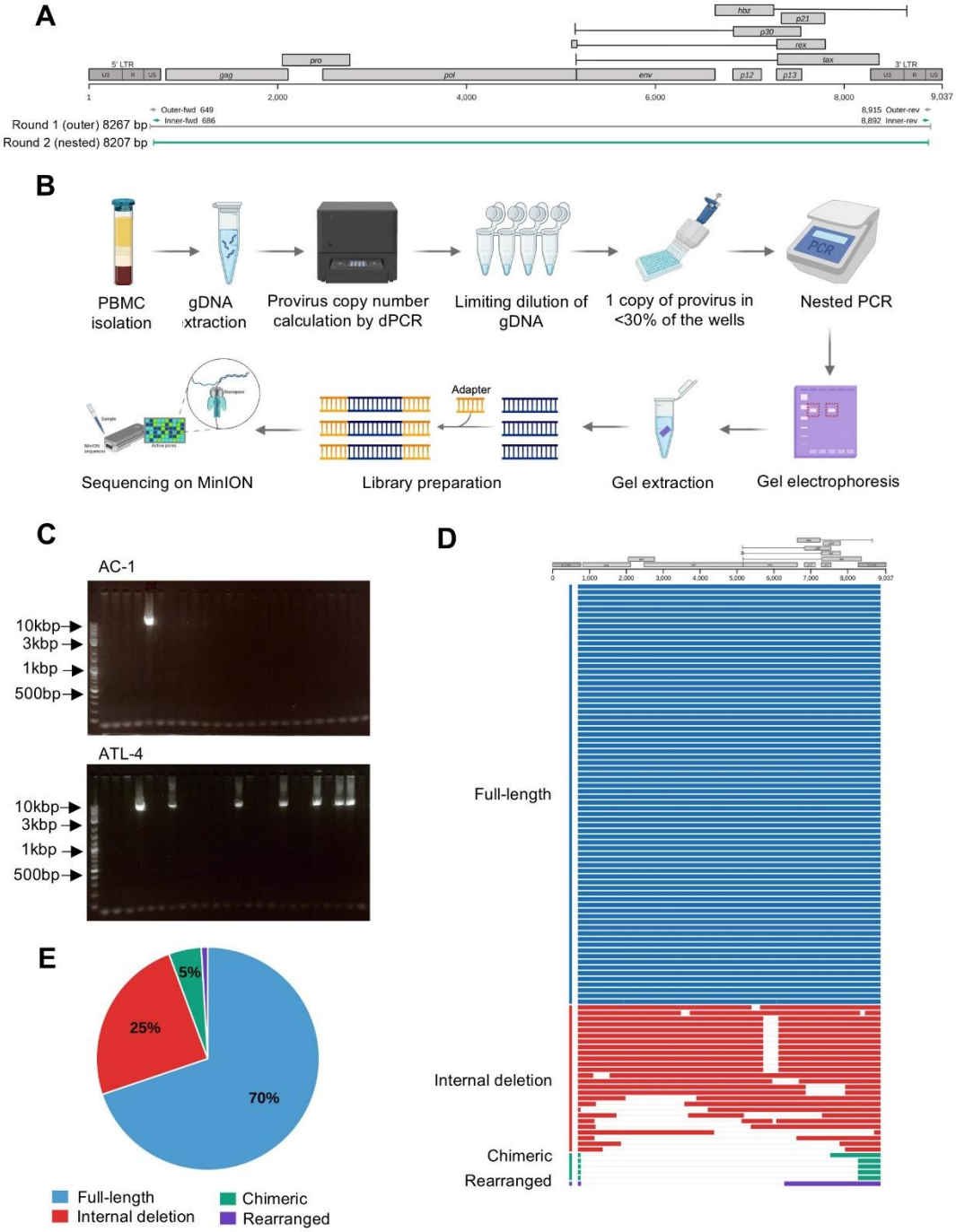
867

Figure 2


A
5'LTR
gag
pol
env
pX
tax/rex
HBZ
3'LTR
0
1,000
2,000
3,000
4,000
5,000
6,000
7,000
8,000
9,037
Coverage
AC-1
AC-2
ATL-1
ATL-2
ATL-3
ATL-4
[0-80]
[0-37]
[0-198]
[0-489]
[0-63]
[0-747]
B
hbz
p21
p30
rex
tax
pro
5' LTR
gag
pol
env
p12
p13
3' LTR
1
2,000
4,000
6,000
8,000
9,037
AC-1
AC-2
ATL-1
ATL-2
ATL-3
ATL-4
C
AC-1
AC-2
ATL-1
ATL-2
ATL-3
ATL-4
Proviral integration sites

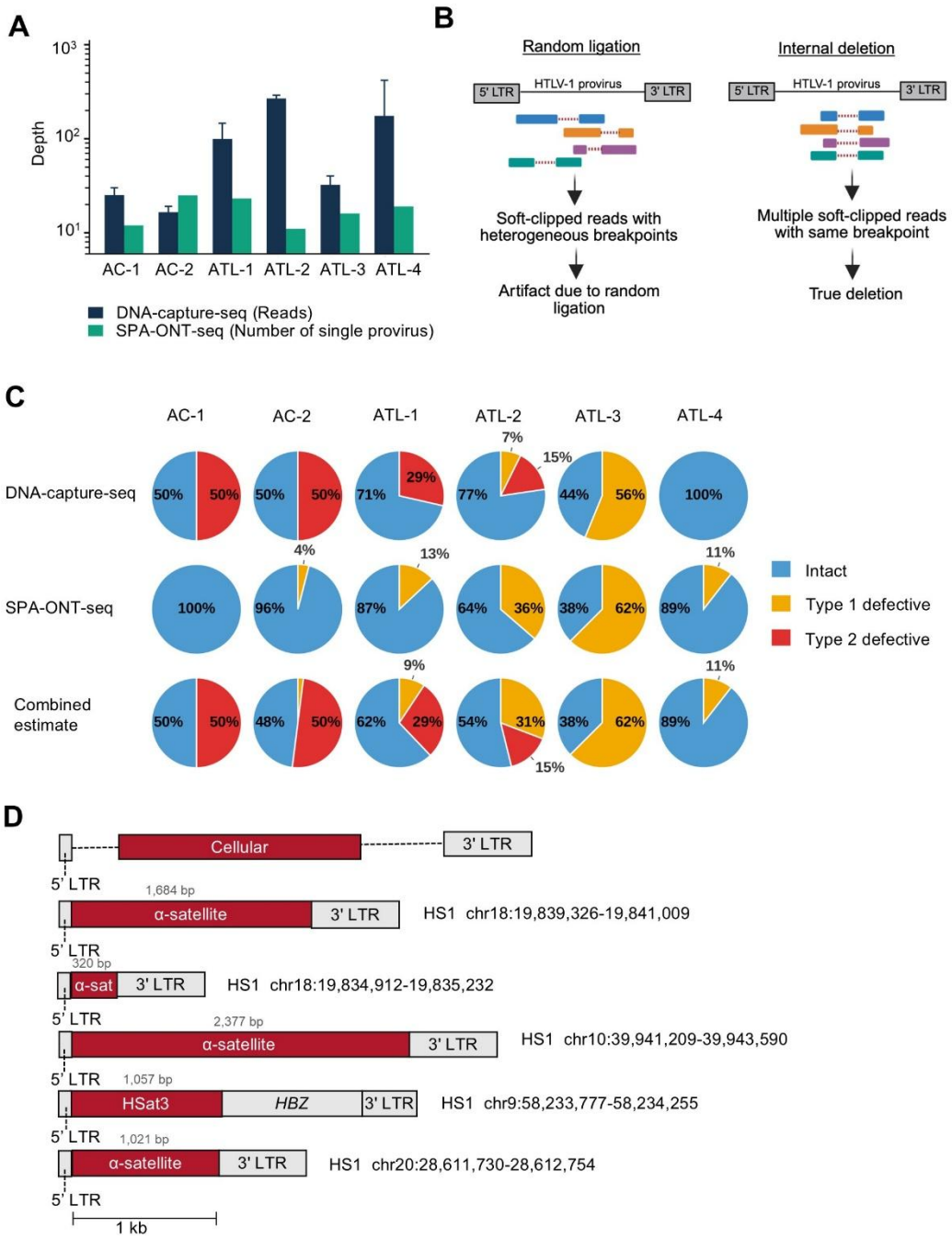
868

Figure 3


A
hbz
p21
p30
rex
tax
5' LTR
pro
gag
pol
env
p12
p13
3' LTR
1
2,000
4,000
6,000
8,000
9,037
Outer-fwd 649
Inner-fwd 686
8,915 Outer-rev
8,892 Inner-rev
Round 1 (outer) 8267 bp
Round 2 (nested) 8207 bp
B
PBMC isolation
gDNA extraction
Provirus copy number calculation by dPCR
Limiting dilution of gDNA
1 copy of provirus in <30% of the wells
Nested PCR
Gel electrophoresis
Gel extraction
Adapter
Library preparation
Sequencing on MinION
C
AC-1
10kbp
3kbp
1kbp
500bp
ATL-4
10kbp
3kbp
1kbp
500bp
D
Full-length
Internal deletion
Chimeric
Rearranged
E
5%
25%
70%
Full-length
Internal deletion
Chimeric
Rearranged

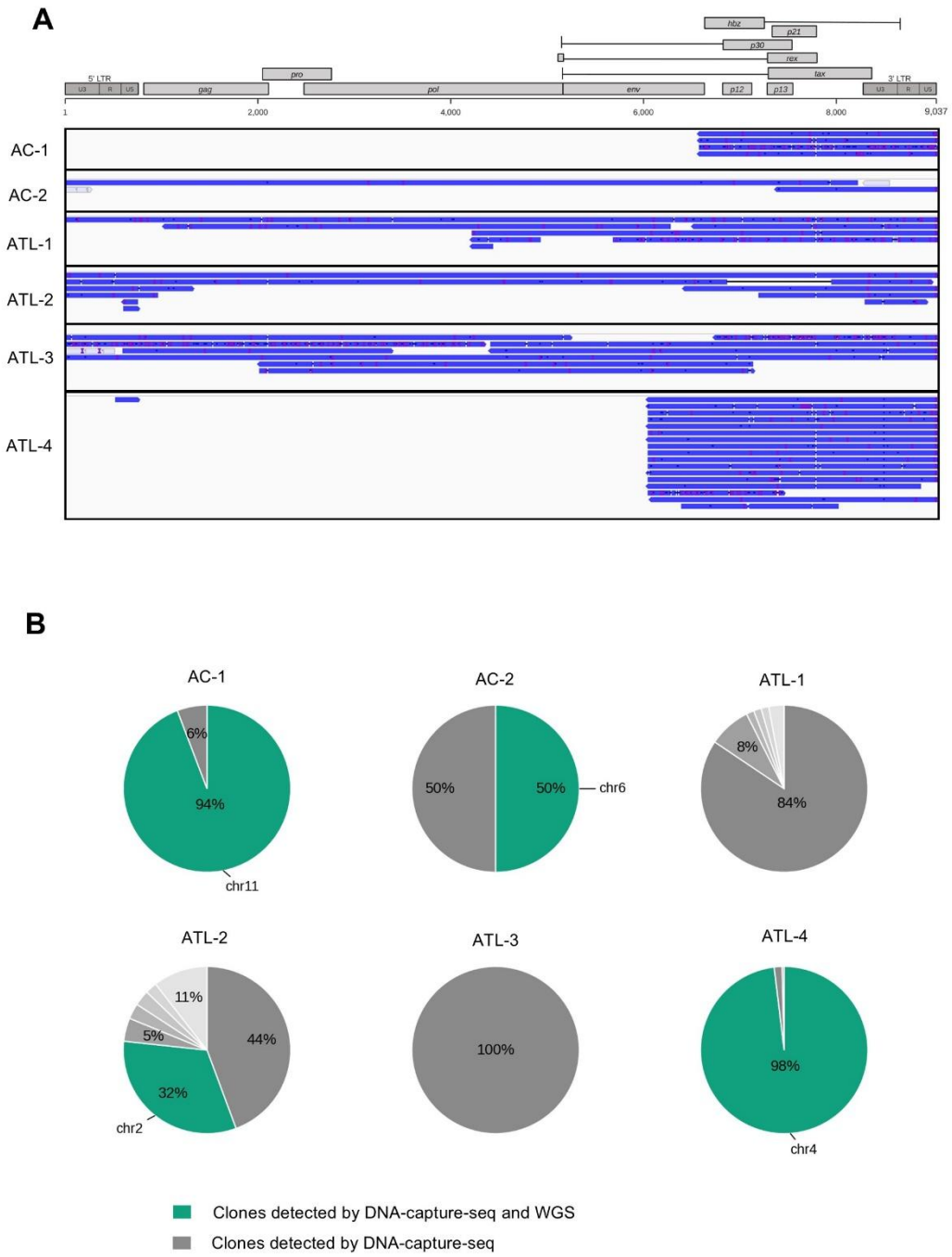
869

Figure 4


A
Depth
AC-1 AC-2 ATL-1 ATL-2 ATL-3 ATL-4
DNA-capture-seq (Reads)
SPA-ONT-seq (Number of single provirus)
B
Random ligation
Internal deletion
5' LTR
HTLV-1 provirus
3' LTR
Soft-clipped reads with heterogeneous breakpoints
Artifact due to random ligation
Multiple soft-clipped reads with same breakpoint
True deletion
C
AC-1 AC-2 ATL-1 ATL-2 ATL-3 ATL-4
DNA-capture-seq
SPA-ONT-seq
Combined estimate
Intact
Type 1 defective
Type 2 defective
D
Cellular
5' LTR
3' LTR
α-satellite
α-sat
HSat3
HBZ
1,684 bp
320 bp
2,377 bp
1,057 bp
1,021 bp
HS1 chr18:19,839,326-19,841,009
HS1 chr18:19,834,912-19,835,232
HS1 chr10:39,941,209-39,943,590
HS1 chr9:58,233,777-58,234,255
HS1 chr20:28,611,730-28,612,754
1 kb

870

Figure 5


A
hbz
p21
p30
rex
tax
5' LTR
U3
R
U5
pro
gag
pol
env
p12
p13
3' LTR
1
2,000
4,000
6,000
8,000
9,037
AC-1
AC-2
ATL-1
ATL-2
ATL-3
ATL-4
B
AC-1
6%
94%
chr11
AC-2
50%
50%
chr6
ATL-1
8%
84%
ATL-2
11%
5%
44%
32%
chr2
ATL-3
100%
ATL-4
98%
chr4
Clones detected by DNA-capture-seq and WGS
Clones detected by DNA-capture-seq

871

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementary1.jpg
- Supplementary2.jpg
- Supplementary1.jpg
- Supplementary2.jpg
- Supplementary4.jpg
- Supplementary3.jpg
- Supplementary5.jpg