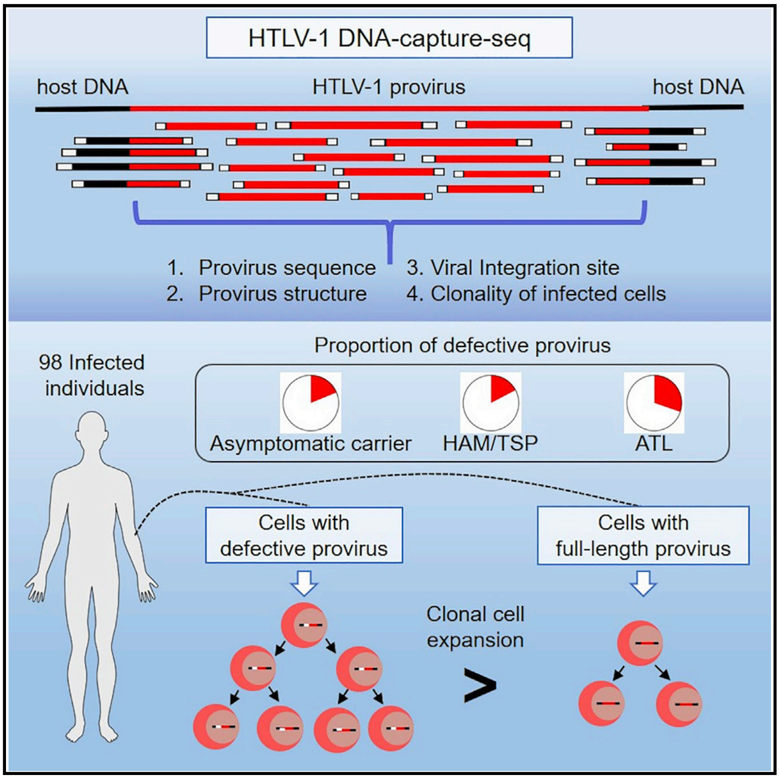
Cell Reports

Article

The Nature of the HTLV-1 Provirus in Naturally Infected Individuals Analyzed by the Viral DNA-Capture-Seq Approach

Graphical Abstract



Highlights

- A method for comprehensive analysis of HTLV-1 proviruses in infected individuals
- The method provides viral sequence, integration site, and degree of cell expansion
- Defective proviruses are present in asymptomatic carriers and HAM/TSP, as well as ATL
- Infected cells with defective proviruses proliferate more than those with intact ones

Authors

Hiroo Katsuya, Saiful Islam,
Benjy Jek Yang Tan, ..., Atae Utsunomiya,
Yoshihisa Yamano, Yorifumi Satou

Correspondence

y-satou@kumamoto-u.ac.jp

In Brief

Katsuya et al. demonstrate that HTLV-1 DNA-capture-seq provides comprehensive information, including the entire viral sequence, integration site, and clonal abundance of infected cells. Infected clones with defective-type proviruses are present in disease states and in asymptomatic carriers, and they proliferate more than full-length proviruses.

Katsuya et al., 2019, Cell Reports *29*, 724–735
October 15, 2019 © 2019 The Authors.
https://doi.org/10.1016/j.celrep.2019.09.016

CellPress

The Nature of the HTLV-1 Provirus in Naturally Infected Individuals Analyzed by the Viral DNA-Capture-Seq Approach

Hiroo Katsuya,[1,2,13] Saiful Islam,[1,2,13] Benjy Jek Yang Tan,[1,2] Jumpei Ito,[3] Paola Miyazato,[1,2] Misaki Matsuo,[1,2] Yuki Inada,[2,4] Saori C. Iwase,[1,2] Yoshikazu Uchiyama,[5] Hiroyuki Hata,[4] Tomoo Sato,[6] Naoko Yagishita,[6] Natsumi Araya,[6] Takaharu Ueno,[7] Kisato Nosaka,[8] Masahito Tokunaga,[9] Makoto Yamagishi,[10] Toshiki Watanabe,[11] Kaoru Uchimaru,[10] Jun-ichi Fujisawa,[7] Atae Utsunomiya,[9,12] Yoshihisa Yamano,[6] and Yorifumi Satou[1,2,14,*]

[1]Joint Research Center for Human Retrovirus Infection, Kumamoto University, 860-0811 Kumamoto, Japan
[2]International Research Center for Medical Sciences (IRCMS), Kumamoto University, 860-0811 Kumamoto, Japan
[3]Laboratory of Systems Virology, Institute for Frontier Life and Medical Sciences, Kyoto University, 606-8507 Kyoto, Japan
[4]Division of Informative Clinical Sciences, Faculty of Medical Sciences, Kumamoto University, 860-0811 Kumamoto, Japan
[5]Department of Medical Physics, Faculty of Life Sciences, Kumamoto University, 860-0811 Kumamoto, Japan
[6]Department of Rare Diseases Research, Institute of Medical Science, St. Marianna University School of Medicine, 211-0063 Kawasaki, Japan
[7]Department of Microbiology, Kansai Medical University, Hirakata, 573-1191 Osaka, Japan
[8]Department of Hematology, Rheumatology and Infectious Disease, Kumamoto University Hospital, 860-0811 Kumamoto, Japan
[9]Department of Hematology, Imamura General Hospital, 890-0064 Kagoshima, Japan
[10]Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, 277-8561 Tokyo, Japan
[11]The Institute of Medical Science Research Hospital, The University of Tokyo, 108-8639 Tokyo, Japan
[12]Graduate School of Medical and Dental Sciences, Kagoshima University, 890-8544 Kagoshima, Japan
[13]These authors contributed equally
[14]Lead Contact
*Correspondence: y-satou@kumamoto-u.ac.jp

https://doi.org/10.1016/j.celrep.2019.09.016

SUMMARY

The retrovirus human T-cell leukemia virus type 1 (HTLV-1) integrates into the host DNA, achieves persistent infection, and induces human diseases. Here, we demonstrate that viral DNA-capture sequencing (DNA-capture-seq) is useful to characterize HTLV-1 proviruses in naturally virus-infected individuals, providing comprehensive information about the proviral structure and the viral integration site. We analyzed peripheral blood from 98 naturally HTLV-1-infected individuals and found that defective proviruses were present not only in patients with leukemia, but also in those with other clinical entities. We further demonstrated that clones with defective-type proviruses exhibited a higher degree of clonal abundance than those with full-length proviruses. The frequency of defective-type proviruses in HTLV-1-infected humanized mice was lower than that in infected individuals, indicating that defective proviruses were rare at the initial phase of infection but preferentially selected during persistent infection. These results demonstrate the robustness of viral DNA-capture-seq for HTLV-1 infection and suggest potential applications for other virus-associated cancers in humans.

INTRODUCTION

It is estimated that around 15% of human cancers are attributed to some viral infections (Plummer et al., 2016). In particular, there are integrated viral genomes detectable in the host cellular DNA in cervical cancer with human papilloma virus (HPV) (Cancer Genome Atlas Research et al., 2017; Hu et al., 2015), hepatocellular carcinoma (HCC) with hepatitis B virus (HBV) (Fujimoto et al., 2012; Sung et al., 2012), and adult T-cell leukemia-lymphoma (ATL) with human T-cell leukemia virus type 1 (HTLV-1), demonstrating the evidence of transformation of the infected cells themselves. HTLV-1 is an exogenous retrovirus endemic in some tropical areas in the world (Gessain and Cassar, 2012; Poiesz et al., 1980). Since HTLV-1 is a retrovirus, the viral RNA genome is reverse transcribed into double-stranded DNA, and the viral DNA is integrated into the host genomic DNA. The integrated virus, called provirus, is transcribed and serves as a template to produce new viral particles. A characteristic of HTLV-1 infection, especially in the chronic phase, is that the virus increases or maintains its copy number not via the production of free viral particles but via the clonal expansion of infected cells. In line with this notion, the proviral sequences are extremely stable during the chronic phase of infection because they are maintained by DNA replication mediated by the DNA polymerase of the host cells, which is much less error-prone compared to the viral reverse transcriptase (Daenke et al., 1990; Van Dooren et al., 2004).

The HTLV-1 genome is approximately 9,000 bp long and has identical long terminal repeat (LTR) sequences at both ends of the provirus (Seiki et al., 1983). The 5′ LTR is the promoter for the transcription in the sense orientation, whereas the 3′ LTR is the promoter for the antisense transcription. Tax, encoded in the pX region in the sense orientation, is a well-characterized viral protein and works as a transactivator of HTLV-1 5′ LTR (Felber et al., 1985). The HTLV-1 bZIP factor (HBZ) is also encoded in the pX region, but in the anti-sense orientation (Gaudray et al., 2002). Both *tax* and *HBZ* are implicated in oncogenesis induced by HTLV-1 (Grossman et al., 1995; Hasegawa et al., 2006; Ma et al., 2016; Satou et al., 2006, 2011).

It is estimated that 10 to 20 million people worldwide are infected with HTLV-1. Although the majority of infected individuals remain asymptomatic, the virus sporadically causes severe diseases, such as ATL, and some inflammatory diseases, such as HTLV-1-associated myelopathy (HAM)/tropical spastic paraparesis (TSP) (Gessain et al., 1985; Osame et al., 1986; Uchiyama et al., 1977; Watanabe, 1997). A classification for ATL patients based on clinical manifestations was proposed in 1991 and classifies ATL into four types: acute, lymphoma, chronic, and smoldering type (Shimoyama, 1991). The acute and lymphoma types are considered aggressive forms, while the chronic and smoldering types follow an indolent clinical course (Katsuya et al., 2015). One criterion for the diagnosis of ATL is the presence of expanded monoclonal infected cells, detected by analyzing viral integration sites (ISs).

HTLV-1 is classified into seven subtypes (a–g) based on the nucleotide diversity in the LTR region (Verdonck et al., 2007). The HTLV-1a subtype, also known as the ‘‘cosmopolitan group,’’ is further divided into five distinct subtypes: Transcontinental, Japanese, West African, North African, and Peruvian Black (Van Dooren et al., 1998; Vidal et al., 1994). It has been reported that the Transcontinental subtype is more frequent than the Japanese subtype among HAM/TSP patients in Japan (Nozuma et al., 2017). Although the HTLV-1 sequence is relatively stable and less prone to mutations, compared to HIV-1, it is a well-known phenomenon that ATL cells frequently contain defective proviruses (Konishi et al., 1984; Manzari et al., 1983). Previous studies reported that there are two types of defective proviruses (Miyazaki et al., 2007; Tamiya et al., 1996). Type 1 defective proviruses contain both 5′ and 3′ LTRs but are missing a part of the proviral sequence between them, while type 2 defective proviruses are lacking the 5′ LTR. Since the 5′ LTR is the promoter of viral sense transcription, type 2 defective proviruses generally lose transcriptional activity of the *tax* gene, which is known as a major target of cytotoxic T-lymphocytes for HTLV-1 (Kannagi et al., 1991). Several previous studies used conventional and high-throughput DNA sequencing techniques to analyze HTLV-1 proviral sequences in infected individuals. Most of the previous reports on defective proviruses analyzed only major clones in ATL patients, while defective proviruses in asymptomatic carriers and HAM/TSP patients have not yet been well characterized.

One limitation of the previous approaches to determine the proviral sequence is the usage of virus-specific primers to amplify the viral sequence. With this conventional method, we cannot obtain the initial sequence of the 5′ LTR or the end of the 3′ LTR. We previously reported that the DNA-probe-based enrichment for retroviral sequence, HIV-1 and HTLV-1, significantly increased the detection sensitivity of retrovirus sequences up to several thousand times more than without enrichment (Miyazato et al., 2016). In that study, we analyzed epigenetic features of the HTLV-1 provirus in HTLV-1-infected cell lines. Another possible application is to characterize the HTLV-1 proviral DNA sequence in a clinical setting, where most of the cells are not infected with HTLV-1, and a very high sensitivity is required to analyze such a scarce target sequence. Here, we applied HTLV-1-targeted DNA-capture sequencing (DNA-capture-seq) to characterize the entire HTLV-1 sequence from the beginning of the 5′ LTR to the end of the 3′ LTR. The protocol gave us additional information, such as the HTLV-1 IS and the type of HTLV-1 provirus structure, because HTLV-1-targeted probes captured not only HTLV-1 fragments, but also chimeric ones containing both HTLV-1 and the human genome. These comprehensive, quantitative, and high-resolution data on HTLV-1 provirus in infected individuals provide fundamental information about persistence and pathogenesis in HTLV-1 infection.

RESULTS

Efficiency of HTLV-1 DNA-Capture-Seq Analysis in Test DNA Samples

We previously demonstrated that DNA-probe-based next-generation sequencing (NGS) library enrichment was useful for HTLV-1 proviral analysis in HTLV-1-infected cell lines (Miyazato et al., 2016; Satou et al., 2016). In the current study, we aimed to evaluate the efficiency of the analytic protocol, especially in clinical materials. The experimental workflow is shown in Figure 1A. Before we analyzed clinical samples, we evaluated the efficacy of the protocol by using test DNA samples. We used an ATL-derived cell line, ED, as a test DNA sample (Maeda et al., 1985). It has been previously reported that the ED cell line contains just one copy of HTLV-1 integrated in the genomic DNA (gDNA) of the host cell (Maeda et al., 1985). We assumed that the detection efficiency of viral sequences in peripheral blood mononuclear cells (PBMCs) of HTLV-1-infected individuals would be much lower than that in cell lines because the average proviral load (PVL) in a clinical setting is only approximately 1% to 2%. Thus, we prepared test DNAs by mixing Jurkat gDNA, an HTLV-1-uninfected T-cell line, with ED gDNA in proportion of 1% or 10%, which we then used to test the efficiency of viral sequence detection by HTLV-1 DNA-capture-seq. We detected a larger number of reads mapped to HTLV-1 in the test DNA sample with 10% ED gDNA than that with 1% ED gDNA, when they were sequenced with the same depth (Figure 1B). The number of HTLV-1 reads correlated with the total sequencing depth. Thus, we tried to estimate how much sequencing depth was required to cover the whole proviral sequence by changing the number of sequencing reads *in silico*. We needed to obtain more than 1,000 HTLV-1 reads to cover the whole viral region in test DNA samples either with 1% or 10% PVL. A total of 500,000 reads were required for the 1% HTLV-1 positive DNA sample, whereas 50,000 total reads were enough for the 10% sample (Figure 1C). This result shows that the requirement of

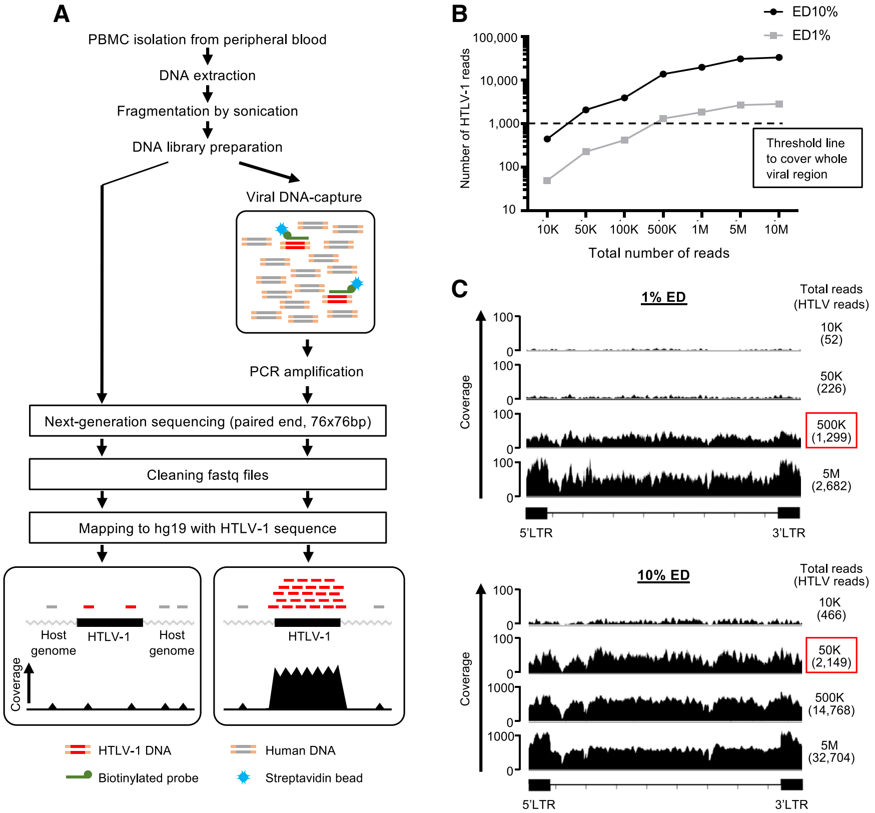


Figure 1. Application of DNA-Capture-Seq to Characterize the HTLV-1 Provirus

(A) Experimental workflow of HTLV-1 DNA-capture-seq for PBMCs from HTLV-1-infected individuals. A DNA-probe-based enrichment step was introduced to increase the detection efficiency of HTLV-1 reads. The schematic figure illustrates the different efficiencies of proviral detection before and after enrichment.

(B) Association between the total number of NGS reads and HTLV-1 reads in the test DNA samples. Jurkat DNA plus 10% and 1% ED DNA, an HTLV-1-infected cell line, were used as test samples. To cover the whole viral sequence, at least 1,000 HTLV-1 mapped reads are required.

(C) HTLV-1 reads obtained from test DNA with 1% and 10% ED gDNA were visualized on integrative genomics viewer (IGV). The numbers of total and HTLV-1 reads are shown on the right side. Red squares indicate the number of reads required to cover a whole proviral sequence.

sequencing depth varies depending on the PVL of the samples we analyze, being higher for lower PVL.

HTLV-1 DNA-Capture-Seq Analysis of PBMCs from HTLV-1-Infected Individuals

We next applied the protocol to the analysis of PBMCs from HTLV-1-infected individuals. We analyzed 98 HTLV-1 naturally infected individuals, including 24 asymptomatic carriers, 29 HAM/TSP cases, and 45 ATL cases. The characteristics of each group and each patient are listed in Table S1. We performed sequencing analysis to obtain a sufficient depth to cover the whole viral sequence (Figure 2A). There were some variations of the HTLV-1 nucleotide sequence among individuals (Figure S1A). On the other hand, there were very few variations within each individual (Figure S1B). Next, we performed phylogenetic tree analysis with each major sequence of all individuals. There were both Japanese and Transcontinental subtypes (Figure 2B). Consistent with previous studies, the frequency of the Transcontinental subtype in HAM/TSP patients was significantly higher than that in asymptomatic carriers or ATL patients (Furukawa et al., 2000; Nozuma et al., 2017) (Figure 2C). It has been reported that there are deletions, insertions, and nonsense mutations in various HTLV-1 genes, except for *HBZ* (Fan et al., 2010). We also identified one case with a nonsense mutation and five cases with a deletion of the *tax* gene among the 45 ATL cases in this study (Table S2).

HTLV-1 DNA-capture-seq efficiently identified defective proviruses at single nucleotide resolution in the ATL cases (Figures 2D and S2). When we analyzed the most dominant clone—which was considered an ATL clone—in each sample, the frequencies of full-length, type 1 defective, and type 2 defective proviruses were 71.1%, 2.2%, and 26.7%, respectively (Figure 2E). When retroviruses integrate into the host cellular genome, short repetitive sequences are generated adjacent to both LTRs by their viral integrase. In the case of HTLV-1, virus integration introduces a 6-bp repeat sequence (Seiki et al., 1983). The 6-bp repeat was present in 8 out of 12 ATL cases with type 2 defective proviruses (Figure 2F). This finding indicated that more than half of the defective proviruses were generated during the process of viral integration, in line with a previous report (Miyazaki et al., 2007). These results demonstrated that the

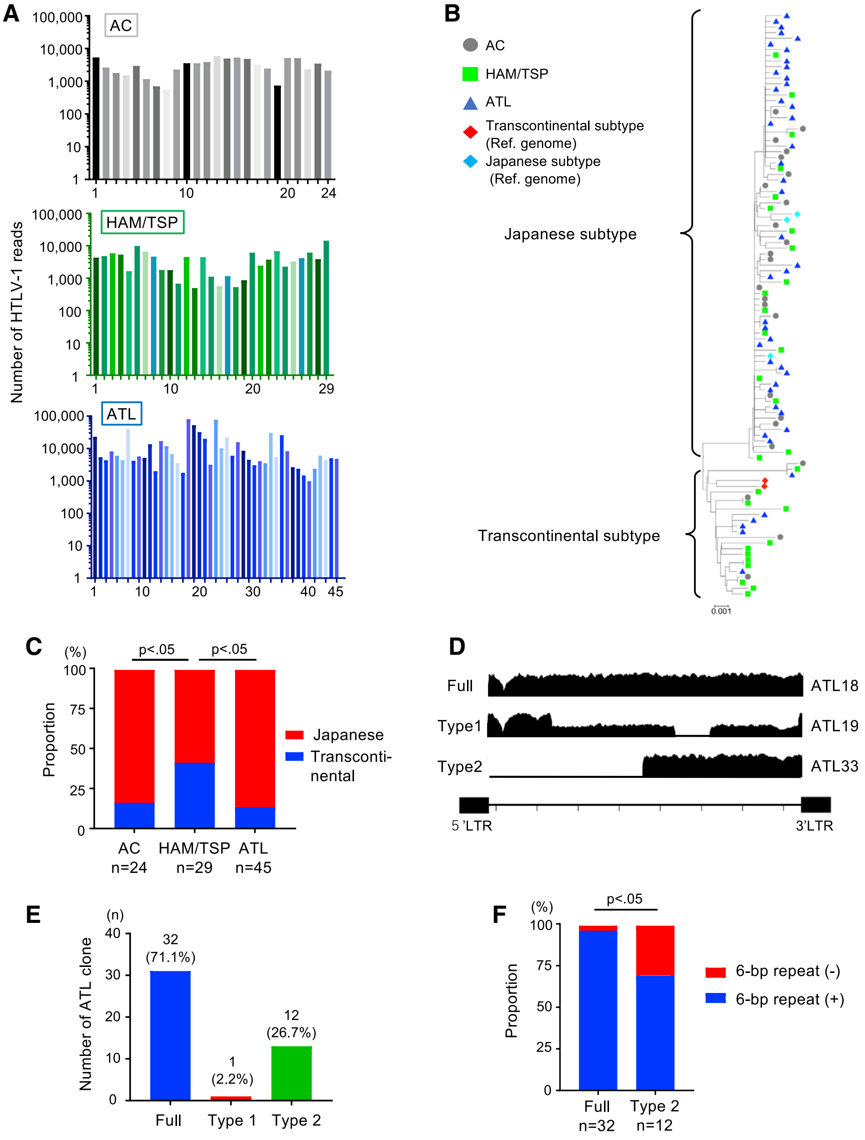


Figure 2. Major Sequence Data Analysis of HTLV-1 Proviruses in PBMCs from the Infected Individuals

(A) The numbers of HTLV-1 reads detected in each infected individual are shown as a bar graph. There were 24 AC, 29 HAM/TSP, and 45 ATL cases enrolled in this study.

(B) Phylogenetic tree analysis was performed with major sequences of each individual.

(C) The proportion of Japanese and Transcontinental subtypes in AC, HAM/TSP, and ATL. The numbers of patients are shown at the bottom.

(D) Representative visualization of aligned reads obtained from ATL patients containing each type of proviral structure. The sample IDs are shown on the right side. Since 5′ and 3′ LTRs are identical and indistinguishable, only HTLV-1 reads mapped to the proviral region between the two LTRs are shown. The type 1 defective provirus contains both 5′ and 3′ LTRs but lacks a part of the proviral region between the two LTRs. The type 2 defective provirus lacks in the 5′ side of the provirus.

(E) The number of dominant clones with full-length, type 1, and type 2 defective proviruses in ATL patients. An ATL clone was defined as the most dominant clone in each ATL patient.

(F) The proportion of ATL clones with or without 6-bp repeat sequences at both ends of the LTRs. The numbers of ATL clones are shown at the bottom.

Statistical significance was analyzed by Fisher's exact test in (C) and (F).

HTLV-1 DNA-capture-seq enables us to characterize the major viral sequence in each infected individual, including HTLV-1 subtypes, genetic changes, and defective proviruses.

Establishment of HTLV-1 IS Analysis by Using HTLV-1 DNA-Capture-Seq

Instead of virus-specific primers, our protocol uses DNA probes covering the entire HTLV-1 in order to capture DNA fragments with viral sequences in DNA libraries prepared from PBMCs (Figure 1A). Therefore, not only viral sequences, but also the flanking host genomic sequences, can be obtained because the DNA probe would hybridize with chimeric amplicons containing both the host cellular and the HTLV-1 genome. The chimeric amplicons generate paired reads containing both virus and host sequences (virus-host reads). In order to establish the viral IS analysis, we used a test DNA sample prepared with the ED cell line, for which we already know the HTLV-1 IS. We also used a negative control sample prepared by mixing DNA containing gDNA of Jurkat T cells and plasmid DNA of the HTLV-1 molecular clone pACH, in which there is no HTLV-1 integration. The result of the test DNA showed that there were plenty of virus-virus and virus-host reads over the HTLV-1 provirus when we used the human genome (hg19) with the integrated HTLV-1 genome as the reference genome for alignment (Figure 3A). In order to see if we could identify the IS without any prior information, we next performed mapping of the same NGS data with hg19 and HTLV-1 as an additional chromosome. A peak of reads was seen clustered around the IS (Figure 3B). Since there was no viral sequence in the reference human genome, only host sequences of virus-host reads were aligned either upstream or downstream of the viral IS. These findings suggested that virus-host reads would be useful to identify unknown ISs in clinical samples. Thus, we generated a program to extract virus-host reads from HTLV-1 DNA-seq data and thereby found that there were not only virus-host reads corresponding to the IS in ED cells, but also many other virus-host reads (Table S3). Since ED is an ATL cell line with monoclonal HTLV-1 integration, virus-host reads different from the viral IS should be experimental artifacts. Virus-host reads derived from the IS were present both upstream and downstream of the

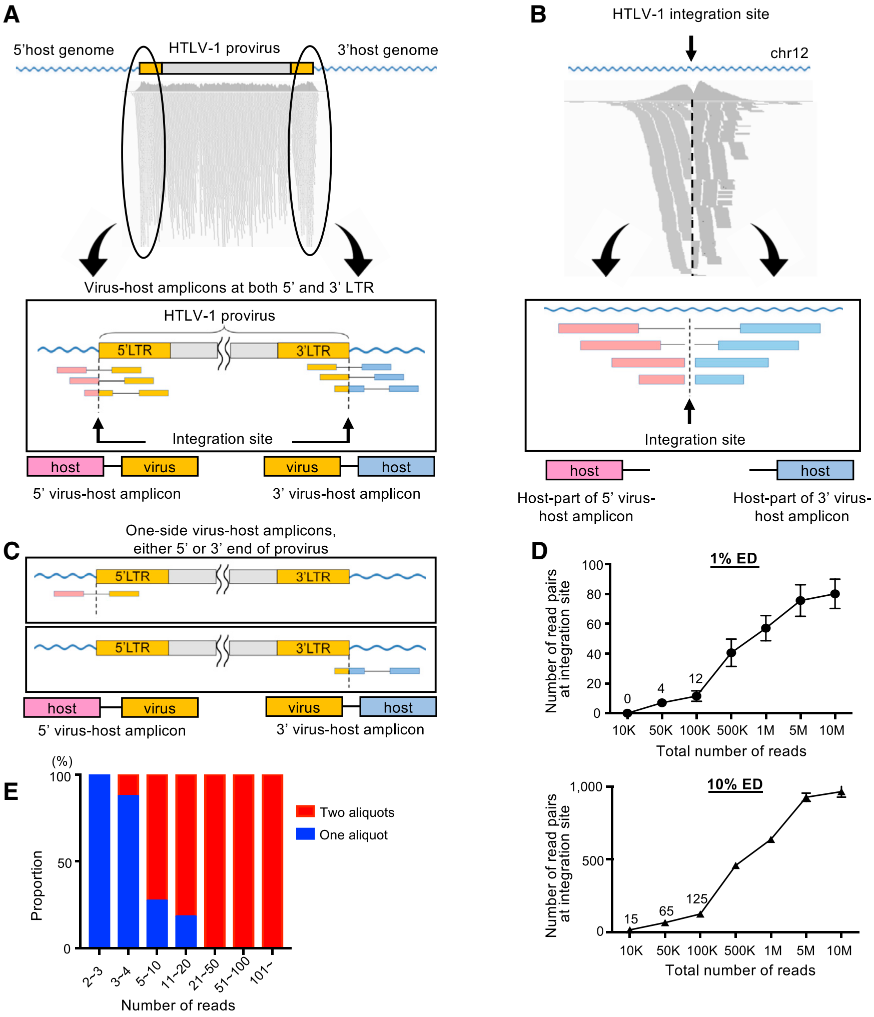


Figure 3. Establishment of HTLV-1 IS Analysis with the DNA-Capture-Seq Data

(A) Visualization of sequencing reads aligned to the HTLV-1 provirus in ED cell line. A schematic figure of the reference genome, integrated HTLV-1, and flanking host genomic region is shown at the top. NGS reads aligned to HTLV-1 are shown below the reference genome. The virus-host chimeric reads are shown in black circles. The viral integration site was defined as the position where virus-host reads were present at both the 5′ and 3′ ends of the provirus.

(B) Visualization of reads derived from ED ISs when NGS data were aligned to the hg19 and HTLV-1 sequence as an additional chromosome. Since there was no viral sequence inserted at the position of the IS, only host sequences derived from virus-host reads were aligned either upstream or downstream of the viral IS. The arrow indicates the location of the IS.

(C) A schematic figure of one-sided virus-host reads generated by experimental artifacts.

(D) Association between the total number of NGS reads and the number of read pairs at ISs (mean value and SD error) in the test DNA samples. Jurkat DNA plus 1% (n = 2) and 10% ED (n = 2) DNA were used as test samples.

(E) Reproducibility of IS analysis by HTLV-1 DNA-capture-seq. The proportions of clones detected in either only one or both aliquots are shown as a cumulative value obtained from PBMCs of five infected individuals.

HTLV-1 IS, but other virus-host reads were not (Table S3). Then we analyzed NGS data from negative control DNA and found that a random ligation between virus and host could be generated during DNA library preparation, resulting in one-sided virus-host reads at either the 5′ or 3′ sides (Figure 3C; Table S4). Thus, in this study, we defined the viral IS as the position where virus-host reads were present at both the 5′ and 3′ ends of the provirus (Figure 3A). To evaluate the sensitivity of the IS detection, we performed the HTLV-1 DNA-capture-seq with mixed gDNA from Jurkat and ED cells. Theoretically, 1% or 10% ED/Jurkat DNA contains one infected clone with the same ISs per 100 cells or 10 cells, respectively. The result demonstrated that 10,000 and 50,000 total sequencing reads were enough to detect the viral IS in 10% or 1% ED/Jurkat DNA, respectively. In other words, the clone expanded to occupy 1% of PBMCs in the infected individuals was detectable by DNA capture-seq if we sequence 50,000 total reads (Figure 3D). Next, we assessed reproducibility of the protocol on detecting viral IS in the analysis of clinical samples. Genomic DNAs from five different infected individuals were divided into two aliquots, and then each aliquot was analyzed by HTLV-1 DNA-capture-seq. There were 128 unique ISs identified, as defined in this study. We found that a HTLV-1 IS with a large number of virus-host reads was more detectable in both aliquots than one with a smaller number of reads was (Figure 3E), indicating that the ISs of minor clones in the sample are stochastically detected, but those of major clones are reproducibly detected in this assay protocol (Figure S3).

HTLV-1 IS Analysis of PBMCs from HTLV-1-Infected Individuals

Using this established method, we performed IS analysis with PBMCs from HTLV-1-infected individuals. Based on the definition of viral IS we used in this study, we identified a total of 207, 306, and 215 unique ISs in asymptomatic carriers, HAM/TSP patients, and ATL patients, respectively. HTLV-1 ISs were distributed broadly in each chromosome (Figure 4A). The frequencies of HTLV-1 ISs inside genes were 42%, 47%, and 47% in asymptomatic carriers, HAM/TSP patients, and ATL patients, respectively, and there were no statistically significant differences among the different clinical entities (Figure 4B). We also analyzed the orientation of the provirus within the host gene. There was no particular tendency in the proportion of the viral orientation relative to the host gene among the three patient cohorts (Figure S4A). We next investigated the epigenetic environments at the IS and found that HTLV-1 ISs disfavored regions

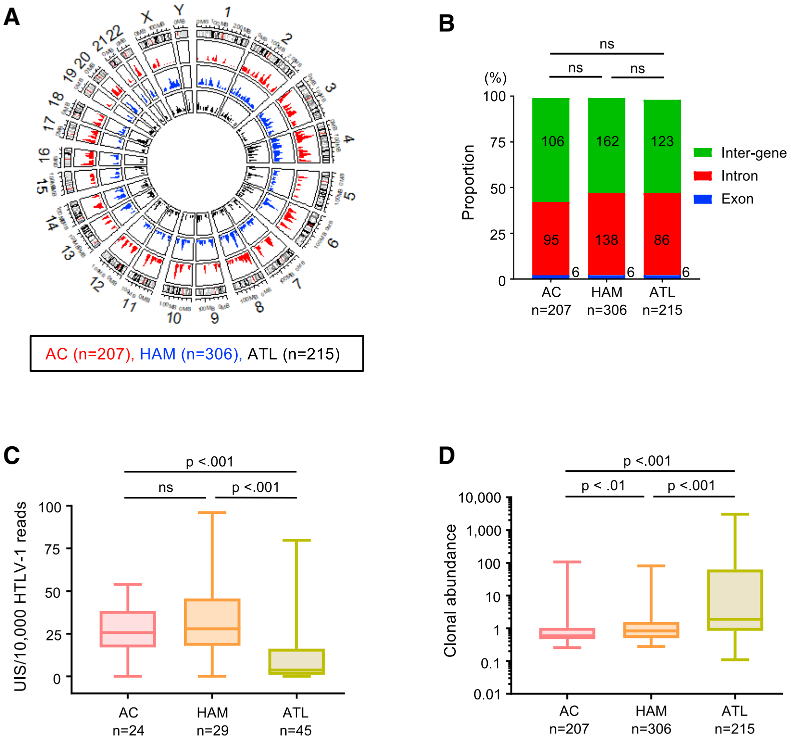


Figure 4. HTLV-1 IS Analysis of PBMCs from HTLV-1-Infected Individuals

(A) The distribution of ISs in each clinical entity is shown in circos plot. The characters on the outermost track represent human chromosomes.

(B) The frequency of ISs within genes or inter-gene in each clinical entity. The numbers of clones analyzed are shown at the bottom, and the number of clones in each category is shown in each fraction within each bar.

(C) The number of unique ISs per 10,000 reads aligned to HTLV-1 from each individual is shown in the box plot. The numbers of patients analyzed are shown at the bottom.

(D) Degrees of abundance of each individual clone are shown in the box plot. The numbers of detected clones are shown at the bottom. Clonal abundance was calculated by considering the proviral load and the number of HTLV-1 reads as follows: clonal abundance of each clone = # of final virus host reads × (10,000 ÷ # of total HTLV1 mapped reads) × {proviral load (%) ÷ 100}. The boxes extend from the 25th to 75th percentiles, and the lines indicate the median. The whiskers go down to the smallest value and up to the largest one.

Statistical significance was obtained by Fisher's exact test (B) and Mann-Whitney U test (C and D).

with H3K36me3, which is a histone mark observed in actively transcribed gene bodies (Figure S4B). There were no significant differences in epigenetic marks, including H3K27ac, a histone mark observed in active promoter and enhancer regions; H3K9me3, a histone mark observed in heterochromatin regions; and H3K36me3 among the different clinical entities and ATL subtypes (Figures S4B and S4C). The number of unique ISs per 10,000 HTLV-1 reads was high in asymptomatic carriers or HAM/TSP, but low in ATL cases, likely due to the presence of clonally expanded ATL clones (Figure 4C). We next evaluated the clonal abundance of each HTLV-1-infected cell based on the number of virus-host reads at each IS (see Methods Details). As expected, the clonal abundance in ATL patients was the highest (Figure 4D). The clonal abundance in HAM/TSP cases was significantly higher than that in asymptomatic carriers. There was a negative correlation between the PVL and the number of unique ISs per 10,000 HTLV-1 reads in HAM/TSP and ATL but not in asymptomatic carriers, also suggesting the presence of expanded clones in HAM/TSP and ATL patients (Figure S4D). These results demonstrated that HTLV-1 DNA-capture-seq can be efficiently applied for clonality analysis of clinical samples as well as the whole viral sequence analysis.

Structural Analyses of the HTLV-1 Provirus and Its Association with Clonal Expansion of the Host Cell

There are several previous reports describing defective proviruses in ATL patients (Konishi et al., 1984; Miyazaki et al., 2007; Tamiya et al., 1996), but none using high-throughput sequencing technology or characterizing defective proviruses in asymptomatic carriers or HAM/TSP patients. As shown in Figure 3A, HTLV-1 DNA-capture-seq provides reads from virus-host amplicons, suggesting that the protocol would enable us to obtain information about proviral DNA structure. Viral sequences of virus-host reads originating from full-length type are theoretically mapped to 5′ or 3′ LTRs (Figure 5A, top panel). Viral sequences of 5′-side virus-host reads derived from type 2 defective proviruses are likely to be mapped to a proviral region downstream of the 5′ LTR (Figure 5A, middle panel). Thus, we defined this provirus as 5′-defective type. If viral sequences of 3′-side virus-host reads are aligned to the provirus upstream of the 3′ LTR, the provirus is defined as 3′-defective type (Figure 5A, bottom panel). We analyzed the proportion of defective proviruses in each clinical entity. High frequencies of defective proviruses were observed in asymptomatic carriers (18.8%) and HAM/TSP (17.3%) as well as ATL (30.2%), showing these are not a specific phenomenon observed only in ATL, but a general feature in HTLV-1-infected individuals (Figure 5B). We analyzed the frequency of HTLV-1 integration within the host genes and found that there were no statistically significant differences between full-length and defective proviruses (Figure S5A). There seems to be no obvious hotspot genes for viral ISs in 5′-defective proviruses (Figure S5B). There was a wide range distribution of deleted regions in defective proviruses detected in the ATL cases we analyzed (Figure 5C). An advantage of HTLV-1 DNA-capture-seq is that we can simultaneously identify proviral structure and the degree of clonal expansion of each individual HTLV-1-infected clone. Thus, we aimed to analyze the relationship between the proviral structure and the clonal abundance of each infected cell. The results showed that HTLV-1-infected clones with defective proviruses exhibited a higher degree of clonal expansion than those with full-length type (Figure 5D). There was a statistically significant difference in cumulative analyses of all data, although no statistically significant difference

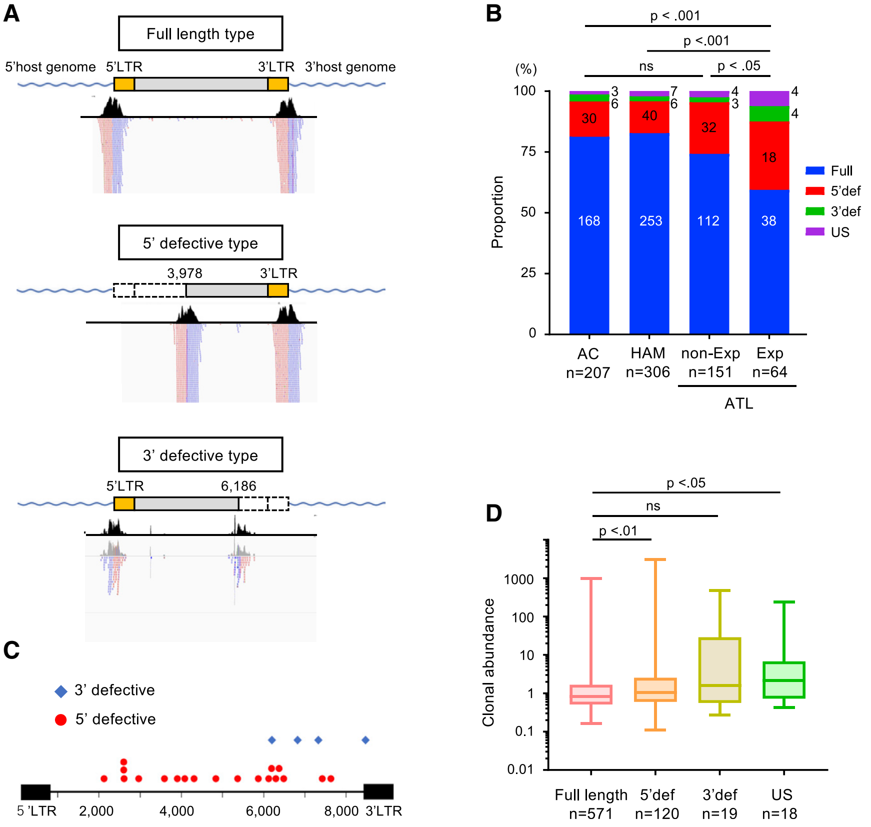


Figure 5. Structural Analyses of the HTLV-1 Provirus and Its Association with Clonal Expansion of the Host Cell

(A) Representative visualization patterns of virus-host reads derived from full-length type proviruses, 5′-defective type proviruses, and 3′-defective type proviruses are shown at the top, middle, and bottom panels, respectively.

(B) Proportion of each proviral type in different clinical entities and the number of each type of proviruses are shown in each fraction within each bar. An expanded clone in ATL cases was defined as one with more than 10 in the clonal abundance.

(C) Distribution of defective sites in 5′- and 3′-defective proviruses detected in ATL cases.

(D) Degree of clonal abundance of each type of provirus. The cumulative results from all cases enrolled in this study are shown.

The numbers of detected clones in each group are shown at the bottom (B and D). Statistical significance was obtained by Fisher's exact test (B) and Mann-Whitney U test (D).

5′ def, 5′-defective type provirus; 3′ def, 3′-defective type provirus; Exp, expanded clone; US, unspecified type provirus.

was observed in the individual group analysis of asymptomatic carriers and ATL cases (Figure S5C).

HTLV-1 DNA-Capture-Seq Analysis of Infected Cells *In Vitro* and in a Humanized Mouse Model

Most infected individuals enrolled in this study are thought to have a long history as HTLV-1 carriers. Thus, a key question that remained to be answered was when and how the high proportion of 5′-defective proviruses was generated during the long course of HTLV-1 infection. To explore this point, we infected Jurkat T cells with HTLV-1, cultivated them for 2, 8, or 16 weeks *in vitro*, and then analyzed their proviruses by HTLV-1 DNA-capture-seq (Figure 6A). There was no significant increase in the proportion of 5′-defective type provirus during *in vitro* cultivation at the different time points (Figure 6B). The infected cells with 5′-defective type proviruses tended to expand more than those with full-length type proviruses (Figure 6C), but the PVL of the infected Jurkat cells was extremely high—338.3%, 155.0%, and 127.0% at 2, 8, and 16 weeks, which is far different from HTLV-1-infected cells *in vivo*. Therefore, we next analyzed an *in vivo* mouse model of HTLV-1 infection (Tezuka et al., 2014). Humanized mice were infected with HTLV-1 by injecting irradiated Jurkat T cells containing full-length HTLV-1 proviruses and maintained for a period of 3 to 12 weeks. We performed HTLV-1 DNA-capture-seq analysis at around 3, 5, 8, and 12 weeks post infection (Figure 6D). The HTLV-1 PVL in these samples ranged from 0.5% to 118.9% (median, 68.7%). The proportion of 5′-defective type was not high at 3 weeks post infection (Figure 6E), suggesting that 5′-defective proviruses were not generated at the initial phase of infection. Next, we analyzed the degree of clonal abundance in each type of proviral structure and found no significant difference between full-length and 5′-defective types of provirus (Figure 6F).

DISCUSSION

In contrast to HIV-1 infection, there is little viremia in the peripheral blood of HTLV-1-infected individuals, even in the absence of anti-retroviral treatment. The viral persistence of HTLV-1 is maintained not by *de novo* infection through viral particles, but by the long life span and/or clonal proliferation of virus-infected cells. As a consequence, approximately 5% of infected individuals develop ATL after a long latent period of time (Bangham and Matsuoka, 2017). Thus, characterization of the HTLV-1 provirus with high resolution is fundamental to understanding the viral persistence and pathogenesis. There are various methods used to characterize HTLV-1 proviral DNA, such as conventional PCR and sequencing technology; modified PCR protocols including inverse long PCR (Takemoto et al., 1994) and ligation-mediated PCR (LM-PCR) (Meekings et al., 2008); and, more recently, high-throughput sequencing technology (Gillet et al., 2011; Rosewick et al., 2017). In this study, we applied the HTLV-1 DNA-capture-seq methodology that we developed recently (Miyazato et al., 2016) to the characterization of the proviruses in PBMCs of the infected individuals.

Here, we demonstrated that the protocol is useful to obtain the entire proviral sequence, proviral structure, and viral IS. There are both advantages and disadvantages of DNA-capture-seq

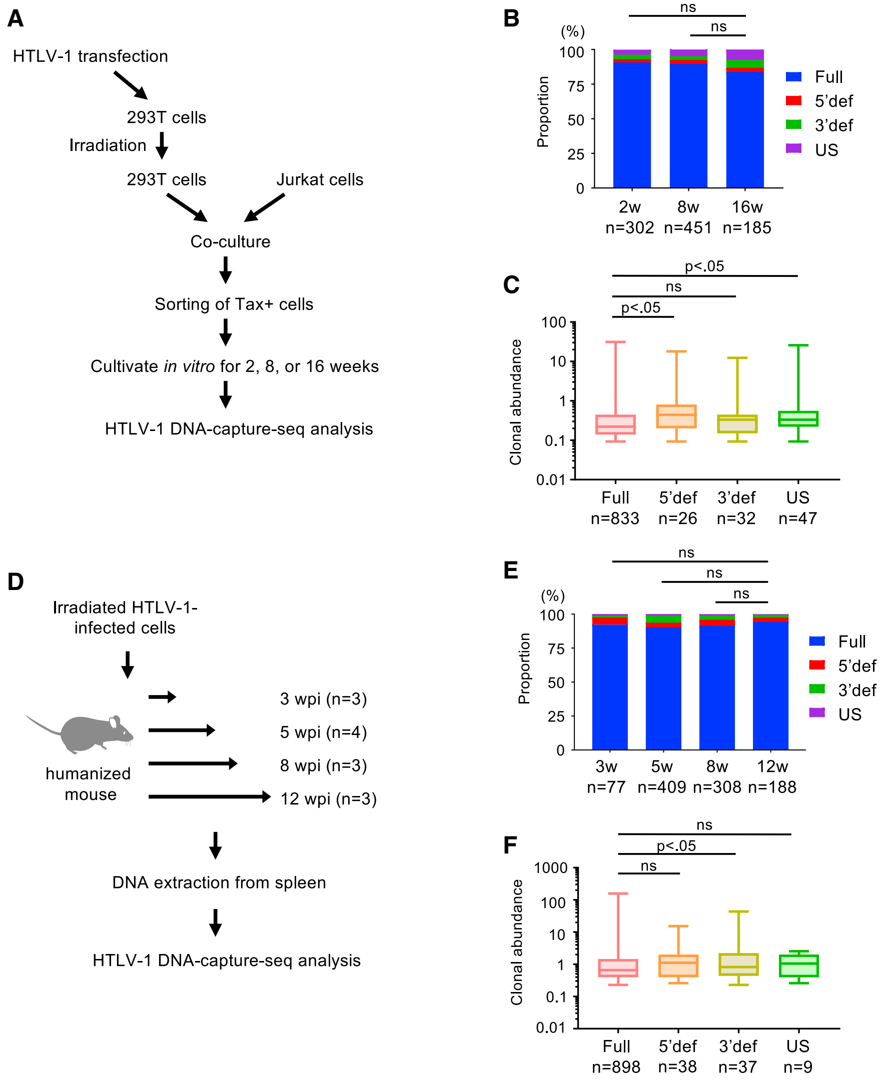


Figure 6. Proviral Structure and Clonal Abundance of HTLV-1-Infected Cells in an *In Vitro* and in a Humanized Mouse Model

(A) Experimental flow of HTLV-1 DNA-capture-seq with Jurkat T cells infected with HTLV-1 *in vitro*.

(B) Proportion of each proviral type at different time points during *in vitro* cultivation. The numbers of clones observed are shown at the bottom.

(C) Degree of clonal abundance of each type of provirus. The numbers of clones observed are shown at the bottom.

(D) Experimental flow of HTLV-1 DNA-capture-seq analysis using the humanized mouse model.

(E) Proportion of each proviral type at different time points. The numbers of detected clones are shown at the bottom.

(F) Degree of clonal abundance in each type of provirus. Statistical significance was obtained by Fisher's exact test in (B and E) and Mann-Whitney U test (C and F).

compared to NGS-based conventional methods for proviral analysis (Table 1). When compared with conventional deep-seq that uses virus-specific primers, one advantage of the HTLV-1 DNA-capture-seq is that we can obtain the entire HTLV-1 sequence from the beginning of the 5′ LTR to the end of the 3′ LTR by virtue of the DNA probes that capture not only viral fragments, but also virus-host chimeric ones (Figure 3A). On the other hand, when we use virus-specific primers targeting the beginning and the end of HTLV-1, only the LTR portion would be amplified by the PCR because they are identical in DNA sequence. To analyze HTLV-1 ISs and quantify the degree of clonal expansion of virus-infected cells, LM-PCR is a well-established experimental approach (Firouzi et al., 2014; Gillet et al., 2011; Rosewick et al., 2017). LM-PCR amplifies virus-host junctions by using two primers targeting the viral LTR and the linker. The LM-PCR method is useful and sensitive to detect ISs (Figure S3), but that does not give us information about viral sequences between LTRs, in contrast to HTLV-1 DNA-capture-seq. We were able to detect expanded clones reproducibly by HTLV-1 DNA-capture-seq (Figure 3D), so the sensitivity was enough to detect ISs of dominant clones, like ATL clones (Figure S3). As discussed, various approaches are available for

Table 1. Characteristics of Different NGS-Based Methods for HTLV-1 Proviral Analysis

	Conventional Deep-Sequencing with Virus-Specific Primers	LM-PCR + NGS	Viral DNA-Capture-Seq
Detection sensitivity of viral sequence	high	NA	moderate
Availability of entire viral sequence	NA	NA	applicable
Susceptibility to sequence variation	high	high	low
Proviral structure analysis	NA for type 2 defective provirus	NA	applicable
Clonality analysis	NA	applicable (high sensitivity)	applicable (moderate sensitivity)
Integration site analysis	NA	applicable (high sensitivity)	applicable (moderate sensitivity)

LM-PCR, ligation-mediated PCR; NGS, next-generation sequencing; NA, not applicable

analysis of the HTLV-1 provirus; therefore, selecting the optimal method depends on research interest and the purpose of each study (Table 1).

Several previous studies have revealed the presence of defective proviruses in ATL cells using the conventional PCR-based approach. However, there were few studies on defective proviruses in asymptomatic carrier (AC) and HAM/TSP patients because of technical limitations. Here, we combined more than 100 tiling DNA probes covering the whole HTLV-1 with NGS, which enabled us to evaluate defective proviruses in a quantitative and comprehensive manner. We analyzed the frequency of defective proviruses in each clinical entity and demonstrated that defective proviruses were not only present in ATL clones, but also frequently observed in asymptomatic carriers and HAM/TSP patients, suggesting that defective proviruses are a general feature of HTLV-1-infected clones *in vivo*. In addition to type 1 and type 2 defective proviruses, we detected a defective provirus without the 3′ LTR. These findings revealed that the nature of the HTLV-1 provirus in naturally infected individuals is more complex than previously thought.

The HTLV-1 DNA-capture-seq provides us with information on both the proviral DNA structure and the degree of clonal expansion of each infected clone. Many researchers in the HTLV-1 research field assume that 5′-defective proviruses, which could not produce viral antigens encoded in sense orientation, should allow the infected cell to escape from the host immune surveillance, resulting in clonal expansion of the infected cells. However, there is no clear evidence to support this assumption. We therefore asked if defective proviruses actually confer an advantage for host cell survival *in vivo*. We found that infected clones with defective proviruses exhibited higher clonal abundance than those with full-length type. Furthermore, we aimed to elucidate when and how the high frequencies of 5′-defective proviruses were generated by analyzing HTLV-1-infection in *in vitro* and *in vivo* models. Both *in vitro* and *in vivo* experiments indicated that the high frequencies of 5′-defective proviruses were not generated during the initial phase of infection. There was no significant increase of 5′-defective proviruses during the persistent infection *in vitro*, at least during the observation period in this study, suggesting that clonal expansion of infected cells with 5′-defective proviruses might not be induced in a cell-intrinsic manner. In addition, we did not detect an accumulation of 5′-defective proviruses during persistent infection in a humanized mouse model. It has been reported that there is an anti-viral immune response in the mice we used in this study (Tezuka et al., 2014), yet the activity might not be strong enough to put selective pressure on the survival of various infected clones. These findings collectively suggested that HTLV-1-infected clones harboring 5′-defective proviruses might have a long life span because they escape from the host immune surveillance, thereby having more chances to accumulate genetic and epigenetic abnormalities. This may explain why we observed the high frequency of 5′-defective proviruses in ATL clones (Figure 5B).

It was clear that 5′-defective proviruses are associated with clonal expansion of the infected cell, but there was a wide distribution of the degree of clonal abundance within each type of proviral structure (Figure 5D). This finding indicates that the structure of HTLV-1 is one of the determinants of clonal abundance of the infected cells, but other factors—such as genetic and epigenetic changes of the host genome and status of the host immunity—may also contribute (Fujikawa et al., 2016; Kataoka et al., 2015; Nosaka et al., 2000). There is substantial evidence supporting the idea that an anti-viral immune response is related to the efficiency of some therapeutic approaches for ATL. ATL patients with long-term remission after allogeneic hematopoietic cell transplantation harbor high frequencies of anti-Tax-specific cytotoxic T lymphocyte (CTL) (Harashima et al., 2004). The immunotherapy with Tax peptide-pulsed dendritic cells exhibited efficacy in some ATL patients and is currently under clinical trial (Suehiro et al., 2015). Furthermore, immunomodulatory agents, such as lenalidomide and anti-CCR4 antibodies, are now approved for the treatment of ATL patients (Ishida et al., 2016, 2017). Therefore, it will be useful to obtain information not only related to the host genome, but also to the HTLV-1 proviral genome to establish a standardized treatment strategy based on the molecular characteristics of ATL cells. For example, ATL clones with defective proviruses may lose viral antigen expression and thereby become insensitive to immunotherapy aiming to enhance the immune response against viral antigens. Thanks to recent advances in DNA sequencing technology, methods such as whole-exome sequencing of human genomes would be useful in the management of ATL patients to predict prognosis and determine an optimal treatment strategy in the near future (Kataoka et al., 2018). Since the HTLV-1

DNA-capture-seq is based on a method similar to whole-exome sequencing, one will be able to analyze the host and proviral genomes simultaneously by combining DNA probes for the host exonic regions and HTLV-1 provirus.

There are several oncogenic viruses with features similar to HTLV-1, such as HPV in cervical cancer (Cancer Genome Atlas Research et al., 2017; Hu et al., 2015) and HBV in HCC (Fujimoto et al., 2012; Sung et al., 2012). Since an integrated virus is present in the host cellular genome, we can utilize the viral IS to provide quantitative information regarding the clonal abundance of the host cells to follow cancer evolution. In this study, we demonstrated that viral DNA-capture-seq was useful to obtain the viral sequence, the viral IS, and the degree of clonal expansion. This method could be applied for the molecular diagnosis of HPV-associated cervical cancer and HBV-associated HCC.

In summary, we have elucidated the nature of HTLV-1 proviral DNA in naturally infected individuals by using HTLV-1 DNA-capture-seq. The result has provided fundamental information on HTLV-1 infection for both basic and clinical research. The viral DNA-capture-seq would be useful to make further progress in HTLV-1 research as well as in other oncogenic viral infections in humans.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Patients Samples
  - ○ Cell lines
  - ○ Humanized mice
- METHOD DETAILS
  - ○ Generation of HTLV-1-infected cells *in vitro*
  - ○ Proviral load measurement by droplet digital PCR
  - ○ Library synthesis, and proviral DNA-capture-seq
  - ○ High-throughput sequencing data analysis
  - ○ Ligation-mediated (LM)-PCR
  - ○ Integration site, proviral structure, and clonal abundance analysis with the DNA-seq data
  - ○ Phylogenetic Analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND CODE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.celrep.2019.09.016.

ACKNOWLEDGMENTS

We thank Michi Miura for the R program to perform the quality check of index reads, Ms. Michiyo Tokunaga for experimental support, Lee Ratner for pACH plasmid, David Derse for pX1 plasmid, and JSPFAD for providing clinical samples. This work was supported by the JSPS KAKENHI (JP26461428, JP17K09016, JP18KK0230, and JP16H06277 to Y.S.; JP16K19580, JP16KK0206, and JP18K16122 to H.K.; JP18K08437 to P.M.; and JP16H06277 to T.W.); Practical Research Project for Rare or Intractable Diseases of the AMED (JP16ek0109026, JP17ek0109083, and JP18ek0109346 to Y.Y.); the AMED (JP19fm0208012 to Y.S.); the Grant for Joint Research Project of the Institute of Medical Science, the University of Tokyo to Y.S.; Kumamoto University Excellent Research Projects to Y.S.; and JST MIRAI to Y.S. The funder has no role in the design of the study, collection of data, its interpretation, or the discussion to submit the work for publication.

AUTHOR CONTRIBUTIONS

Study Conception, Y.S.; Methodology and Formal Analysis, H.K., S.I., and Y.S.; Software, B.J.Y.T., J.I., and Y.U.; Investigation, H.K., S.I., B.J.Y.T., P.M., Y.I., M.T., S.C.I., M.M., N.A., and T.U.; Resources, T.S., N.Y., N.A., T.U., J.F., K.N., M.T., M.Y., T.W., K.U., A.U., and Y.Y.; Data Curation, H.K. and Y.S.; Writing – Original Draft, H.K., S.I., and Y.S.; Writing – Review and Editing, H.K., S.I., B.J.Y.T., J.I., P.M., M.M., S.C.I., Y.U., H.H., T.S., N.A., T.U., J.F., K.N., M.T., M.Y., T.W., K.U., A.U., Y.Y., and Y.S.; Supervision, H.H. and Y.S.; Project Administration and Funding Acquisition, H.K., P.M., T.W., Y.Y. and Y.S.

DECLARATION OF INTERESTS

K.U. has received research funding from Daiichi Sankyo Co., Ltd. Y.Y. has received research funding from Kyowa Hakko Kirin Co., Ltd., outside the submitted work, and has a patent under the title “MEDICINE FOR TREATING OR PREVENTING HTLV-I-RELATED MYELOPATHY, AND METHOD FOR FORECASTING THE EFFECT OF ANTIBODY THERAPY FOR PATIENT OF HTLV-I-RELATED MYELOPATHY” (Japan patent 5552630) licensed to St. Marianna University School of Medicine, and another under “PREVENTIVE OR THERAPEUTIC AGENT FOR HTLV-1-ASSOCIATED MYELOPATHY USING LOW-DOSE ANTI-CCR4 ANTIBODY” (Japan patent 6430082), licensed to St. Marianna University School of Medicine and Kyowa Hakko Kirin Co., Ltd.. A patent describing the therapeutic method and medicament for htlv-1 associated myelopathy (ham) PCT/JP2013/068296 is pending. All other authors declare no competing interests.

Received: March 6, 2019
Revised: June 13, 2019
Accepted: September 6, 2019
Published: October 15, 2019

REFERENCES

Bangham, C.R.M., and Matsuoka, M. (2017). Human T-cell leukaemia virus type 1: parasitism and pathogenesis. Philos. Trans. R. Soc. Lond. B Biol. Sci. *372*, 20160272.

Berry, C.C., Gillet, N.A., Melamed, A., Gormley, N., Bangham, C.R., and Bushman, F.D. (2012). Estimating abundances of retroviral insertion sites from DNA fragment length data. Bioinformatics *28*, 755–762.

Cancer Genome Atlas Research Network; Albert Einstein College of Medicine; Analytical Biological Services; Barretos Cancer Hospital; Baylor College of Medicine; Beckman Research Institute of City of Hope; Buck Institute for Research on Aging; Canada’s Michael Smith Genome Sciences Centre; Harvard Medical School; and Helen F. Graham Cancer Center &Research Institute at Christiana Care Health Services, et al. (2017). Integrated genomic and molecular characterization of cervical cancer. Nature *543*, 378–384.

Consortium, E.P.; ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Daenke, S., Nightingale, S., Cruickshank, J.K., and Bangham, C.R. (1990). Sequence variants of human T-cell lymphotropic virus type I from patients with tropical spastic paraparesis and adult T-cell leukemia do not distinguish neurological from leukemic isolates. J. Virol. *64*, 1278–1282.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Fan, J., Ma, G., Nosaka, K., Tanabe, J., Satou, Y., Koito, A., Wain-Hobson, S., Vartanian, J.P., and Matsuoka, M. (2010). APOBEC3G generates nonsense mutations in human T-cell leukemia virus type 1 proviral genomes *in vivo*. J. Virol. *84*, 7278–7287.

Felber, B.K., Paskalis, H., Kleinman-Ewing, C., Wong-Staal, F., and Pavlakis, G.N. (1985). The pX protein of HTLV-I is a transcriptional activator of its long terminal repeats. Science *229*, 675–679.

Firouzi, S., López, Y., Suzuki, Y., Nakai, K., Sugano, S., Yamochi, T., and Watanabe, T. (2014). Development and validation of a new high-throughput method to investigate the clonality of HTLV-1-infected cells based on provirus integration sites. Genome Med. *6*, 46.

Fujikawa, D., Nakagawa, S., Hori, M., Kurokawa, N., Soejima, A., Nakano, K., Yamochi, T., Nakashima, M., Kobayashi, S., Tanaka, Y., et al. (2016). Polycomb-dependent epigenetic landscape in adult T-cell leukemia. Blood *127*, 1790–1802.

Fujimoto, A., Totoki, Y., Abe, T., Boroevich, K.A., Hosoda, F., Nguyen, H.H., Aoki, M., Hosono, N., Kubo, M., Miya, F., et al. (2012). Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. Nat. Genet. *44*, 760–764.

Furukawa, Y., Yamashita, M., Usuku, K., Izumo, S., Nakagawa, M., and Osame, M. (2000). Phylogenetic subgroups of human T cell lymphotropic virus (HTLV) type I in the tax gene and their association with different risks for HTLV-I-associated myelopathy/tropical spastic paraparesis. J. Infect. Dis. *182*, 1343–1349.

Furuta, R., Yasunaga, J.I., Miura, M., Sugata, K., Saito, A., Akari, H., Ueno, T., Takenouchi, N., Fujisawa, J.I., Koh, K.R., et al. (2017). Human T-cell leukemia virus type 1 infects multiple lineage hematopoietic cells *in vivo*. PLoS Pathog. *13*, e1006722.

Gaudray, G., Gachon, F., Basbous, J., Biard-Piechaczyk, M., Devaux, C., and Mesnard, J.M. (2002). The complementary strand of the human T-cell leukemia virus type 1 RNA genome encodes a bZIP transcription factor that down-regulates viral transcription. J. Virol. *76*, 12813–12822.

Gessain, A., and Cassar, O. (2012). Epidemiological Aspects and World Distribution of HTLV-1 Infection. Front. Microbiol. *3*, 388.

Gessain, A., Barin, F., Vernant, J.C., Gout, O., Maurs, L., Calender, A., and de Thé, G. (1985). Antibodies to human T-lymphotropic virus type-I in patients with tropical spastic paraparesis. Lancet *2*, 407–410.

Gillet, N.A., Malani, N., Melamed, A., Gormley, N., Carter, R., Bentley, D., Berry, C., Bushman, F.D., Taylor, G.P., and Bangham, C.R. (2011). The host genomic environment of the provirus determines the abundance of HTLV-1-infected T-cell clones. Blood *117*, 3113–3122.

Grossman, W.J., Kimata, J.T., Wong, F.H., Zutter, M., Ley, T.J., and Ratner, L. (1995). Development of leukemia in mice transgenic for the tax gene of human T-cell leukemia virus type I. Proc. Natl. Acad. Sci. USA *92*, 1057–1061.

Harashima, N., Kurihara, K., Utsunomiya, A., Tanosaki, R., Hanabuchi, S., Masuda, M., Ohashi, T., Fukui, F., Hasegawa, A., Masuda, T., et al. (2004). Graft-versus-Tax response in adult T-cell leukemia patients after hematopoietic stem cell transplantation. Cancer Res. *64*, 391–399.

Hasegawa, H., Sawa, H., Lewis, M.J., Orba, Y., Sheehy, N., Yamamoto, Y., Ichinohe, T., Tsunetsugu-Yokota, Y., Katano, H., Takahashi, H., et al. (2006). Thymus-derived leukemia-lymphoma in mice transgenic for the Tax gene of human T-lymphotropic virus type I. Nat. Med. *12*, 466–472.

Hu, Z., Zhu, D., Wang, W., Li, W., Jia, W., Zeng, X., Ding, W., Yu, L., Wang, X., Wang, L., et al. (2015). Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism. Nat. Genet. *47*, 158–163.

Ishida, T., Fujiwara, H., Nosaka, K., Taira, N., Abe, Y., Imaizumi, Y., Moriuchi, Y., Jo, T., Ishizawa, K., Tobinai, K., et al. (2016). Multicenter Phase II Study of Lenalidomide in Relapsed or Recurrent Adult T-Cell Leukemia/Lymphoma: ATLL-002. J. Clin. Oncol. *34*, 4086–4093.

Ishida, T., Utsunomiya, A., Jo, T., Yamamoto, K., Kato, K., Yoshida, S., Takemoto, S., Suzushima, H., Kobayashi, Y., Imaizumi, Y., et al. (2017). Mogamulizumab for relapsed adult T-cell leukemia-lymphoma: Updated follow-up analysis of phase I and II studies. Cancer Sci. *108*, 2022–2029.

Kannagi, M., Harada, S., Maruyama, I., Inoko, H., Igarashi, H., Kuwashima, G., Sato, S., Morita, M., Kidokoro, M., Sugimoto, M., et al. (1991). Predominant recognition of human T cell leukemia virus type I (HTLV-I) pX gene products by human CD8+ cytotoxic T cells directed against HTLV-I-infected cells. Int. Immunol. *3*, 761–767.

Kataoka, K., Nagata, Y., Kitanaka, A., Shiraishi, Y., Shimamura, T., Yasunaga, J., Totoki, Y., Chiba, K., Sato-Otsubo, A., Nagae, G., et al. (2015). Integrated molecular analysis of adult T cell leukemia/lymphoma. Nat. Genet. *47*, 1304–1315.

Kataoka, K., Iwanaga, M., Yasunaga, J.I., Nagata, Y., Kitanaka, A., Kameda, T., Yoshimitsu, M., Shiraishi, Y., Sato-Otsubo, A., Sanada, M., et al. (2018). Prognostic relevance of integrated genetic profiling in adult T-cell leukemia/lymphoma. Blood *131*, 215–225.

Katsuya, H., Ishitsuka, K., Utsunomiya, A., Hanada, S., Eto, T., Moriuchi, Y., Saburi, Y., Miyahara, M., Sueoka, E., Uike, N., et al.; ATL-Prognostic Index Project (2015). Treatment and survival among 1594 patients with ATL. Blood *126*, 2570–2577.

Konishi, H., Kobayashi, N., and Hatanaka, M. (1984). Defective human T-cell leukemia virus in adult T-cell leukemia patients. Mol. Biol. Med. *2*, 273–283.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol. Biol. Evol. *33*, 1870–1874.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Ma, G., Yasunaga, J., and Matsuoka, M. (2016). Multifaceted functions and roles of HBZ in HTLV-1 pathogenesis. Retrovirology *13*, 16.

Maeda, M., Shimizu, A., Ikuta, K., Okamoto, H., Kashihara, M., Uchiyama, T., Honjo, T., and Yodoi, J. (1985). Origin of human T-lymphotrophic virus I-positive T cell lines in adult T cell leukemia. Analysis of T cell receptor gene rearrangement. J. Exp. Med. *162*, 2169–2174.

Manzari, V., Wong-Staal, F., Franchini, G., Colombini, S., Gelmann, E.P., Oroszlan, S., Staal, S., and Gallo, R.C. (1983). Human T-cell leukemia-lymphoma virus (HTLV): cloning of an integrated defective provirus and flanking cellular sequences. Proc. Natl. Acad. Sci. USA *80*, 1574–1578.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. *28*, 495–501.

Meekings, K.N., Leipzig, J., Bushman, F.D., Taylor, G.P., and Bangham, C.R. (2008). HTLV-1 integration into transcriptionally active genomic regions is associated with proviral expression and with HAM/TSP. PLoS Pathog. *4*, e1000027.

Mitchell, M.S., Bodine, E.T., Hill, S., Princler, G., Lloyd, P., Mitsuya, H., Matsuoka, M., and Derse, D. (2007). Phenotypic and genotypic comparisons of human T-cell leukemia virus type 1 reverse transcriptases from infected T-cell lines and patient samples. J. Virol. *81*, 4422–4428.

Miyazaki, M., Yasunaga, J., Taniguchi, Y., Tamiya, S., Nakahata, T., and Matsuoka, M. (2007). Preferential selection of human T-cell leukemia virus type 1 provirus lacking the 5′ long terminal repeat during oncogenesis. J. Virol. *81*, 5714–5723.

Miyazato, P., Katsuya, H., Fukuda, A., Uchiyama, Y., Matsuo, M., Tokunaga, M., Hino, S., Nakao, M., and Satou, Y. (2016). Application of targeted enrichment to next-generation sequencing of retroviruses integrated into the host human genome. Sci. Rep. *6*, 28324.

Nosaka, K., Maeda, M., Tamiya, S., Sakai, T., Mitsuya, H., and Matsuoka, M. (2000). Increasing methylation of the CDKN2A gene is associated with the progression of adult T-cell leukemia. Cancer Res. *60*, 1043–1048.

Nozuma, S., Matsuura, E., Kodama, D., Tashiro, Y., Matsuzaki, T., Kubota, R., Izumo, S., and Takashima, H. (2017). Effects of host restriction factors and the HTLV-1 subtype on susceptibility to HTLV-1-associated myelopathy/tropical spastic paraparesis. Retrovirology *14*, 26.

Osame, M., Usuku, K., Izumo, S., Ijichi, N., Amitani, H., Igata, A., Matsumoto, M., and Tara, M. (1986). HTLV-I associated myelopathy, a new clinical entity. Lancet *1*, 1031–1032.

Plummer, M., de Martel, C., Vignat, J., Ferlay, J., Bray, F., and Franceschi, S. (2016). Global burden of cancers attributable to infections in 2012: a synthetic analysis. Lancet Glob. Health *4*, e609–e616.

Poiesz, B.J., Ruscetti, F.W., Gazdar, A.F., Bunn, P.A., Minna, J.D., and Gallo, R.C. (1980). 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. USA *77*, 7415–7419.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26.

Rosewick, N., Durkin, K., Artesi, M., Marçais, A., Hahaut, V., Griebel, P., Arsic, N., Avettand-Fenoel, V., Burny, A., Charlier, C., et al. (2017). Cis-perturbation of cancer drivers by the HTLV-1/BLV proviruses is an early determinant of leukemogenesis. Nat. Commun. *8*, 15264.

Satou, Y., Yasunaga, J., Yoshida, M., and Matsuoka, M. (2006). HTLV-I basic leucine zipper factor gene mRNA supports proliferation of adult T cell leukemia cells. Proc. Natl. Acad. Sci. USA *103*, 720–725.

Satou, Y., Yasunaga, J., Zhao, T., Yoshida, M., Miyazato, P., Takai, K., Shimizu, K., Ohshima, K., Green, P.L., Ohkura, N., et al. (2011). HTLV-1 bZIP factor induces T-cell lymphoma and systemic inflammation *in vivo*. PLoS Pathog. *7*, e1001274.

Satou, Y., Miyazato, P., Ishihara, K., Yaguchi, H., Melamed, A., Miura, M., Fukuda, A., Nosaka, K., Watanabe, T., Rowan, A.G., et al. (2016). The retrovirus HTLV-1 inserts an ectopic CTCF-binding site into the human genome. Proc. Natl. Acad. Sci. USA *113*, 3054–3059.

Satou, Y., Katsuya, H., Fukuda, A., Misawa, N., Ito, J., Uchiyama, Y., Miyazato, P., Islam, S., Fassati, A., Melamed, A., et al. (2017). Dynamics and mechanisms of clonal expansion of HIV-1-infected cells in a humanized mouse model. Sci. Rep. *7*, 6913.

Seiki, M., Hattori, S., Hirayama, Y., and Yoshida, M. (1983). Human adult T-cell leukemia virus: complete nucleotide sequence of the provirus genome integrated in leukemia cell DNA. Proc. Natl. Acad. Sci. USA *80*, 3618–3622.

Shimoyama, M. (1991). Diagnostic criteria and classification of clinical subtypes of adult T-cell leukaemia-lymphoma. A report from the Lymphoma Study Group (1984-87). Br. J. Haematol. *79*, 428–437.

Strain, M.C., Lada, S.M., Luong, T., Rought, S.E., Gianella, S., Terry, V.H., Spina, C.A., Woelk, C.H., and Richman, D.D. (2013). Highly precise measurement of HIV DNA by droplet digital PCR. PLoS ONE *8*, e55943.

Suehiro, Y., Hasegawa, A., Iino, T., Sasada, A., Watanabe, N., Matsuoka, M., Takamori, A., Tanosaki, R., Utsunomiya, A., Choi, I., et al. (2015). Clinical outcomes of a novel therapeutic vaccine with Tax peptide-pulsed dendritic cells for adult T cell leukaemia/lymphoma in a pilot study. Br. J. Haematol. *169*, 356–367.

Sung, W.K., Zheng, H., Li, S., Chen, R., Liu, X., Li, Y., Lee, N.P., Lee, W.H., Ariyaratne, P.N., Tennakoon, C., et al. (2012). Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. Nat. Genet. *44*, 765–769.

Takemoto, S., Matsuoka, M., Yamaguchi, K., and Takatsuki, K. (1994). A novel diagnostic method of adult T-cell leukemia: monoclonal integration of human T-cell lymphotropic virus type I provirus DNA detected by inverse polymerase chain reaction. Blood *84*, 3080–3085.

Tamiya, S., Matsuoka, M., Etoh, K., Watanabe, T., Kamihira, S., Yamaguchi, K., and Takatsuki, K. (1996). Two types of defective human T-lymphotropic virus type I provirus in adult T-cell leukemia. Blood *88*, 3065–3073.

Tezuka, K., Xun, R., Tei, M., Ueno, T., Tanaka, M., Takenouchi, N., and Fujisawa, J. (2014). An animal model of adult T-cell leukemia: humanized mice with HTLV-1-specific immunity. Blood *123*, 346–355.

Uchiyama, T., Yodoi, J., Sagawa, K., Takatsuki, K., and Uchino, H. (1977). Adult T-cell leukemia: clinical and hematologic features of 16 cases. Blood *50*, 481–492.

Van Dooren, S., Gotuzzo, E., Salemi, M., Watts, D., Audenaert, E., Duwe, S., Ellerbrok, H., Grassmann, R., Hagelberg, E., Desmyter, J., and Vandamme, A.M. (1998). Evidence for a post-Columbian introduction of human T-cell lymphotropic virus [type I] [corrected] in Latin America. J. Gen. Virol. *79*, 2695–2708.

Van Dooren, S., Pybus, O.G., Salemi, M., Liu, H.F., Goubau, P., Remondegui, C., Talarmin, A., Gotuzzo, E., Alcantara, L.C., Galvão-Castro, B., and Vandamme, A.M. (2004). The low evolutionary rate of human T-cell lymphotropic virus type-1 confirmed by analysis of vertical transmission chains. Mol. Biol. Evol. *21*, 603–611.

Verdonck, K., González, E., Van Dooren, S., Vandamme, A.M., Vanham, G., and Gotuzzo, E. (2007). Human T-lymphotropic virus 1: recent knowledge about an ancient infection. Lancet Infect. Dis. *7*, 266–281.

Vidal, A.U., Gessain, A., Yoshida, M., Tekaia, F., Garin, B., Guillemain, B., Schulz, T., Farid, R., and De Thé, G. (1994). Phylogenetic classification of human T cell leukaemia/lymphoma virus type I genotypes in five major molecular and geographical subtypes. J. Gen. Virol. *75*, 3655–3666.

Watanabe, T. (1997). HTLV-1-associated diseases. Int. J. Hematol. *66*, 257–278.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
Asymptomatic carriers and ATL patients' PBMCs	Imamura General Hospital	N/A
HAM/TSP patients' PBMCs	St. Marianna University School of Medicine	N/A
DNA extracted from ATL PBMCs	Joint Study on Predisposing Factors of ATL Development (JSPFAD)	N/A
Mouse splenocyte	Kansai Medical University	N/A
Chemicals, Peptides, and Recombinant Proteins		
Ficoll-Paque PLUS	GE Healthcare	Product code#:17-1440-02
DNeasy blood & tissue kit	QIAGEN	Cat#:69504
Critical Commercial Assays		
NEBNext Ultra II DNA library prep kit for Illumina	New England Biolabs	Cat#: E7645S
xGen lockdown reagents	Integrated DNA Technologies	Cat#:1072281
GenNext NGS library quantification kit	TOYOBO	Product code#: NLQ-101
Library quality check by TapeStation	Agilent Technologies	https://www.agilent.com/en/promotions/agilent-2200-tapestation-system
Deposited Data		
Fastq and bam files of all samples	This paper	SRA accession: PRJNA520252
Experimental Models: Cell Lines		
ED cells	Dr Michiyuki Maeda	Maeda et al., 1985
Jurkat cells	ATCC	TIB-152
293T cells	ATCC	CRL-3216
JET cells	Kansai Medical University	N/A
Experimental Models: Organisms/Strains		
Mouse: NOD.Cg-Prkdc^{scid} $\mathit{Il2rg}^{tm1Wjl}$/SzJ	Charles River	https://www.criver.com
Oligonucleotides		
See Table S5	This Paper	NA
DNA probes for enrichment	Miyazato et al., 2016	https://www.nature.com/articles/srep28324
Recombinant DNA		
pX1 plasmid	Laboratory of Dr. David Derse	N/A
pACH plasmid	Laboratory of Dr. Lee Ratner	N/A
Software and Algorithms		
BWA-MEM algorithm	Li and Durbin., 2009	http://bio-bwa.sourceforge.net/
Samtools	Li and Durbin, 2009	http://samtools.sourceforge.net/
Picard	Broad Institute of MIT and Harvard	http://broadinstitute.github.io/picard/
Integrative Genomics Viewer	Robinson et al., 2011	http://software.broadinstitute.org/software/igv/
MEGA7	Kumar et al., 2016	https://www.ncbi.nlm.nih.gov/pubmed/27004904
Prism 7	GraphPad Software	https://www.graphpad.com/

LEAD CONTACT AND MATERIALS AVAILABILITY

- Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Yorifumi Satou (y-satou@kumamoto-u.ac.jp).
- This study did not generate new unique reagents.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Patients Samples

All protocols involving human subjects were reviewed and approved by the Kumamoto University institutional review board (Approval ID: Genome no. 263). The study was carried out in accordance with the guidelines proposed in the Declaration of Helsinki. Informed written consent from human subjects was obtained in this study. The asymptomatic HTLV-1 carriers and patients with adult T-cell leukemia-lymphoma (ATL) were seen at Imamura General Hospital in Kagoshima-city. The patients with HTLV-1-associated myelopathy/tropical spastic paraparesis (HAM/TSP) were seen at St. Marianna University School of Medicine in Kawasaki-city. Twenty-six DNA samples from ATL patients were provided by the Joint Study on Predisposing Factors of ATL Development (JSPFAD). PBMCs were obtained from asymptomatic HTLV-1 carriers (n = 24), HAM/TSP (n = 29), and ATL patients (n = 45). The ATL cases consist of 21 acute, 13 chronic and 11 smoldering types (Shimoyama, 1991). Characteristics of each group are presented in Table S1. PBMCs were separated by density-gradient centrifugation using Ficoll-Paque (GE Healthcare) and cryopreserved in Cell Banker (Juji Field Inc.) at −80°C until use. Genomic DNA was extracted from PBMCs using the DNeasy kit (QIAGEN) according to the manufacture's protocol.

Cell lines

We used an ATL-derived cell line, ED cells and three HTLV-1-uninfected T-cell lines, Jurkat, and JET, which is a subline of Jurkat cells expressing tdTomato under the control of 5 tandem repeats of Tax responsive element. 293T cells are derived from human embryonic kidney 293 cells. These cells were cultured in RPMI supplemented with 10% FBS, 100 U/ml penicillin and 100 μg/ml streptomycin. The extraction of DNA was performed as described above for PBMCs from HTLV-1-infected individuals. To estimate the efficiency of our protocol, we prepared test DNAs by mixing ED genomic DNA (gDNA) in the amount of 1% and 10% with Jurkat gDNA.

Humanized mice

The experiment with humanized mice was approved by the Animal Care Committee of Kansai Medical University. Humanize mice, including both male and female, were generated by intra–bone marrow injection of human CD133^{+} stem cells into NOD.Cg-*Prkdc*^{scid} *Il2rg*^{tm1Wjl}/SzJ mice (Charles River). The HTLV-1-infected Jurkat cells were irradiated with 10 Gy from a ^{137}Cs source irradiator and were inoculated intraperitoneally into 24- to 28-week-old huNOG mice (Tezuka et al., 2014). The splenocytes were isolated from the sacrificed mice at around 3, 5, 8, and 12 weeks after infection. The extraction of gDNA was performed by the same method as for PBMCs from HTLV-1-infected individuals as described above.

METHOD DETAILS

Generation of HTLV-1-infected cells *in vitro*

293T cells transfected with pX1 plasmid, an HTLV-1 molecular clone (Mitchell et al., 2007), by Polyethylenimine (PEI), and then were irradiated with 30 Gy. The irradiated 293T cells were co-cultured with JET cells for 3 days (Furuta et al., 2017). Then, tdTomato positive cells, which is induced by Tax in the cells, were sorted by FACSAria, and cultured in RPMI supplemented with 10% FBS, 100 U/ml penicillin and 100 μg/ml streptomycin for 2, 8, or 16 weeks. The extraction of DNA was performed as described above.

Proviral load measurement by droplet digital PCR

Droplet digital PCR (ddPCR) was performed by using primers and a probe targeting a conserved region in HTLV-1 pX region and the *ALB* gene, according to previous reports with minor modifications (Strain et al., 2013). ddPCR droplets were generated by the QX200 droplet generator (Bio-Rad). Generated droplets were then transferred to a 96-well PCR plate and sealed with a pre-heated PX1 PCR plate sealer (Bio-Rad) for 5 s at 180°C. PCR cycles were performed in a C1000 Touch thermal cycler (Bio-Rad) with the following settings: 95°C for 10 minutes followed by 39 cycles of 94°C for 30 s, 58°C for 60 s, and final 98°C for 10 minutes and 4°C for hold. The plate was then placed in the QX200 droplet reader (Bio-Rad) for quantification of the number of positive and negative droplets based on their fluorescence. Threshold values for ddPCR were determined based on the highest level of droplet fluorescence in the no-template-control sample (NTC) to provide an objective cut-off with maximum sensitivity. Data were analyzed using QuantaSoft software (Bio-Rad). Then proviral load was calculated as follows,

$$\text{proviral load}(\%) = (\text{copy number of HTLV1 pX DNA}) \div \{(\text{copy number of ALB}) \div 2\} \times 100$$

Library synthesis, and proviral DNA-capture-seq

The HTLV-1 DNA-capture-seq was performed as previously described with minor modifications (Miyazato et al., 2016). gDNA was sheared by sonication with a Picoruptor (Diagenode) to obtain fragments of an average size of 300 bp. Libraries for NGS were prepared using the NEBNext Ultra II DNA library prep kit for Illumina (NEB) following the manufacturer's instructions. We next performed a probe-based enrichment step as previously reported (Miyazato et al., 2016). Briefly, multiplexed libraries were mixed with the 148 biotinylated DNA probes targeting HTLV-1 proviral sequences (GenBank: AB513134). Hybridization was performed by incubating the mixture at 65°C for four hours. After the addition of streptavidin-coated beads, and several wash steps (xGen lockdown reagents, IDT), the captured DNA fragments were amplified by PCR with P5 and P7 primers for Illumina sequencing. The enriched DNA libraries were quantified by TapeStation instrument (Agilent Technologies), and quantitative PCR (GenNext NGS library quantification kit, Toyobo). Sequencing was carried out on Illumina MiSeq or NextSeq instrument with 2 × 75 bp reads.

High-throughput sequencing data analysis

Three fastq files, Read1, Read2 and Index Read, were obtained from Illumina MiSeq or NextSeq. We first performed a data-cleaning step by using an in-house Perl script (kindly provided by Dr Michi Miura, Imperial College London), which extracts reads with a high Index Read sequencing quality (Phred score > 20 in each position of the 8-bp index read). We next removed adaptor sequences from Read1 and Read2 followed by a cleaning step to remove reads with too short or with too low Phred score as previously described (Satou et al., 2017). The cleaned sequencing reads were aligned to human reference genome (hg19) with HTLV-1 (GenBank: AB513134) as a separate chromosome or integrated provirus using the BWA-MEM algorithm (Li and Durbin, 2009). We then used Samtools (Li and Durbin, 2009) and Picard (http://broadinstitute.github.io/picard/) for further data processing and cleanup such as removal of reads with multiple alignments and duplicated reads. Visualization of the aligned reads was performed using Integrative Genomics Viewer (IGV) (Robinson et al., 2011).

Ligation-mediated (LM)-PCR

HTLV-1 IS analysis was performed using LM-PCR and high-throughput sequencing. About 1 μg of genomic DNA was sheared by sonication with a Picoruptor (Diagenode, S.A., Belgium) instrument to a size of 300-400bp in length. DNA end was repaired and addition of adenosine at the 3′ end of the DNA was performed with NEBNext Ultra II End Repair/dA-Tailing Module (New England Biolabs). Linker was then ligated to the ends of DNA using NEBNext Ultra II Ligation Module (New England Biolabs). Ligated products were amplified by a first PCR targeting the 3′-LTR in one end and the linker in the other end. First PCR amplicons were cleaned by QIAquick PCR purification kit (QIAGEN) and performed second PCR. The following thermal cycler conditions were used for both PCRs: 96°C for 30sec (1 cycle); 94°C for 5sec, 72°C for 1min (7 cycles); 94°C for 5 s, 68°C 1 min (13 cycles); 68°C 9 min (1 cycle) and hold at 4°C. Second PCR amplicons were purified using the QIAquick PCR Purification Kit (QIAGEN) and followed by Ampure XP bead purification. Purified PCR amplicons were quantified using Agilent 2200 TapeStation and quantitative PCR (GenNext NGS library quantification kit, Toyobo). LM-PCR libraries were sequenced on the Illumina MiSeq as paired-end read, and the resulting fastq files were analyzed. Oligonucleotides used to perform LM-PCR and sequencing were listed in the key resources table.

Integration site, proviral structure, and clonal abundance analysis with the DNA-seq data

To analyze IS and proviral structure, we aligned the cleaned fastq files to the reference genome containing all human chromosomes (chr 1–22, X, Y and M) and HTLV-1 as 2 separate chromosomes – the viral LTR sequence (HTLV_LTR) and the whole viral sequence excluding the LTRs (HTLV_noLTR). We extracted virus-host reads by an in-house Python script, which generates list of all virus-host reads in the sample. Random ligations between virus and host DNAs were generated during DNA library preparation, resulting in one-sided virus-host reads at either 5′ or 3′ sides (Figure 3C; Table S4). Thus, in this study we defined viral integrations sites by the presence of virus-host reads at both 5′ and 3′ end of provirus (Figure 3A). Then locations of IS in the human genome were determined by the host-virus junction that was present within sequencing reads. When the host-virus junction was not present within sequencing reads, we defined the integration sites as the position at the center of 5′-side host read and 3′-side host read of virus-host chimeric reads (Figure 3A).

The degree of clonal abundance for each infected clone was calculated by the number of virus-host reads. After PCR replicate removal, the number of final virus-host reads in a certain genomic region reflects the initial cell number of the clone. Thus, it would be possible to estimate the initial copy number of each clone by counting the number of virus-host reads after removal of PCR replicates, as is the case with linker-mediated PCR (Berry et al., 2012; Gillet et al., 2011).To compare the clonality between the samples, the number of the final reads were normalized using the number of the total HTLV-1-aligned reads and the proviral load as follows,

$$\text{clonal\ \ abundance of each clone} = \#\text{of final virus host reads} \times (10,000 \div \#\text{of total HTLV1 mapped reads}) \\ \times \{\text{proviral load } (\%) \div 100\}$$

Histone modifications of primary helper memory T cells from peripheral blood were obtained from ChIP-Seq datasets provided by ENCODE project (Consortium, E.P., 2012). The relationship between HIV-1 IS and histone modification was analyzed as reported previously (Satou et al., 2017). We made bed files with ISs of 5′-defective proviruses to perform gene enrichment analysis using GREAT, an online software application for gene annotations (McLean et al., 2010).

Phylogenetic Analysis

HTLV-1 sequence alignments were performed using MUSCLE (Edgar, 2004). Genetic distances were calculated between and within isolates, and neighbor-joining trees were generated using a maximum composite likelihood algorithm and default parameters using MEGA7 software (Molecular Evolutionary Genetics Analysis Program) (Kumar et al., 2016).

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analysis was performed using Prism 7 software (GraphPad Software, Inc., CA). Fisher's exact test was used in a two-tailed test to compare the differences of proportions between HTLV-1 strains and diseases, and provirus structure and diseases. A nonparametric test (Mann-Whitney *U* test) was used in a two-tailed test to examine the association between clonal abundance and provirus structures.

DATA AND CODE AVAILABILITY

The bam files analyzed from all patients' samples have been deposited at NCBI BioProject (accession: PRJNA520252).