Selective clonal persistence of human retroviruses in vivo: Radial chromatin organization, integration site, and host transcription

Anat Melamed1*, Tomas W. Fitzgerald2, Yuchuan Wang3, Jian Ma3, Ewan Birney2, Charles R. M. Bangham1*

The human retroviruses HTLV-1 (human T cell leukemia virus type 1) and HIV-1 persist in vivo as a reservoir of latently infected T cell clones. It is poorly understood what determines which clones survive in the reservoir. We compared >160,000 HTLV-1 integration sites (>40,000 HIV-1 sites) from T cells isolated ex vivo from naturally infected individuals with >230,000 HTLV-1 integration sites (>65,000 HIV-1 sites) from in vitro infection to identify genomic features that determine selective clonal survival. Three statistically independent factors together explained >40% of the observed variance in HTLV-1 clonal survival in vivo: the radial intranuclear position of the provirus, its genomic distance from the centromere, and the intensity of local host genome transcription. The radial intranuclear position of the provirus and its distance from the centromere also explained ~7% of clonal persistence of HIV-1 in vivo. Selection for the intranuclear and intrachromosomal location of the provirus and host transcription intensity favors clonal persistence of human retroviruses in vivo.

INTRODUCTION

Human T cell leukemia virus type 1 (HTLV-1) persists in the host chiefly by clonal proliferation (1, 2). A typical HTLV-1–infected host has 10^6 to 10^7 HTLV-1–infected T cell clones (3); each clone can be distinguished by the unique integration site of the single-copy provirus in the host genome (4). Every clone has its own characteristics of proviral expression, host gene expression, chromatin structure, and equilibrium abundance; each of these attributes is influenced by the genomic integration site (5, 6).

In primary infection, the initial virus spread is rapid (7). The proviral load [PVL; percentage of HTLV-1–infected peripheral blood mononuclear cells (PBMCs)] reaches an equilibrium or set point in each host. The PVL can vary between hosts by more than 1000-fold (8) and is proportional to the number of different HTLV-1+ T cell clones (9). The PVL is partly determined by the host immune response; the force of selection exerted by the HTLV-1–specific cytotoxic T lymphocytes depends on the level of expression of HTLV-1 antigens (10). Both HTLV-1 and HIV-1 persist in a reservoir in vivo that depends partly on continued clonal proliferation (1, 11, 12).

We previously reported that, whereas initial integration of HTLV-1 shows no preference for any given chromosome, the HTLV-1+ clones that persist in vivo are found more often than by chance in the acrocentric chromosomes (13, 14, 15, 21, and 22) (13). The centromere-proximal regions of these chromosomes lie in the transcriptionally repressive environment around the nucleolus. This observation (13) suggested that repression of proviral expression in the nucleolar periphery minimizes the exposure of the infected cell to the strong anti-HTLV-1 immune response and so favors the survival of that clone.

Chromosomes are not randomly distributed in the nucleus: Each chromosome occupies a characteristic position known as a chromosome territory (CT) (14). The CTs are not static but rather represent an average in the cell population. Both the CTs and individual chromosomes are radially organized in the nucleus (15): For example, chromosomes 18 and X often lie near the nuclear periphery, whereas chromosomes 17 and 19 are usually found in the center of the nucleus (16–18).

The spatial distribution of transcriptional activity of the genome is also nonrandom. Intranuclear bodies, known as nuclear speckles, are associated with transcription and pre-mRNA processing (19, 20). Two types of genomic domain are associated with particularly low transcriptional activity: lamina-associated domains (LADs) (21), near the nuclear periphery, and nucleolus-associated domains (NADs) (22–24). These domains are characterized by a relatively low gene density, a high density of repressive histone marks (H3K9me2/ H3K9me3 in particular), and low guanine–cytosine (GC) content (25). There is a strong overlap between LADs and NADs, and certain domains stochastically reassociate with either the nuclear lamina or nucleolus in the daughter cells after mitosis (23, 26).

The selective persistence of HTLV-1 proviruses in certain chromosomes and the persistence in vivo of very different numbers of HTLV-1+ T cell clones in different hosts imply that HTLV-1 infection results in the selective survival of certain clones, which then persist for the remainder of the host’s life. The aim of this study was to test two hypotheses that arise from these observations: first, that specific genomic attributes of the proviral integration site in an HTLV-1–infected clone determine its survival in the host during chronic infection, and second, that the intranuclear position of the provirus determines clonal survival in vivo. Last, we applied the same approach to analyze the data on HIV-1 integration sites.

RESULTS

Analysis of integration sites identifies genome-wide correlates of clone survival

To identify the features of the HTLV-1 proviral integration site associated with clonal persistence in vivo, we compared specific...
Preferential survival of HTLV-1 and HIV-1 is nonrandomly distributed between and within chromosomes

While initial integration (in vitro) of HTLV-1 occurs in proportion to chromosome size (fig. S1) (13), analysis of the much larger datasets in the present study confirmed the previous observation (13) of the preferential survival of HTLV-1 in vivo in acrocentric chromosomes as a group. On closer inspection, this preference is seen to be due to a strong bias for survival in chromosomes 13, 14, and 15; no survival bias was observed in chromosome 21, and survival was counterselected in chromosome 22, the smallest acrocentric chromosome. The chromosome most preferred for survival in vivo is chromosome 18, and the chromosome most disfavored for survival in vivo is chromosome 19 (Fig. 1A and fig. S1). By contrast, in HIV-1–infected cells, there was a stronger bias for or against initial integration in particular chromosomes than in HTLV-1–infected cells: The gene-rich chromosome 19 was the most favored for initial HIV-1 integration, whereas the similarly sized, gene-poor chromosome 18 was disfavored (fig. S2, A and B). However, the rank order of chromosomes preferred for integration in vitro was very similar between HTLV-1 and HIV-1 (Kendall’s τ = 0.53, P < 0.001; fig. S2C). The rank order of chromosome preference for in vivo survival was also correlated between the two viruses, albeit less strongly (Kendall’s τ = 0.32, P < 0.05; fig. S2D).

Within each chromosome, certain regions are either favored or disfavored for HTLV-1 survival in vivo. On most chromosomes, areas close to the centromere are more strongly favored, whereas areas more distant from the centromere and closer to the telomere are disfavored in vivo, in contrast with the more uniform distribution of initial integration in vitro (Fig. 1B).

HTLV-1 initial integration favors accessible, active chromatin

Previous work has shown that initial integration of HTLV-1 was strongly preferred in close proximity to specific transcription factor binding sites (TFBS) (5). To extend this observation to additional TFBS, we used all transcription factor chromatin immunoprecipitation sequencing (ChIP-seq) datasets published by the Encyclopedia of DNA Elements (ENCODE) project (28, 29) using the B cell line GM12878, comprising 156 ChIP-seq datasets, from 135 transcription factors (an equivalent dataset is not available on ENCODE for T cells; table S2). Each integration site (or random site as control) was annotated with respect to each TFBS dataset, and the minimum distance to any TFBS was calculated. The results show that integration sites were significantly enriched within 1 and 10 kb of any TFBS both in vivo and, more strongly, in vitro, compared with random sites. This observation suggests that initial integration is more frequent in accessible chromatin, available for transcription factor binding (fig. S3A). To corroborate this conclusion, we identified the deoxyribonuclease (DNase) I hypersensitive site (DHS) nearest to each integration site. We find that HTLV-1 sites within 10 kb of a DHS are more frequent than random expectation, both in vivo and especially in vitro (Fig. 2A and fig. S3B).

Last, we used histone mark datasets from the ENCODE project to annotate the histone mark density in fixed windows, either 1-kb windows up to 1 Mb on either side of the integration site (Fig. 2B

![Image](https://www.science.org)
sites was calculated in discrete windows upstream and downstream across all integration sites. 

frequent near histone marks associated with transcriptional repression (e.g., H3K9me3). Integration sites in vivo were less frequent than in vitro near all histone marks evaluated, except for H3K9me3, suggesting counterselection in vivo of proviruses that lie near those marks. Whereas the histone mark distribution around initial (in vitro) integration sites was symmetrical, the corresponding distribution around in vivo sites was, in some cases, asymmetrical. For example, H3K27me3 appeared to be counterselected in vivo more strongly downstream of the integration site than upstream (Fig. 2C and fig. S3D).

**HTLV-1 survival and position along the chromosome**

To identify the within-chromosome features that favor or disfavor survival of integrated HTLV-1 proviruses in vivo, we divided the human genome (GRCh38; see Methods) into discrete 1-Mb windows. We define the HTLV-1 survival index in each window as \( \text{Ln}(F_{\text{in vivo}}/F_{\text{in vitro}}) \), where \( F \) denotes the respective frequency of integration sites in that window.

We define the distance from each window to the centromere as the absolute distance between the midpoint of each window and the midpoint of the centromere. The results show a strong negative correlation between the HTLV-1 survival index and the distance from the centromere (Fig. 3) on both the \( p \) and the \( q \) arms of the chromosomes (Pearson’s \( R = -0.26 \) and \(-0.36 \), respectively): Survival is favored when the provirus is integrated closer to the centromere and progressively disfavored toward the telomeres (fig. S4). Survival of the integrated provirus correlated significantly less strongly with the distance from the telomere than with the distance to the centromere (fig. S5).

**HTLV-1 proviruses selectively survive in chromatin near the nuclear lamina and distant from nuclear speckles**

The observed preferential survival of proviruses integrated in particular chromosomes (13, 30) raised the question whether the physical position of the provirus in the nucleus influences HTLV-1 clonal survival in vivo. To answer this question, we estimated the distance of the provirus from specific intranuclear sites by using tyramide signal amplification-sequencing (TSA-seq) data published by Chen et al. (19) to estimate the distance of a given genomic location from the nuclear lamina or from nuclear speckles. We aligned and processed the integration site data according to the protocol described (19) and determined the mean TSA-seq signal in each 1-Mb window, in which the integration site frequency was quantified. The results (Fig. 4A) showed a bias toward initial integration (in vitro) in chromatin that lies near nuclear speckles. By contrast, the HTLV-1 in vivo survival index showed a strong positive correlation with proximity to the lamin proteins and a strong negative correlation with proximity to the SON protein (Fig. 4B and figs. S6 to S8). A similar analysis of HIV integration sites showed a marked preference toward the integration near SON (fig. S9), consistent with a recent report of frequent HIV-1 integration in nuclear speckle-associated domains (31). However, there was a trend toward increased survival away from nuclear speckles and near the lamin proteins, similar to that observed in HTLV-1 (Fig. 4C and fig. S10). Genome-wide analysis confirmed a significant positive correlation between proximity to lamin proteins and the survival of HTLV-1 and HIV-1 proviruses in vivo and a significant negative correlation between proximity to nuclear speckles and survival (Fig. 5A and fig. S11, A and B).

LADs are usually identified using DNA adenine methyltransferase identification (DamID). To corroborate the observation that proviral
Fig. 3. HTLV-1 survival versus distance from the centromere. The HTLV-1 clone survival index $\text{CSI}_{\text{HTLV-1}}$ [defined as $\text{Ln}([\text{int}]/\text{vivo}/\text{in vitro}$), the natural logarithm of the ratio between HTLV-1 in vivo and HTLV-1 in vitro site frequencies] is significantly negatively correlated with the absolute genomic distance from the centromere in both the short and long arms of the chromosomes (Pearson’s correlation test).

HTLV-1 survival is associated with proximity to the nuclear lamina by an independent approach, we used a DamID dataset produced in T cells (32). The results show that both HTLV-1 and HIV-1 integration sites in vivo are enriched in LADs compared to integration sites in vitro (Fig. 5B). Further analysis showed that integration sites identified in patients on antiretroviral therapy (ART) were significantly enriched in LADs compared with those from patients who were not on ART (fig. S11C).

Last, we used the recently reported spatial position inference of the nuclear genome (SPIN) method, which combines data from TSA-seq, DamID, and Hi-C to build a model that defines a set of spatial localization states of chromatin relative to nuclear bodies, reflecting a gradient of radial position from the nuclear lamina to the nuclear speckles (33). This analysis shows a monotonic increase in the HTLV-1 survival index toward the lamina (Fig. 5C).

HTLV-1+ clone survival in vivo independently correlates with the expression status of the genomic region

The highest gene density is often located in “T” bands of the human genome, many of which are telomeric (34, 35). We therefore investigated whether the observed decrease in survival associated with greater distance from the centromere could be attributed to an increase in gene density. In each 1-Mb window along the genome, we quantified gene density using the Ensembl database and compared this density against the survival index. The results show that the gene density correlates with the distance to nuclear speckles (fig. S12, A and B) and is strongly negatively correlated with the survival index (fig. S12C).

Because HTLV-1 is primarily found in vivo in CD4+ T cells (36), we wanted to test whether the expression of genes, specific to those cells, plays a role in the selective survival of integrated proviruses. Using expression data on primary T cells from the BLUEPRINT Epigenome project (37), we defined a mean expression level (regardless of the number or position of genes) in 1-Mb windows across the genome. In the observed bimodal distribution of expression intensity, we used the local minimum to define genomic windows that are low expressing or high expressing (fig. S13). We find that the HTLV-1 survival index is
significantly lower in high-expressing genomic sites \((P < 10^{-16}, \text{Wilcoxon rank sum test})\). Similarly, at the level of 1-Mb windows, there was a strong negative correlation between expression intensity and the survival index for both HTLV-1 and HIV-1 (Fig. 6).

Since several genomic features considered here are known to be correlated, for example, mean expression intensity and distance from the centromere, we carried out multivariate linear regression to identify the independent correlates of survival of HTLV-1 \(^+\) clones in vivo. Three factors—distance to nuclear speckles (TSA-seq signal), the expression intensity, and the distance to the centromere—remain significant independent predictors, together explaining \(\sim 40\%\) of the observed variation in the HTLV-1 survival index (Table 2, table S3, and fig. S14). A similar analysis of the HIV-1 data identified two of these factors—the distance to nuclear speckles and distance to the centromere—as independent correlates, together explaining \(\sim 7\%\) of the observed variation in HIV-1 \(^+\) clone survival in vivo (Table 2 and table S4).

**DISCUSSION**

Similar to other persistent viruses, HTLV-1 establishes an equilibrium between viral replication and the host immune response. HTLV-1 does this by two chief mechanisms: first, by replicating mainly by clonal proliferation of infected cells rather than by de novo infection, thus minimizing the need for viral antigen expression and consequent immune-mediated killing, and second, by expressing the proviral plus strand (which encodes the most immunogenic viral antigens) in rare, self-limiting bursts (38, 39). The resulting reservoir of long-lived HTLV-1 clones is very large: The PVL frequently exceeds 10\% of PBMCs in nonmalignant HTLV-1 infection.

The optimal strategy of survival for persistence of HTLV-1 in vivo is therefore to minimize proviral expression during most of the lifetime of the infected cell, while retaining the ability to reexpress the provirus in intense bursts, either to infect a new host or to create a new clone in the same host (1). The characteristics of proviral expression differ from clone to clone and appear to be determined largely by the proviral integration site (5, 6). We therefore hypothesized that local features of the chromatin flanking the provirus, such as epigenetic modifications associated with transcriptional activity, would correlate with the selective clonal survival of HTLV-1 \(^+\) cells in vivo.

The results presented here show that certain epigenetic marks are associated with in vivo survival of an HTLV-1 \(^+\) T cell clone; however, these effects are relatively weak. By contrast, we found a remarkably strong correlation between selective in vivo clone survival and three
High expression regions with a tendency to locate near the nuclear lamina is more likely to persist in vivo than one whose provirus occupies a central position in the nucleus.

We conclude that integration of an HTLV-1 provirus into a genomic region that typically occupies a transcriptionally repressive compartment in the nucleus—near the nuclear lamina or the nucleolar periphery—favors the survival of that clone in vivo. Initial proviral integration favors transcriptionally active, accessible regions of the genome (Fig. 4A) (5, 9), but the results reported here show that proviruses in regions of high transcriptional activity are counterselected during the subsequent chronic infection. However, the radial intranuclear position of the provirus and its intrachromosomal location influence selective clonal survival of the virus independently of the local host transcription intensity. The importance of the spatial intranuclear position of the provirus in vivo clone survival of human retroviruses is summarized in the model in Fig. 7.

We previously showed that the HTLV-1 provirus binds the chromatin architectural protein CTCF (40) and thereby deregulates the higher-order structure and transcription of the flanking host genome (6). CTCF contributes to the localization of chromatin to the nucleolar periphery (41). It is therefore possible that CTCF binding provides an advantage to the virus by promoting association of the provirus with this transcriptionally repressive compartment.

HIV-1 differs strongly from HTLV-1 in its strategy of persistence in vivo. HTLV-1 expression is noncytolytic, allowing clones to persist by intermittent proviral expression. By contrast, HIV-1 expression is cytolytic, and the virus persists in the host mainly by sustained de novo infection, that is, creation of new (albeit mostly short-lived) clones. However, the reservoir of HIV-1–infected cells that can persist indefinitely during highly active ART is maintained partly by clonal proliferation (11, 12), perhaps driven by normal homeostatic mechanisms (42). We applied the methods described above to analyze data on the HIV-1 proviral integration site, again from both in vitro infection and from cells isolated from infected individuals, both pre-ART and on ART (27).

The results show that HIV-1 clone survival in vivo, similar to that of HTLV-1, is correlated with the nuclear position of the provirus (distance from nuclear speckles) and the distance from the centromere. This observation contrasts with the fact that the nuclear speckle-associated domains, which are enriched in transcriptionally active genes, are strongly favored for the initial integration of HIV-1. However, in contrast with HTLV-1, only approximately 7% of the variation in proviral survival of HIV-1 can be explained by these factors: We postulate that this difference is due to the difference between the two retroviruses in the relative importance of infectious spread and mitotic spread (1) in the persistence of the virus in the host. The power of this analysis of HIV-1 data is limited by two factors: first, the smaller sample size (a total of 110,338 integration sites of HIV-1, cf. 397,910 integration sites of HTLV-1), and second, the uncertainty in the proportion of HIV-1 integration sites identified in vivo that represent the true persistent reservoir, rather than short-lived clones. It is possible that the importance of the intranuclear spatial location of the HIV-1 provirus in persistence in vivo exceeds the estimate of 7% obtained here.

Jiang et al. (43) examined the integration sites of HIV-1 in a group of elite controllers and individuals on ART (number of genomes analyzed = 1385 and 2388, respectively). They observed an overrepresentation of integration sites of genome-intact HIV-1 proviruses in centromeric satellite DNA, especially in elite controllers. These authors concluded that persistence of intact HIV-1 proviruses favors

---

**Table 2. Linear model, significant predictors of CSI.**

| Predictor                                      | HTLV-1 | HIV-1  |
|-----------------------------------------------|--------|--------|
| (Intercept)                                   | −0.12  | 0.25   |
| Proximity P to nuclear speckle*               | −0.38  | −0.33  |
| Distance D from centromere (Gb)               | −4.36  | −1.78  |
| $P_i \times D_i$                              | 2.87   | NS     |
| Local host expression intensity†              | −0.04  | NS     |
| $N$(1-Mb windows)                             | 2677   | 2135   |
| $R^2$ adjusted                                | 0.41   | 0.07   |

*TSA-seq signal for SON. †Ln(expression signal in T cells).
It is less clear why the absolute genomic distance between the provirus and the centromere is strongly correlated with survival, independently of the distance from nuclear speckles or the nuclear lamina. Average transcriptional intensity, gene density, GC content, and early DNA replication all tend to increase toward the telomere. However, neither gene density nor GC content remains as a significant independent correlate of HTLV-1 survival in the multivariate regression analysis, and even after taking the transcriptional intensity into account, the distance from the centromere remains as a strong correlate of survival. The functional importance of DNA replication timing is not well understood (46). Our results suggest that HTLV-1 exploits some other, unidentified feature, independent of intranuclear position and transcriptional activity, which varies with the absolute genomic distance from the centromere.

**METHODS**

**Integration site datasets, cells, and patients**

The HTLV-1 integration site datasets used here (either in vivo or in vitro) are detailed in table S1. Raw FASTQ data were used and processed in parallel to ensure consistency and comparability of data.

FASTQ files were filtered to exclude potential spurious mapping events by selecting sequences that contain the final five bases of the HTLV-1 LTR (47) using Cutadapt (48). Filtered sequencing reads were trimmed using trim galore (www.bioinformatics.babraham.ac.uk/projects/trim galore) to remove low-quality (q ≥ 20) and adapter bases (using minimum overlap of three bases) and subsequently aligned against a combined reference of hg38 human genome and HTLV-1 upstream sequence using the Burrow-Wheeler Aligner (BWA) (49). Aligned reads were filtered using SAMtools (50) to include only uniquely mapped proper read pairs (mapQ ≥ 10). Read pairs were further processed using a bespoke R script to correct the mapped position based on the CIGAR string and grouped based on unique pairs between integration sites and shear sites. Last, integration site abundance was estimated using the R package sonicLength (51) and cleaned to correct for mapping and barcode errors using sequence similarity.

In vitro integration sites and donor cell line (MT-2) integration sites were sequenced in parallel. Any integration sites found in the donor cell line or in >1 infection assay were excluded from analysis.

Random sites were selected from the hg38 genome reference using the R package intSiteRetriever (52), and a mock FASTQ file was generated from these positions and hg38 sequence to simulate integration site raw data. Subsequently, this mock FASTQ file was processed through the same pipeline described above to ensure compatibility with integration site data. Custom scripts used to extract the data are available at https://doi.org/10.5281/zenodo.6353523 and https://github.com/ImperialCollegeLondon/intsite-2022.

The three main types of integration site were combined (HTLV-1 in vivo sites, HTLV-1 in vitro sites, and random sites), and repeatedly observed sites were removed from each dataset to ensure nonredundancy. See Table 1 for summary of integration site counts and data file S1 for integration site datasets.

**Integration site frequency and clone survival index**

For the analysis of 1-Mb windows across the human genome, discrete windows along each chromosome were defined from position 1 to the chromosome terminus. One-megabase windows that overlap the end of the chromosome were excluded. To improve mapping
CSI = \ln\left(\frac{F_{\text{in vivo}}}{F_{\text{in vitro}}}\right)

where CSI is undefined (either \(F_{\text{in vivo}}\) or \(F_{\text{in vitro}} = 0\)), and the corresponding 1-Mb window is excluded from statistical analysis.

**HIV integration sites**

We used data on HIV integration sites published by Coffin et al. (27), detailed in table S1. In vitro [phytohemagglutinin (PHA)—untreated only] and in vivo (both pre-ART and post-ART) integration sites from PBMCs were compiled and remapped to hg38 using the liftOver tool included in the R package rtracklayer (55). In keeping with the processing of the HTLV-1 data, if any two HIV-1 integration sites were mapped within 5 bp of each other (~1.2% of integration sites), one of the pair was removed to create a unique list of integration sites. Of the 13,142 HIV-1 integration sites from patients pre-ART and the 32,569 sites from patients on ART, a unique nonredundant list of 44,367 HIV-1 in vivo integration sites was compiled.

**Chromatin modification and accessibility annotation**

To map the presence of integration sites with respect to histone mark density, TFBSs, and DHSs, we used data from experiments carried out and analyzed by the ENCODE project (table S2) (28, 29).

**DNase hypersensitive sites**

DNase-seq data were retrieved from the ENCODE project using the following criteria: organ, blood; cell, leukocyte; biosample, GM12878 or CD4-positive, alpha-beta memory T cell; genome assembly, GRCh38; and file type, “bigwig.” For GM12878, two replicate experiments are reported; a site is recorded as within N bases of a DHS if this condition is satisfied in both experimental replicates. Genomic distances are cumulative; e.g., the integration sites within 10 kb of a DHS also include the integration sites within 1 kb. Annotation of the nearest DHS to each integration site was done using the hiAnnotator R package (https://bioconductor.org/packages/release/bioc/html/hiAnnotator.html).

**Transcription factor binding sites**

Transcription factor ChIP-seq datasets were retrieved from the ENCODE project using the following criteria: organ, blood; cell, leukocyte; biosample, GM12878; genome assembly, GRCh38; file type, “bed narrowpeak”; output type, “optimal idr threshold peaks”; and audit category excluding “extremely low read depth” and “extremely low read length.” At the time of retrieval (August 2019), 156 datasets were available from 135 targets. Where more than one dataset was available for the same target, the larger dataset was used. Annotation of the nearest TFBS to each integration site from each type was made using the hiAnnotator R package.

**Histone mark data**

Histone modification ChIP-seq datasets were retrieved from the ENCODE project using the following criteria: organ, blood; cell, leukocyte; biosample, GM12878 or CD4-positive, alpha-beta memory T cell; genome assembly, GRCh38; file type, “bigwig”; and output type, “fold change over control.” Only those based on two replicates are used. Where more than one dataset was available for the same target, the larger dataset was used. Histone modification signal was averaged over fixed windows in the regions flanking each integration (or random) site using the UCSC bigWigAverageOverBed tool (56) and averaged across all integration sites.

**Nuclear position annotation**

**TSA-seq data analysis**

For consistency of reference (hg38), raw FASTQ data reported by Chen et al. (19) (table S2) were realigned and processed according to the authors’ protocol (https://github.com/ma-compbio/TSA-Seq-toolkit). A Y-excluded reference genome (hg38F) was used (K562 is a female cell line); Bowtie2 (57) was used to align raw data + controls, followed by normalization using the authors’ script. Wig output was converted to BigWig format using UCSC bigToBigwig (56), which was then quantified in fixed 1-Mb windows across the human genome using the UCSC bigWigAverageOverBed tool (56).

**DAM-ID data preparation**

LAD data (table S2) were mapped by Robson et al. (32) using the hg19 reference genome. LAD genomic positions were converted to hg38 using the liftOver tool included in the R package rtracklayer (55). Integration sites in LADs were annotated using the hiAnnotator R package. Here, we show results from annotation against LADS in activated T cells. Using the data from resting T cells does not qualitatively alter these results.

**Chromosome length**

Chromosome positions, chromosome arms, gaps, and centromere data were retrieved from the UCSC table browser (58).

**SPIN data analysis**

We compared the HTLV-1 survival ratio on different SPIN states (33) identified in the K562 cell line. The SPIN states were calculated at a 25-kb resolution and analyzed in the human hg38 genome assembly. First, we assigned SPIN states to each 1-Mb genomic bin. We used bedtools intersect (54) to calculate the overlap between SPIN states and predetermined the 1-Mb genomic bins. We only kept the 1-Mb genomic bins where the majority (≥75%) of regions were covered by a single SPIN state; genomic bins with less than 75% of regions registered with one single SPIN state were discarded. We then calculated the natural logarithm of the ratio of in vivo/in vitro proportion of integration sites (“survival ratio”) for each 1-Mb bin. Genomic bins with zero integration or missing data were also discarded. Last, we plotted the distribution of survival ratio on different SPIN states in the boxplot (Fig. 5C).

**Gene density and T cell expression data**

A gene position list for gene density quantification was retrieved from the Ensembl BioMart database (version 80) using the biomaRt R package (59). The number of genes overlapping each 1-Mb window (overlaps of any length) was counted using the GenomicRanges R package (60).

For analysis of the local genomic expression signal, ribodepleted RNA-seq expression signal datasets were retrieved in bigwig format from the BLUEPRINT project site (http://dcc.blueprint-epigenome.eu/#/files) based on the following criteria: cell type, central memory CD4-positive, alpha-beta T cell; and tissue, venous blood. Using
discrete 1-Mb windows across the human genome, the mean expression signal from uniquely mapped reads per window was calculated using the UCSC bigWigAverageOverBed tool (56) and the mean of two available samples calculated per window.

This study makes use of data generated by the BLUEPRINT Consortium. A full list of the investigators who contributed to the generation of the data is available at www.blueprint-epigenome.eu. Funding for the project was provided by the European Union’s Seventh Framework Programme (FP7/2007-2013) under grant agreement no 282510 - BLUEPRINT.

SUPPLEMENTARY MATERIALS
Supplementary material for this article is available at https://science.org/doi/10.1126/sciadv.abm6210

View/request a protocol for this paper from Bio-protocol.

REFERENCES AND NOTES
1. D. J. Laydon, V. Sunkara, L. Boelen, C. R. M. Bangham, B. Asquith, The relative
View/request a protocol for this paper from Bio-protocol.

1. D. J. Laydon, V. Sunkara, L. Boelen, C. R. M. Bangham, B. Asquith, The relative
View/request a protocol for this paper from Bio-protocol.

1. D. J. Laydon, V. Sunkara, L. Boelen, C. R. M. Bangham, B. Asquith, The relative
View/request a protocol for this paper from Bio-protocol.

1. D. J. Laydon, V. Sunkara, L. Boelen, C. R. M. Bangham, B. Asquith, The relative
View/request a protocol for this paper from Bio-protocol.

1. D. J. Laydon, V. Sunkara, L. Boelen, C. R. M. Bangham, B. Asquith, The relative
View/request a protocol for this paper from Bio-protocol.

1. D. J. Laydon, V. Sunkara, L. Boelen, C. R. M. Bangham, B. Asquith, The relative
View/request a protocol for this paper from Bio-protocol.

1. D. J. Laydon, V. Sunkara, L. Boelen, C. R. M. Bangham, B. Asquith, The relative
View/request a protocol for this paper from Bio-protocol.

1. D. J. Laydon, V. Sunkara, L. Boelen, C. R. M. Bangham, B. Asquith, The relative
View/request a protocol for this paper from Bio-protocol.

1. D. J. Laydon, V. Sunkara, L. Boelen, C. R. M. Bangham, B. Asquith, The relative
View/request a protocol for this paper from Bio-protocol.

1. D. J. Laydon, V. Sunkara, L. Boelen, C. R. M. Bangham, B. Asquith, The relative
View/request a protocol for this paper from Bio-protocol.

1. D. J. Laydon, V. Sunkara, L. Boelen, C. R. M. Bangham, B. Asquith, The relative
View/request a protocol for this paper from Bio-protocol.

1. D. J. Laydon, V. Sunkara, L. Boelen, C. R. M. Bangham, B. Asquith, The relative
View/request a protocol for this paper from Bio-protocol.

1. D. J. Laydon, V. Sunkara, L. Boelen, C. R. M. Bangham, B. Asquith, The relative
View/request a protocol for this paper from Bio-protocol.

1. D. J. Laydon, V. Sunkara, L. Boelen, C. R. M. Bangham, B. Asquith, The relative
View/request a protocol for this paper from Bio-protocol.

1. D. J. Laydon, V. Sunkara, L. Boelen, C. R. M. Bangham, B. Asquith, The relative
View/request a protocol for this paper from Bio-protocol.

1. D. J. Laydon, V. Sunkara, L. Boelen, C. R. M. Bangham, B. Asquith, The relative
View/request a protocol for this paper from Bio-protocol.

1. D. J. Laydon, V. Sunkara, L. Boelen, C. R. M. Bangham, B. Asquith, The relative
View/request a protocol for this paper from Bio-protocol.

1. D. J. Laydon, V. Sunkara, L. Boelen, C. R. M. Bangham, B. Asquith, The relative
View/request a protocol for this paper from Bio-protocol.
41. A. Bizhanova, P. D. Kaufman, Close to the edge: Heterochromatin at the nucleolar
39. M. Mahgoub, J.-I. Yasunaga, S. Iwami, S. Nakaoka, Y. Koizumi, K. Shimura, M. Matsuoka,
42. N. Chomont, M. El-Far, P. Ancuta, L. Trautmann, F. A. Procopio, B. Yassine-Diab,
43. C. Jiang, X. Lian, C. Gao, X. Sun, K. B. Einkauf, J. M. Chevalier, S. M. Y. Chen, S. Hua, B. Rhee,
45. L. Zhang, Y. Zhang, Y. Chen, O. Gholamalamdari, Y. Wang, J. Ma, A. S. Belmont, TSA-seq
46. C. Marchal, J. Sima, D. M. Gilbert, Control of DNA replication timing in the 3D genome.
47. N. A. Gillet, A. Melamed, C. R. M. Bangham, High-throughput mapping and clonal
48. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing
50. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis,
49. H. Li, R. Durbin, Fast and accurate short read alignment with burrows-wheeler transform.
52. C. C. Berry, C. Nobles, E. Six, Y. Wu, N. Malani, E. Sherman, A. Dryga, J. K. Everett, F. Male,
53. A. Bailey, K. Bittinger, M. J. Drake, L. Caccavelli, P. Bates, S. Hacein-Bey-Abina,
54. A R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features.
55. S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, W. Huber, BioMart
56. A. N. Engelman, M. Carrington, R. F. Siliciano, J. M. Siliciano, S. G. Deeks, B. D. Walker,
57. L. E. Finlan, D. Sproul, I. Thomson, S. Boyle, E. Kerr, P. Perry, B. Ylstra, J. R. Chubb,
58. T. W. F., E. B., and C. R. M. B. Writing—original draft: A. M., Y. W., and C. R. M. B. Project administration:
59. S. Durinck, P. Collas, Lamina-associated domains: Peripheral matters and internal affairs.
60. H. M. Amemiya, A. Kundaje, A. P. Boyle, The ENCODE blacklist: Identification
61. L. Game and colleagues in the Genomics Facility in the MRC London Institute of Medical
62. N. Briand, P. Collas, Lamina-associated domains: Peripheral matters and internal affairs.

H. G. Stunnenberg, O. Stegle, K. Downes, T. Pastinen, N. Soranzo, Genetic drivers of epigenetic and transcriptional variation in human immune cells. Cell 167, 1398–1414. e24 (2016).
38. M. R. Billman, D. Rueda, C. R. M. Bangham, Single-cell heterogeneity and cell-cycle-related viral gene bursts in the human leukaemia virus HTLV-1. Wellcome Open Research. 2, 87 (2017).
39. M. Mahgoub, J.-I. Yasunaga, S. Iwami, S. Nakaoka, Y. Koizumi, K. Shimura, M. Matsuoka, Sporadic on/off switching of HTLV-1 tax expression is crucial to maintain the whole population of virus-induced leukemic cells. Proc. Natl. Acad. Sci. U.S.A. 111, E1269–E1278 (2018).
40. Y. Satou, P. Miyazato, K. Ishihara, H. Yaguchi, A. Melamed, M. Miura, A. Fukuda, K. Nosaka, T. Watanabe, A. G. Rowan, M. Nakao, C. R. M. Bangham, The retrovirus HTLV-1 inserts an ectopic CTCF-binding site into the human genome. Proc. Natl. Acad. Sci. U.S.A. 113, 3054–3059 (2016).
41. A. Bizhanova, P. D. Kaufman, Close to the edge: Heterochromatin at the nucleolar and nuclear peripheries. Biochim. Biophys. Acta Gene Regul. Mech. 1864, 194666 (2021).
42. N. Chomont, M. El-Far, P. Ancuta, L. Trautmann, F. A. Procopio, B. Yassine-Diab, G. Boucher, M.-R. Boulassel, G. Ghattas, J. M. Brenchley, T. W. Schacker, B. J. Hill, D. C. Douek, J.-P. Routy, E. K. Haddad, R.-P. Sekaly, HIV reservoir size and persistence are driven by T cell survival and homeostatic proliferation. Nat. Med. 15, 893–900 (2009).
43. C. Jiang, X. Lian, C. Gao, X. Sun, X. Xion, E. M. Parsons, G. T. Gladkov, K. W. Seiger, J. E. Blackmer, M. Osborn, M. J. Peluso, R. Hoh, M. Somsouk, J. Milush, L. N. Bertagnolli, S. E. Sweet, J. A. Varriale, P. D. Burbelo, T.-W. Chun, G. M. Laird, E. Serrao, A. N. Engelman, M. Carrington, R. F. Siliciano, J. M. Siliciano, S. G. Deeks, B. D. Walker, M. Lichterfeld, X. G. Yu, Distinct viral reservoirs in individuals with spontaneous control of HIV-1. Nature 585, 261–267 (2020).
44. K. B. Einkauf, M. R. Osborn, C. Gao, W. Sun, X. Xion, X. Lian, E. M. Parsons, G. T. Gladkov, K. W. Seiger, J. E. Blackmer, M. Osborn, M. J. Peluso, R. Hoh, M. Somsouk, J. Milush, L. N. Bertagnolli, S. E. Sweet, J. A. Varriale, P. D. Burbelo, T.-W. Chun, G. M. Laird, E. Serrao, A. N. Engelman, M. Carrington, R. F. Siliciano, J. M. Siliciano, S. G. Deeks, B. D. Walker, M. Lichterfeld, X. G. Yu, Distinct viral reservoirs in individuals with spontaneous control of HIV-1. Nature 585, 261–267 (2020).
45. N. A. Gillet, A. Melamed, C. R. M. Bangham, High-throughput mapping and clonal quantification of retroviral integration sites. Methods Mol. Biol. 1582, 127–141 (2017).
46. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 17, 10–12 (2011).
47. H. Li, R. Durbin, Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics 25, 1754–1760 (2009).
48. H. Li, A. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 genome project data processing subgroup, the sequence alignment/ map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
49. C. C. Berry, N. A. Gillet, A. Melamed, N. Gormley, C. R. M. Bangham, F. D. Bushman, Estimating abundances of retroviral insertion sites from DNA fragment length data. Bioinformatics 28, 755–762 (2012).
Selective clonal persistence of human retroviruses in vivo: Radial chromatin organization, integration site, and host transcription

Anat Melamed, Tomas W. Fitzgerald, Yuchuan Wang, Jian Ma, Ewan Birney, Charles R. M. Bangham

Sci. Adv., 8 (17), eabm6210. • DOI: 10.1126/sciadv.abm6210

View the article online
https://www.science.org/doi/10.1126/sciadv.abm6210
Permissions
https://www.science.org/help/reprints-and-permissions