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 108 to 109 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 viral 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 presentation 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 this study confirmed the previous observation (13) of the preferential survival of HTLV-1 in vitro 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 tau = 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 tau = 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 of 10

<p>| Table 1. Datasets used in analysis. |</p>
<table>
<thead>
<tr>
<th>Virus</th>
<th>Dataset type</th>
<th>Integration sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTLV-1</td>
<td>In vitro</td>
<td>234,607</td>
</tr>
<tr>
<td>HTLV-1</td>
<td>In vivo</td>
<td>162,401</td>
</tr>
<tr>
<td>HIV-1</td>
<td>In vitro</td>
<td>65,924</td>
</tr>
<tr>
<td>HIV-1</td>
<td>In vivo</td>
<td>44,367</td>
</tr>
<tr>
<td>Random</td>
<td></td>
<td>108,976</td>
</tr>
</tbody>
</table>

Fig. 1. HTLV-1 integration site survival is biased to specific chromosomes and is nonrandomly distributed in each chromosome. (A) For each chromosome, two ratios of integration site frequencies (F) were calculated: F_{in vivo}/F_{in vitro} and F_{in vitro}/F_{random sites} (logarithmic scales). HTLV-1 survival is most strongly favored in chromosome 18. The least squares regression line is shown (dotted). (B) Heatmap of F_{in vivo}/F_{in vitro} in fixed windows across each chromosome (1-Mb wide, 10-kb steps): Blue indicates preferential clone survival in vivo, and red indicates counterselection in vivo.
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 $\ln(F_{\text{in vivo}}/F_{\text{in vitro}})$, 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 $CS_{HTLV-1}$ [defined as $\ln(F_{in\,vivo}/F_{in\,vitra})$, 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).

![Graph showing HTLV-1 survival versus distance from the centromere](image)

The HTLV-1 clone survival index $CS_{HTLV-1}$ is plotted against the logarithm of the ratio of integration site frequency in vivo and in vitro. The correlation is significant for both the short and long arms of the chromosomes.

Fig. 4. Preferential survival in vivo of infected T cell clones whose provirus lies near the nuclear lamina and distant from nuclear speckles. Data on two mid-sized chromosomes (11 and 12) are shown; data on all chromosomes are shown in Figs. S6 to S10. In each panel, the TSA-seq data on lamin A/C and SON from Chen et al. (19) are plotted against $\ln$ (integration site frequency, $F$). (A) HTLV-1 integration site frequency in vivo and in vitro. (B) The HTLV-1 clone survival index $CS_{HTLV-1}$ closely tracks the lamin A/C TSA-seq signal. (C) HIV-1 clone survival index, $CS_{HIV-1}$.

![Graph showing preferential survival in vivo](image)

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 (\(P < 10^{-16}\), 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 ~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 ~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
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 the 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 genome-intact HIV-1 proviruses in centromeric satellite DNA, especially in elite controllers). These authors concluded that persistence of intact HIV-1 proviruses favors

![Image](https://www.science.org/at/Imperial-College-London-on-May-23,2022)

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

<table>
<thead>
<tr>
<th>Predictor</th>
<th>HTLV-1</th>
<th>HIV-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>−0.12</td>
<td>0.25</td>
</tr>
<tr>
<td>Proximity P to nuclear speckle*</td>
<td>−0.38</td>
<td>−0.33</td>
</tr>
<tr>
<td>Distance D from centromere (Gb)</td>
<td>−4.36</td>
<td>−1.78</td>
</tr>
<tr>
<td>(P \times D)</td>
<td>2.87</td>
<td>NS</td>
</tr>
<tr>
<td>Local host expression intensity†</td>
<td>−0.04</td>
<td>NS</td>
</tr>
<tr>
<td>(N) (1-Mb windows)</td>
<td>2677</td>
<td>2135</td>
</tr>
<tr>
<td>(R^2) adjusted</td>
<td>0.41</td>
<td>0.07</td>
</tr>
</tbody>
</table>

*TSA-seq signal for SON. †Ln(expression signal in T cells).
of local host transcription intensity in HIV-1 infection. These positional effects appear to be stronger than the influence of the provirus: These positional effects appear to be stronger than the influence of the provirus. LADs, are reversibly associated with the nuclear lamina (43). Thus, facultative heterochromatin may be the optimal site for in vivo survival, minimizing exposure to immune selection while retaining the potential to be reexpressed. Clonal survival by eliminating the clones that most frequently transcribed, and they suggested that the HIV-1 reservoir may resemble that of HTLV-1. Our conclusions are consistent with these observations and conclusions, and extend them by demonstrating that in each virus the genome-wide importance of the position of the provirus both in the nucleus and within the chromosome. While we find that survival in HIV-1 significantly correlates with the local host expression (Fig. 6B) in univariate analysis, this correlation is not significant in the multivariate model. This observation suggests that the correlation with expression intensity is largely due to nuclear positioning (44).

In this study, we used the data on the intranuclear distribution of chromatin from the TSA-seq analysis of K562 cells by Chen et al. (19) as well as the K562 SPIN states (33). K562 cells are an erythroleukemia line, and details of the intranuclear chromatin distribution in K562 may differ from that in T cells, which are the chief host cell infected by human retroviruses. Approximately 10% of the genome differs significantly between human cell lines in intranuclear position (relative to speckles) (45). The clonal survival of these human retroviruses may therefore correlate even more strongly with the spatial distribution of chromatin in T cells.

In HTLV-1 infection, it is likely that the host immune response (10) is a major force that results in the observed pattern of selective clonal survival by eliminating the clones that most frequently express the HTLV-1 provirus, because the provirus is integrated either in a region of the host genome of intense transcriptional activity or in a region that typically lies near the center of the nucleus, near nuclear speckles, and distant from the nuclear lamina and the nucleolus. 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).

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
confidence, 1-Mb windows with at least 1-kb overlap with an ENCODE exclusion list region (53) or which include at least 1000 ambiguous bases [counted using bedtools nuc (54)] in the hg38 reference were also excluded.

The proviral integration site frequency (F) is calculated in each defined genomic region (e.g., 1-Mb window) as the proportion of all integration sites of a given dataset present in that region. We define a clone survival index (CSI) for each specified genomic region (e.g., 1-Mb window) as

\[
\text{CSI} = \ln \left( \frac{F_{\text{in vivo}}}{F_{\text{in vitro}}} \right)
\]

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, “bed narrowpeak.” 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 local coverage.” 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.6b2m6210.


45. L. Zhang, Y. Zhang, Y. Chen, O. Ghosalamandari, Y. Wang, J. Ma, A. S. Belmont, TSA-seq reveals a largely conserved genome organization relative to nuclear speckles with small position changes tightly correlated with gene expression changes. *Genome Res.*, gr.266239.120 (2020).


**Acknowledgments:** We thank A. Belmont, W. Bickmore, A. Lamond, B. McStay, and members of the Bangham group for helpful discussions. We thank G. Taylor, L. Cook, the donors and research nurses in the National Centre for Human Retrovirology, Imperial College London, and L. Game and colleagues in the Genomics Facility in the MRC London Institute of Medical Sciences, London, UK. A large dataset of in vitro integration sites of HTLV-1 was obtained by A. McCallin; we thank G. Maertens and A. McCallin for these data. Funding: This work was supported by the Wellcome Trust (https://wellcome.ac.uk/) (CRMB Investigator Award WT207477). **Author contributions:** Conceptualization: A.M., E.B., and C.R.M.B. Data curation: A.M. and T.W.F. Investigation: A.M. and T.W.F. Visualization: A.M. and Y.W. Methodology: A.M., E.B., and C.R.M.B. Resources: C.R.M.B. Supervision: J.M., E.B., and C.R.M.B. Funding acquisition: A.M. and T.W.F. Writing—review and editing: A.M., T.W.F., J.M., E.B., and C.R.M.B. Competing interests: The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Submitted 29 September 2021
Accepted 16 March 2022
Published 29 April 2022
10.1126/sciadv.abm6210
Selective clonal persistence of human retroviruses in vivo: Radial chromatin organization, integration site, and host transcription
Anat MelamedTomas W. FitzgeraldYuchuan WangJian MaEwan BirneyCharles 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