**The impact of HTLV-1 expression on the 3D structure and expression of host chromatin**

**Hiroko Yaguchi, Anat Melamed, Saumya Ramanayake, Helen Kiik, Aviva Witkover, and Charles R M Bangham**

Department of Infectious Diseases, Faculty of Medicine, Imperial College London, London, United Kingdom

Correspondence: c.bangham@imperial.ac.uk

**Abstract**

A typical HTLV-1-infected individual carries >$10^4$ different HTLV-1-infected T cell clones, each with a single-copy provirus integrated in a unique genomic site. We previously showed that the HTLV-1 provirus causes aberrant transcription in the flanking host genome and, by binding the chromatin architectural protein CTCF, forms abnormal chromatin loops with the host genome. However, it remained unknown whether these effects were exerted simply by the presence of the provirus or were induced by its transcription. To answer this question, we sorted HTLV-1-infected T-cell clones into cells positive or negative for proviral plus-strand expression, and then quantified host and provirus transcription using RNA-seq and aberrant chromatin looping using quantitative chromosome conformation capture (q4C) in each cell population. We found that proviral plus-strand transcription induces aberrant transcription and splicing in the flanking genome but suppresses aberrant chromatin loop formation with the nearby host chromatin. Reducing provirus-induced host transcription with an inhibitor of transcriptional elongation allows recovery of chromatin loops in the plus-strand-expressing population. We conclude that aberrant host transcription induced by proviral expression causes temporary, reversible disruption of chromatin looping in the vicinity of the provirus.

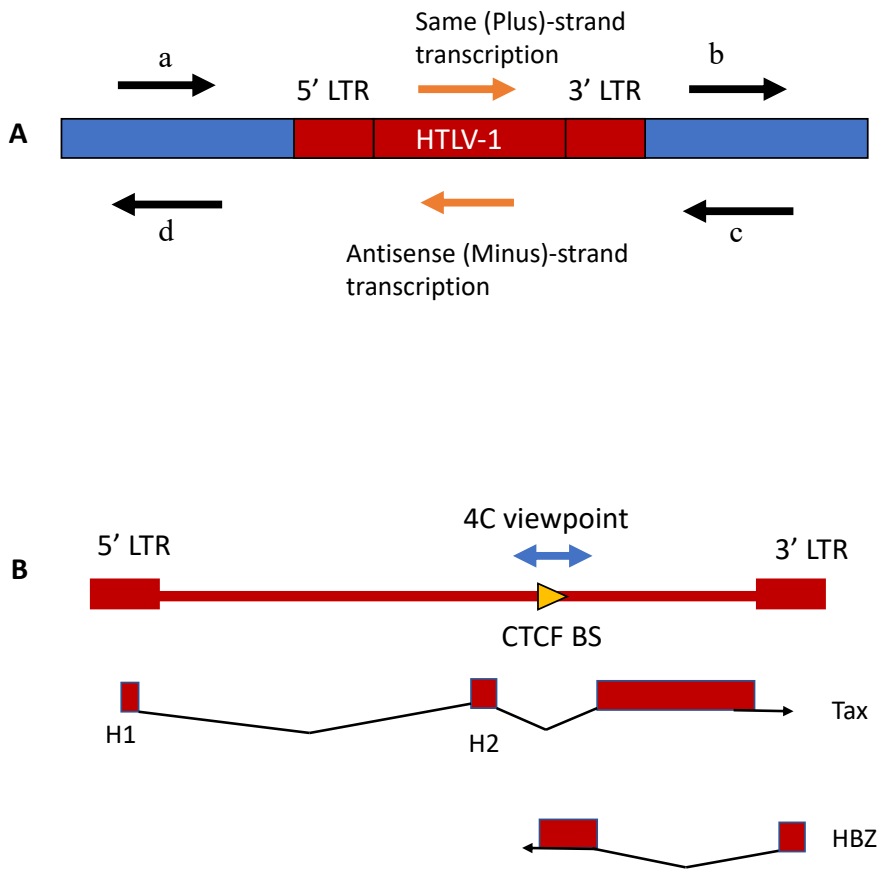
## Introduction

The 3D structure of chromatin is important in many biological processes, including transcription regulation, DNA replication, cell cycle and differentiation. The CCCTC-binding factor (CTCF) is a key player in chromatin loop formation [1, 2]. CTCF binds a non-palindromic 20-nucleotide DNA motif at ~50,000 sites in the human genome [2]; two CTCF molecules bound to different genomic sites can dimerize, forming the boundary of chromatin loops extruded through the cohesin complex. Chromatin looping plays a critical part in gene regulation by regulating the contacts between specific enhancers and promoters.

Human T cell leukemia virus type 1 (HTLV-1, also known as human T-lymphotropic virus type 1), mainly infects $CD4^+$ T-cells *in vivo.* Most carriers remain asymptomatic, but in 5% to 10% of the infected individuals, HTLV-1 infection leads to either an aggressive T-cell malignancy known as adult T-cell leukemia/lymphoma (ATL), or a chronic progressive neuro-inflammatory condition called HTLV-1-associated myelopathy/tropical spastic paraparesis (HAM/TSP; referred to hereafter as HAM) [3]. HTLV-1 persists in the host mainly by clonal proliferation of infected cells, and a typical host carries $>10^4$ long-lived HTLV-1-infected T-cell clones [4]: each clone can be distinguished by the unique integration site of the single-copy provirus in the host genome [5].

The HTLV-1 provirus has two genomic strands **(Fig 1)**. The plus strand, transcribed from the 5′ LTR, encodes the structural proteins, the transcriptional transactivator protein Tax, the regulator of mRNA splicing and transport Rex, and minor accessory proteins. The minus strand, transcribed from the 3′ LTR, encodes the regulatory protein HBZ [6].

**Fig 1**




Single-cell heterogeneity in HTLV-1 proviral expression has been shown in naturally HTLV-1-infected T cell clones [7-9]. At a given time, a small proportion of cells express intense bursts of Tax, the proportion varying between the clones [7] **(Tables 1 and 2).** By contrast, the minus strand is expressed in approximately 50% of circulating infected T cells at a given time [8]. Each HTLV-1-infected T-cell clone has its own pattern of proviral expression, that is, the frequency, intensity and duration of the transient transcriptional burst of the proviral

plus-strand [10-12]. Since the HTLV-1 proviral sequence varies little within the host, these clone-specific differences are thought to be largely due to the unique genomic integration site of the provirus; other factors may include the antigen specificity and epigenetic modifications of the host cell.

**Table 1 Clone list**

| **Clone** | **Genomic location of dominant integration site, hg 19*** | **Provirus orientation** | **Subject** | **Clone derived from** | **tax expression reporter** | **Reference** |
|---|---|---|---|---|---|---|
| HA1 | chr07:18905519 | R | HEZ | Asymptomatic carrier | - | Newly established |
| 11.63 | chr19:33829548 | F | TBW | HAM/TSP patient | - | [5, 10] |
| TBX4B | chr22:44323198 | F | TBX | HAM/TSP patient | - | [5, 10] |
| 3.60 | chr04:70567285 | F | TBJ | HAM/TSP patient | - | [5, 10] |
| d2EGFP-TBX4B | chr22:44323198 | F | TBX | HAM/TSP patient | GFP | [12] |
| d2EGFP-11.63 | chr19:33829548 | F | TBW | HAM/TSP patient | GFP | Newly established |
| d2EGFP-11.50 | chr19:28282587 | R | TBW | HAM/TSP patient | GFP | [11] |
| Timer - TBX4B | chr22:44323198 | F | TBX | HAM/TSP patient | Timer protein | [11] |
| Timer-3.60 | chr04:70567285 | F | TBJ | HAM/TSP patient | Timer protein | [11] |

**Table 2**

**Tax expression of HTLV-1-infected T cell clones**

| Clone | Tax-positive (%) |
|---|---|
| 3.60 | 4 ± 0.8 |
| TBX4B | 8 ± 0.1 |
| 11.63 | 36 ± 5.5 |
| 11.50 | 40 ± 2.5 |
| HA1 | 56 ± 3.7 |

T cell clones (clone details as shown in Table 1) were stained for Tax protein and analysed by flow cytometry. Mean ± SD (n=2 or 3) of Tax-positive fraction of viable cells are shown.

Disorganization of 3D chromatin structure can cause diseases by rewiring interactions between genes and regulatory elements [13], and certain mutations in CTCF are linked with human disease [14-16]. We previously showed that the HTLV-1 provirus contains a CTCF binding site (BS): the provirus creates novel loops with the host genome, inducing transcriptional deregulation in the host genome flanking the provirus [10, 17]. However, it remained unknown whether these effects were exerted simply by the presence of the provirus or by its transcription.

To answer this question, we sorted naturally-infected HTLV-1$^{+}$ T cell clones to obtain subsets of provirus-expressing cells and non-expressing cells, and analysed the effects of proviral expression on chromatin looping, using quantitative circular chromosome conformation capture (q4C) [10], and on host transcription, using RNA-seq.

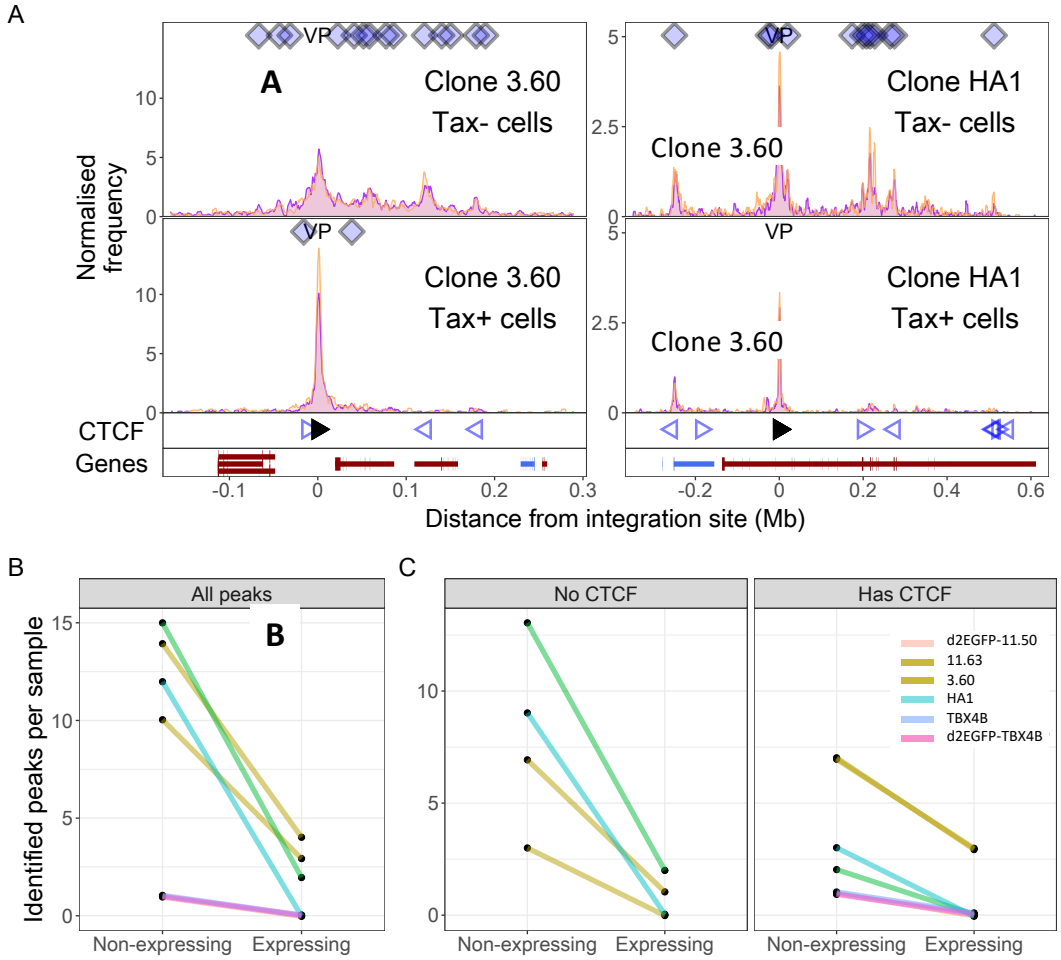
## Results

### HTLV-1 expression proviral plus-strand expression reduces chromatin looping between the provirus and the host genome.

We hypothesized that the chromatin loops between the provirus and host genome region are regulated by expression of the provirus. To test the hypothesis, we focused on the proviral plus-strand, which is expressed in intense intermittent transcriptional bursts [7]. We used the HTLV-1 transcriptional transactivator protein Tax as a marker of plus-strand expression. T cell clones naturally infected with HTLV-1, each of which has a unique integration site (**Table 1**), were crosslinked with paraformaldehyde, flow-sorted into $Tax^{+}$ and $Tax^{-}$ populations (**S1 Fig**), and subjected to q4C assay, using a fragment of the proviral sequence containing its CTCF binding site as the q4C viewpoint [10] (Fig 1B).

In the $Tax^{-}$ population of clones 3.60 and HA1, we identified q4C peaks (long-range chromatin loops between the provirus and the host genome) **(Fig 2A):** the identified peaks often overlapped CTCF binding sites, as previously reported in the study using unsorted T cells [10].

**Fig 2**




Unexpectedly, the frequency of the chromatin loops identified in Tax⁻ cells was much lower in Tax⁺ cells, although the technical peak at the viewpoint (VP), which is always present in a successful 4C-seq assay [18], remained in Tax⁺ cells (**Fig 2A**).

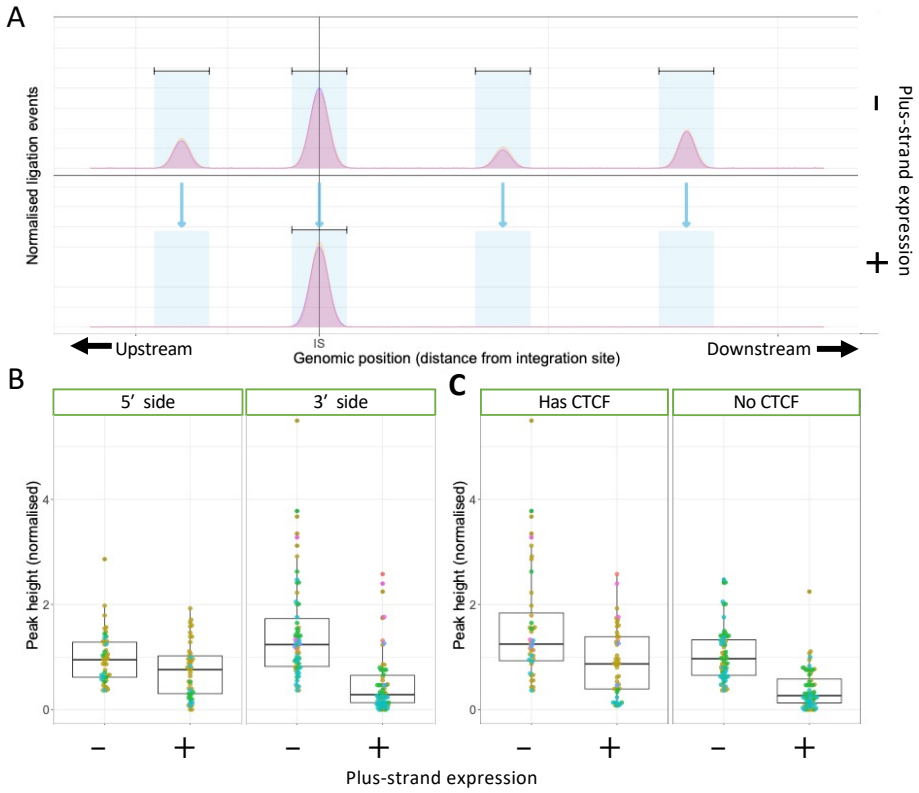
Tax protein is a transcriptional transactivator that stimulates transcription of both HTLV-1 proviral genes and many host genes. To exclude the possibility that the anti-Tax antibody used in intracellular staining altered chromatin looping, we performed q4C assays using T-cell clones transduced with a reporter construct which expresses a modified EGFP with a half-life of ~2h (d2EGFP-TBX4B) when stimulated by Tax protein [11]. The GFP signal intensity was positively correlated with Tax expression [12]. Provirus-expressing ($GFP^+$) cells and non-expressing ($GFP^-$) cells were flow-sorted after fixation in 1% paraformaldehyde and subjected to the q4C assay. Consistent with the results from intracellular Tax staining (Fig 2A), $GFP^+$ cells also showed fewer chromatin loops compared with $GFP^-$ cells (**S2 Fig**), and the 4C profiles of $GFP^+$ d2EGFP-TBX4B cells were consistent with those of sorted $Tax^+$ TBX4B cells. We conclude that the lower frequency of chromatin looping observed in $Tax^+$ cells was associated with proviral transcription and was not caused by staining Tax protein.

A decrease in the number of q4C peaks at distant chromatin contacts in the host genome was observed in provirus plus-strand-expressing subsets in all T cell clones investigated (**Fig 2B, S2 Fig, S3 Fig**), in regions either with or without CTCF binding sites (**Fig 2C**).

To corroborate these observations by an independent technique, 3C-qPCR was performed to quantify the frequency of the interaction between the provirus and the two host regions where the principal 4C peaks were identified in the $Tax^-$ population of clone 11.63 (**S3 Fig**), using specific primers designed to amplify the junction between the provirus and the host genome (**S4 Fig A, Table 3**). The results showed a reproducible significant decrease in chromatin contact frequency between the provirus and both peaks in $Tax^+$ cells (**S4 Fig C, D**), although there was no difference in the contact frequency observed within the provirus (**S4 Fig B**).

Next, to quantify the difference in chromatin loop formation between provirus-expressing and non-expressing cells, peak regions were defined in the non-expressing cell population, and peaks were then sought at the same sites in the provirus-expressing cell population. (**Fig 3A**) At each site, we then compared the normalized peak height of the q4C profile in non-expressing cells with that in provirus-expressing cells. The results show that HTLV-1 plus-strand expression resulted in a significant reduction in the frequency of these chromatin loops, especially on the 3′ side of the integration site (**Fig 3B**). This reduction was observed in both regions with and without CTCF binding sites (**Fig 3C**).

**Fig 3**




**HTLV-1 proviral plus-strand expression drives host transcription downstream *in cis*.** Throughout this work, we refer to the direction of flanking host transcription relative to the orientation of the provirus. Thus, host transcription upstream (5′) of the 5′LTR, in the same sense as the proviral plus-strand, is denoted as ‘same-sense, 5′ side of provirus’ **(Fig 1 arrow**

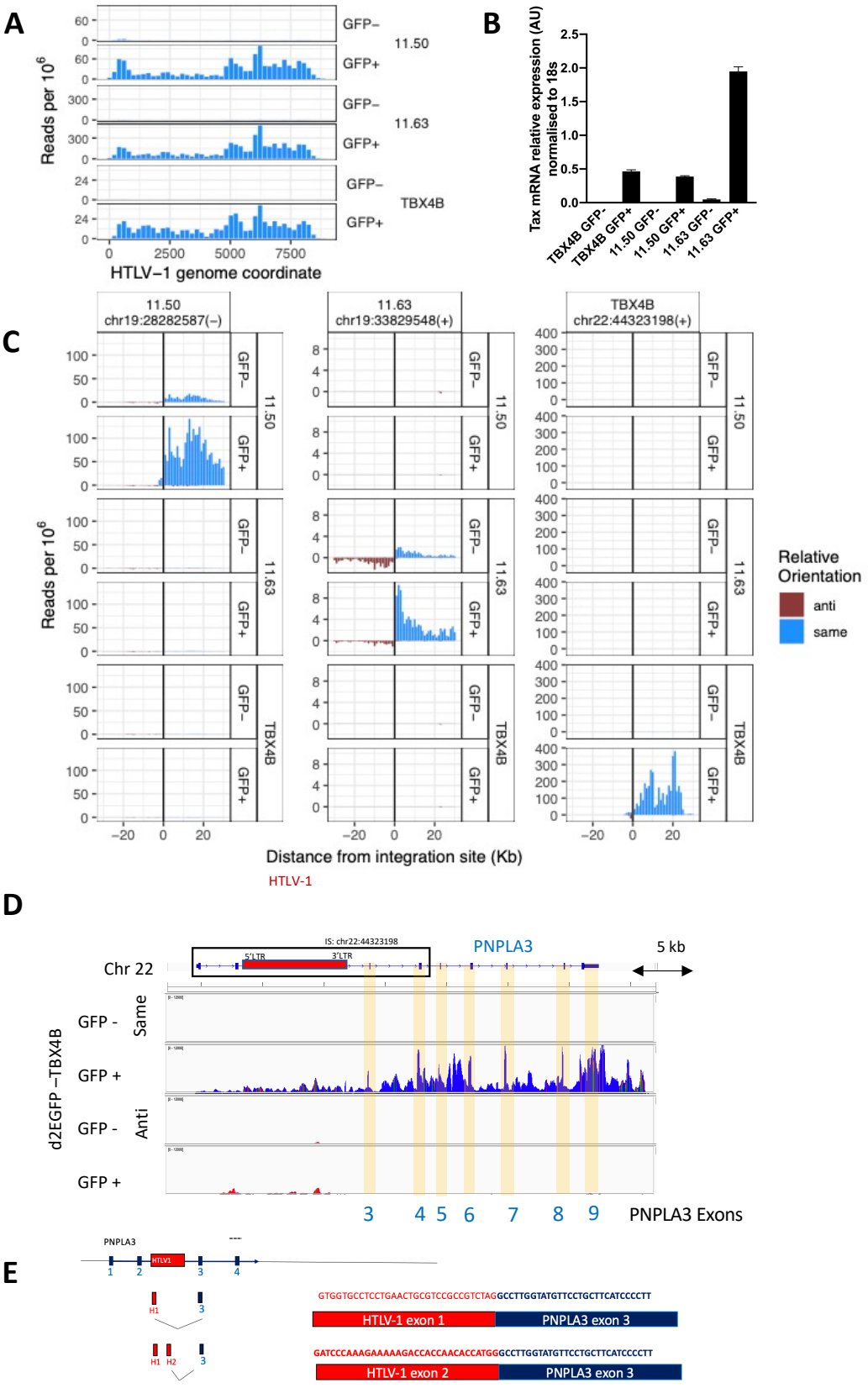
**a)**, and transcription in the same sense 3′ of the 3′LTR as 'same-sense, 3′ side of provirus' **(Fig 1 arrow b)**. Similarly, we refer to host transcription in the same sense as the proviral minus-strand as 'antisense, 3′ side of provirus' **(Fig 1 arrow c)** and 'antisense, 5′ side of provirus' **(Fig 1 arrow d)**.

Next, we asked whether the aberrant host transcription is associated with proviral transcription or simply with the presence of the provirus, regardless of transcription. The landscape of host transcription flanking the provirus has been investigated in ATL cases [19, 20], in which there is a dominant single malignant clone. However, there are typically widespread epigenetic and transcriptional abnormalities in malignant cells, and expression of the proviral plus-strand is frequently lost in ATL clones [21]. We previously reported abnormal *in cis* host transcription near the integration site in naturally-infected, non-malignant HTLV-1-positive T-cell clones [10], but since only a fraction of cells express the proviral plus-strand at a given instant, we could not distinguish whether the abnormal host transcription was associated with proviral plus-strand expression, or whether there was constitutive activation irrespective of the plus-strand burst, perhaps by Tax-mediated transactivation or by the proviral enhancer.

RNA isolated from fresh, unfixed cells is required for optimal RNA-seq results. However, intracellular staining of Tax protein requires prior fixation. Therefore, to elucidate whether the aberrant host transcription was induced by proviral plus-strand transcription, we used d2EGFP-transduced clones (11.63, 11.50 and TBX4B) (see Table 1) [11], from which $GFP^+$ (Tax-expressing) and $GFP^-$ (Tax-non-expressing) cells were isolated by live-cell sorting; total cellular RNA was extracted and subjected to a stranded RNA-seq analysis. The proviral plus-strand was highly transcribed in all 3 clones in $GFP^+$ cells but not in $GFP^-$ cells (**Fig 4A and**

**B)**, confirming that $GFP^+$ cell-sorting enriched a cell population with active HTLV-1 plus-strand transcription.

**Fig 4**


A
Reads per 10^6
GFP−
GFP+
11.50
11.63
TBX4B
HTLV−1 genome coordinate
B
Tax mRNA relative expression (AU) normalised to 18s
TBX4B GFP-
TBX4B GFP+
11.50 GFP-
11.50 GFP+
11.63 GFP-
11.63 GFP+
C
11.50 chr19:28282587(−)
11.63 chr19:33829548(+)
TBX4B chr22:44323198(+)
Distance from integration site (Kb)
HTLV-1
Relative Orientation
anti
same
D
Chr 22
IS: chr22:44323198
5'LTR
3'LTR
PNPLA3
5 kb
d2EGFP −TBX4B
GFP -
GFP +
Same
Anti
3 4 5 6 7 8 9
PNPLA3 Exons
E
PNPLA3
HTLV1
H1
H2
GTGGTGCCTCCTGAACTGCGTCCGCCGTCTAGGCCTTGGTATGTTCCTGCTTCATCCCCTT
HTLV-1 exon 1
PNPLA3 exon 3
GATCCCAAAGAAAAAGACCACCAACACCATGGGCCTTGGTATGTTCCTGCTTCATCCCCTT
HTLV-1 exon 2
PNPLA3 exon 3

Host transcription in the same sense on the 3′ side of the provirus near the clone's respective proviral integration site was greater in GFP$^+$ cells than in GFP$^-$ cells in all 3 clones (**Fig 4C**). For example, in clone d2EGFP-TBX4B, HTLV-1 is integrated between exon 2 and exon 3 of the *PNPLA3* gene (**Fig 4D**) , which is not normally expressed in T cells (**Fig 4C**). The *PNPLA3* gene (same-sense, 3′ side of the provirus) was highly transcribed in the GFP$^+$ population, but was not transcribed in GFP$^-$ cells of clone d2EGFP-TBX4B, or in two other clones (d2EGFP-11.50 and d2EGFP-11.63) regardless of Tax expression, suggesting that the observed increase in *PNPLA3* mRNA did not result from Tax-mediated transactivation (**Fig 4C**). Aberrant transcription in the same-sense on the 5′ side of the provirus was also seen in two clones (11.50 and TBX4B) at a low level (**Fig 4C, D**).

In addition to splicing between the 3′LTR and the host genome (antisense, 5′ side of provirus) (data not shown), which has also been reported in ATL cases [19], in plus-strand-expressing cells we found that the exons of 5′LTR (exon H1/exon H2 in **Fig 1**) were spliced out to form not only normal viral mRNAs but also viral-host fusion transcripts, by fusion to downstream host plus-strand transcripts. For example, in the GFP$^+$ subset of d2EGFP-TBX4B cells, proviral exons H1 or H2 were found to be fused to exon 3 of *PNPLA3* (**Fig 4D, E**). The observed events always occurred between a canonical HTLV-1 splice donor and a canonical host splice acceptor. Such splicing events were observed even when no host gene was present in the flanking region (**S5 Fig**). We then reanalysed the RNA-seq data obtained previously [11] and found a novel transcript in clone Timer-3.60 cells expressing the plus-strand, formed by splicing out of the proviral plus-strand exon H1 and fusion to novel host exons (**S5 Fig B**). These splicing events again followed the canonical GT|AG mRNA processing rule. Fusion was observed between a host splice acceptor and proviral splice donor up to 120 kb away. By contrast, splicing was not observed between a host splice donor and an HTLV-1 splice

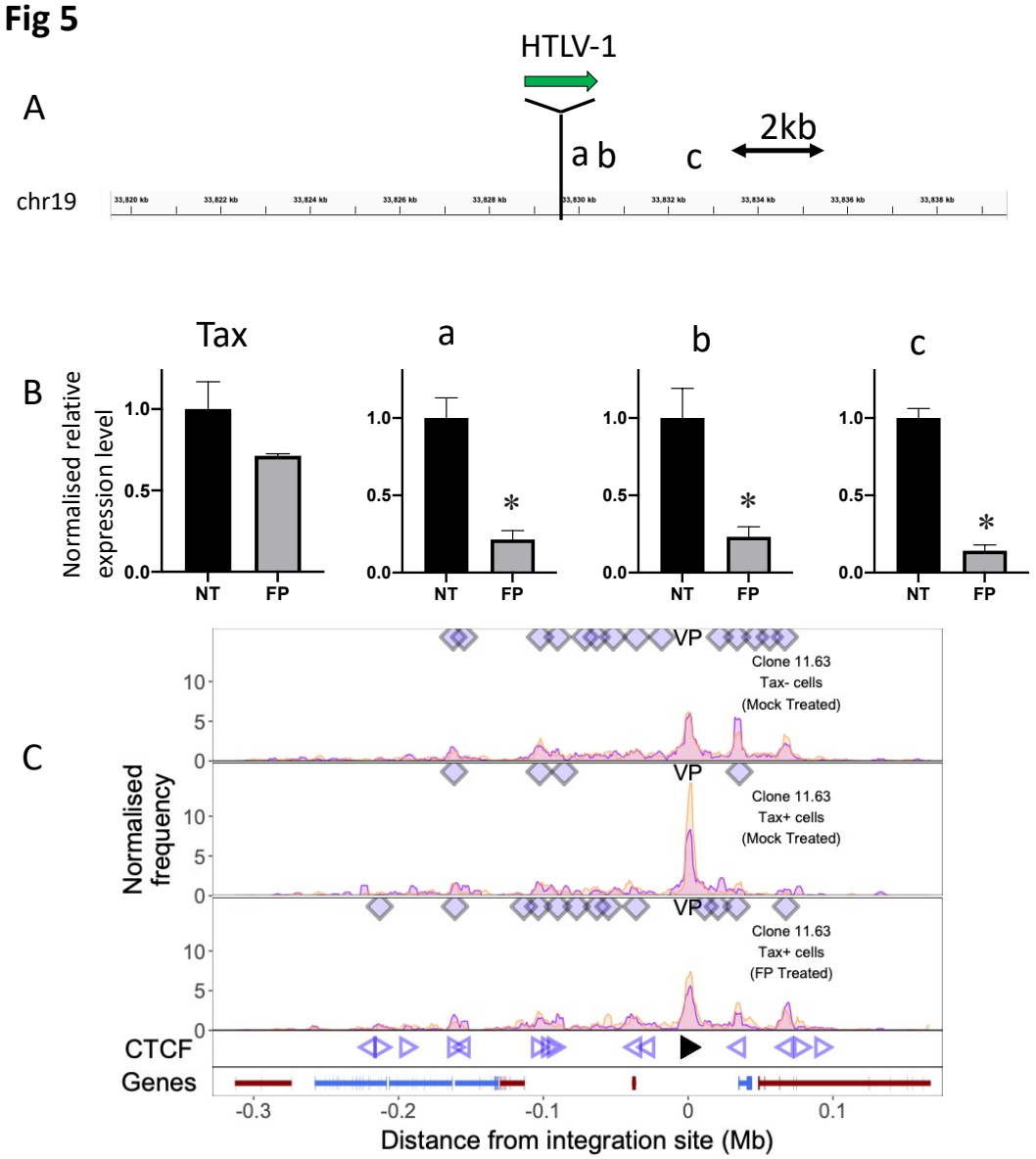
acceptor; although host transcription in the same-sense on the 5′ side of the provirus was induced by proviral plus-strand expression. (**Fig 4 C, D**).

**Treatment of T cell clones with an inhibitor of transcriptional elongation allows recovery of chromatin loops in the Tax$^+$ population.**

The observation of upregulated host transcription near the integration site and decreased chromatin loops in the provirus-expressing cell populations raised the question whether provirus-induced host transcription disrupts chromatin loop formation between the provirus and host genome. To test this hypothesis, we treated the infected cells with flavopiridol, an inhibitor of transcriptional elongation. Flavopiridol was reported to inhibit both elongation and readthrough transcription induced by influenza virus [22]. Treatment of cells with 1 nM flavopiridol for 1.5 hours did not affect Tax protein expression (S6 Fig) so the cells could still be sorted for Tax protein. Flavopiridol treatment also did not reduce *tax* mRNA expression significantly. However, the level of aberrant host transcription in the same sense, 3′ side of the provirus was significantly decreased (**Fig 5A, B**).

After flavopiridol treatment, the cells were stained intracellularly, sorted for Tax protein, and then subjected to q4C assay. While the frequency of chromatin loops between the provirus and the host genome was lower in Tax$^+$ cells than in Tax$^-$ cells (as in Fig.1 above), chromatin loops remained in Tax$^+$ cells when they were treated with flavopiridol (**Fig 5C**). We conclude that inhibition of transcription elongation by flavopiridol reduced host transcription and allowed the preservation and restoration of chromatin loops.

**Fig 5**


A
HTLV-1
a b
c
2kb
chr19
B
Tax
a
b
c
Normalised relative expression level
1.0
0.5
0.0
NT
FP
*
C
VP
Clone 11.63
Tax- cells
(Mock Treated)
Clone 11.63
Tax+ cells
(Mock Treated)
Clone 11.63
Tax+ cells
(FP Treated)
Normalised frequency
10
5
0
CTCF
Genes
-0.3
-0.2
-0.1
0
0.1
Distance from integration site (Mb)

## Discussion

The results presented here lead to two main conclusions. First, 4C analysis of sorted T cell clones for plus-strand expression revealed that the chromatin loops formed between the provirus and the immediately flanking host genome are reduced or lost during plus-strand proviral transcription. Second, plus-strand proviral transcription causes (1) aberrant host transcription on the same-strand, chiefly on the 3′ side of the provirus and to a lesser extent on the 5′ side; (2) fusion transcripts between provirus and host RNA; the 5′LTR (exon H1/exon H2) was fused to host exons in the same sense on the 3′ side of the provirus, although intra–HTLV-1 splicing remained intact.

In clone TBX4B we observed that transcriptional activation of a host gene more distant (1.4 Mb) from the provirus was associated with expression of the wild-type provirus, but not with the provirus in which the CTCF binding site was knocked out (**S7 Fig**). We postulate that, while the nearby chromatin loops are diminished or lost during HTLV-1 proviral plus-strand transcription, more long-range CTCF-dependent loops can remain intact.

Previous studies reporting RNA-seq analysis in ATL cases [19, 20] showed aberrant antisense expression in the host genome flanking the 5′ side of the provirus; the 3′LTR sequence was spliced out and fused to host exons on the antisense strand on the 5′ side of the provirus. However, aberrant host transcription in the same sense as the proviral plus-strand (either 5′ or 3′ to the provirus), and splicing between the 5′LTR and host exons in the same sense on the 3′ side of the provirus, were rarely observed. In the non-malignant clones in the

present study, although viral-host transcripts in the minus-strand were still observed (data not shown), plus-strand transcription always dominated.

We observed splicing between the 5′LTR and host exons in the plus-strand on the 3′ side of the provirus **(Fig 4D, E, S5 Fig),** even when there was no identified host gene in either strand. The observation that proviral-host chimeric RNA splicing events can activate cryptic host exons was recently reported in HIV-1-infected Jurkat T cell clones carrying reporter proviruses integrated into the introns of three cancer-related genes [23]. Our observations show that HTLV-1 plus-strand expression has the potential to disrupt host cell functions, by creating novel transcripts, with potential biological functions, under the transcriptional control of the HTLV-1 5′LTR.

Whereas frequent splicing was observed between a proviral donor and host acceptor, we did not observe splicing between a host donor and proviral acceptor, even if the host genome 5′ side region was transcribed. This observation implies that same-sense host transcription on the 5′ side of the provirus stops at the integration site, whereas LTR-driven transcription can result in read-through into the host genome in the same sense on the 3′ side of the provirus. CTCF, which binds to the HTLV-1 provirus, can regulate mRNA splicing: further work is required to test the possibility that CTCF bound to the provirus causes aberrant splicing in the flanking host genes.

The results reported here show that the aberrant host transcription induced by the HTLV-1 provirus differs from the host gene transcriptional landscape at the HIV-1 integration site [23]. No aberrant host transcription was observed in the antisense strand in HIV-1-infected cells. By contrast with HIV-1, HTLV-1 can induce host transcription in the same sense on the

5′ side of the provirus (**Fig 4C**), whereas both proviruses induce strong same-sense transcription on the 3′ side of the provirus. Whereas downstream transcription can be at least partially explained as readthrough transcripts from the provirus, same-sense transcription on the 5′ side of the provirus in an HTLV-1$^+$ clone cannot be explained by readthrough from the 5′LTR. Two observations suggest that this same-sense transcription on the 5′ side of the provirus is due to an enhancer effect of the HTLV-1 5′LTR. First, this aberrant same-sense transcription was seen only in GFP$^+$ cells (**Fig 4C, D**); second, the transcription correlated with the level of plus-strand proviral transcription in timer protein reporter clones (S8 Fig).

An unexpected finding in this study was that aberrant loop formation between the provirus and host chromatin is suppressed by proviral plus-strand transcription. Since CTCF appears to remain bound to the provirus during expression of *tax* [8], the observed decrease in chromatin looping cannot be attributed to loss of CTCF. The observations that proviral transcription increases aberrant transcription and splicing in the flanking host genome (**Fig 4 and S5 Fig**) and that reducing proviral-expression-induced host transcription with an inhibitor of transcriptional elongation allows recovery of chromatin loops in the Tax$^+$ population (**Fig 5**) suggest that aberrant host transcription induced by the proviral expression disrupts chromatin looping. This inference is consistent with the report by Heinz et al. (2018) [22] that during influenza A virus (IAV) infection the IAV non-structural protein 1 (NS1) induces global inhibition of transcription termination of highly transcribed genes and causes readthrough transcription for hundreds of kilobases, resulting in disruption of chromatin interactions.

Chromatin looping between the HTLV-1 provirus and the host genome is mediated by CTCF binding to the provirus [14]. We postulated that CTCF binding confers a selective advantage

on the provirus; however, the mechanism of this putative advantage is not known. At least three non-mutually-exclusive possibilities can be identified. First, through its function as an epigenetic barrier, CTCF may regulate epigenetic modification of the provirus [24] Second, CTCF (again through its barrier function) might block unwanted activation of plus-strand proviral transcription by the recently reported intragenic enhancer [25]. Third, CTCF may control the intranuclear position of the chromatin containing the provirus, which is correlated with selective survival of HTLV-1-infected T cell clones in vivo [26].

## Methods

### Cells

The details of HTLV-1-infected T-lymphocyte clones used in this study are shown in Table 1. All clones were derived as previously described [11] from peripheral blood mononuclear cells (PBMCs) of donors attending the National Centre for Human Retrovirology (NCHR) at Imperial College Healthcare NHS Trust, St Mary's Hospital, London. All donors gave written informed consent in accordance with the Declaration of Helsinki to donate blood samples to the Communicable Diseases Research Tissue Bank, approved by the UK National Research Ethics Service (15/SC/0089). The identification of genomic insertion sites by LMPCR was described elsewhere [27]. The cells were maintained in RPMI-1640 (Sigma, R0883) supplemented with L-glutamine, penicillin+streptomycin and 20% fetal bovine serum (Gibco, 10500-064) in 5% $CO_2$ at 37ºC. IL-2 (Miltenyi Biotec, 130-097-745) was supplemented (100 unit/ml) into the culture twice a week. The integrase inhibitor raltegravir (Selleck Chemicals, MK-0518) was used at 10 μM throughout the culture to prevent secondary infection.

### 3C and q4C assay of sorted cells

HTLV-1- infected T cell clones were stained with LIVE/DEAD Fixable Near-IR Dead Cell Stain kit (Invitrogen, L34976) to enabling gating on live cells and then crosslinked in phosphate-buffered saline (PBS) containing 1% formaldehyde for 10 min at room temperature. Then the Tax protein was stained intracellularly with anti-Tax-AF647 (clone LT-4; 0.4 μg/ml), using a Foxp3 staining kit (eBioscience, 00-5523-00) and viable $Tax^+$ and Tax- cells were sorted (**S1 Fig)**. For the Tax expression reporter clones (d2EGFP clones), the cells were stained with LIVE/DEAD Fixable Near-IR Dead Cell Stain kit, crosslinked as above, and viable $GFP^+$ cells and GFP- cells were sorted. After sorting, q4C was performed [10].

q4C analysis was carried out as previously described [10]. Reads spanning 100 to 150 bp were inspected for quality using FastQC, and filtered for correct amplification using the NlaIII restriction sequence and preceding 4 bases (total 8 bases) as an identifying barcode for the q4C viewpoint using Cutadapt. Reads where the first NlaIII site was incompletely digested (therefore containing an additional NlaIII fragment) were further trimmed by Cutadapt. Finally, Trimgalore was used to trim low quality reads (<20) and remove the Ilumina adaptor sequence, keeping reads of minimum 30 b in length. Trimmed reads were aligned to a combined reference of human (hg19) and viral (AB513134) genomes using Bowtie2. Ligation sites were quantified from aligned reads using Perl and R scripts. Ligation sites were quantified in windows of 10 kb across the chromosomes in order of base coordinates. Peaks were called as described previously [10] using a three-state hidden Markov model, requiring agreement between both samples (biological replicates).

Real-time 3C qPCR was described elsewhere [17]. Primer pairs and probe (**Table 3**) were used to detect long-range chromatin contacts between the provirus and host genome region at Peak 1 (34.5 kb downstream of the provirus) and Peak 2 (68 kb downstream of the provirus)

of chromosome 19. As control, we used a primer pair and Taqman probe to detect the contacts of two regions in the provirus, and a 101 bp sequence from the provirus was amplified to normalize the qPCR using internal control primer sets and probe (**Table 3**). Data analysis was done with LinRegPCR software (version 2014.5). Six replicates were performed on each library, using a QuantStudio 7 Flex Real-Time PCR System utilizing TaqMan Gene Expression Master Mix (Applied Biosystems).

**RNA-seq and RT-qPCR**

d2EGFP clones were stained with LIVE/DEAD, and live cells were sorted with a BD FACSAria lll cell sorter under containment level 3 (CL3) conditions, as described elsewhere [11]. Total RNA was extracted from the sorted cells using an miRNeasy kit (Qiagen) and RNA-seq libraries were prepared using the Ribo-Zero Plus rRNA Depletion kit and TruSeq Stranded mRNA HT Sample Prep Kit and sequenced with the NovaSeq6000 (150 bp paired-end reads).

RNA-seq reads were inspected for quality using FastQC, and trimmed using Trimgalore to remove low-quality reads and Ilumina adapter sequences. Trimmed reads were aligned against a combined reference of human (hg19) and viral (AB513134) genomes using GSNAP v. 2019-06-10. Read coverage was counted using Bedtools coverage after split to exons using Bedtools bamtobed tool using the resolution denoted in each figure.

Reads were aligned to a combined reference of human (hg19) and viral (AB513134) inserted into the integration site at chr22:44323198, chr19:33829548 and chr04:70567285 to identify splice sites of fusion transcripts in TBX4B, 11.63 and 3.60, respectively.

RNA-seq of timer protein reporter clones was described elsewhere [11]. RNA-seq of TBX4B clones containing wild-type CTCF Binding site (BS) in the provirus and mutant (Mut) clone containing a mutated CTCF-BS were performed as described elsewhere [10].

For RT-qPCR, first-strand cDNA was synthesized with First Strand cDNA Synthesis Kit using random primers. Primers used are shown **in Table 3**. Real-time qPCR was then performed in 6 replicates on each library, using a QuantStudio 7 Flex Real-Time PCR System utilizing Fast SYBR™ Green Master Mix (Applied Biosystems).

**Acknowledgements**

We thank Parisa Amjadi from the CL3 Cell Sorting Facility at The Centre for Immunology and Vaccinology at Imperial College London for cell sorting. We are grateful to Tomas Fitzgerald at the European Bioinformatics Institute and Aris Aristodemou for helpful discussion. We thank Laurence Game at MRC London Institute of Medical Sciences for q4C sequencing, and Oxford Genomics Centre for cDNA library preparation and RNA sequencing.

## References


1. Baranello L, Kouzine F, Levens D. CTCF and cohesin cooperate to organize the 3D structure of the mammalian genome. Proc Natl Acad Sci U S A. 2014;111(3):889-90. Epub 2014/01/09. doi: 10.1073/pnas.1321957111. PubMed PMID: 24398527; PubMed Central PMCID: PMCPMC3903212.
2. Phillips JE, Corces VG. CTCF: master weaver of the genome. Cell. 2009;137(7):1194-211. Epub 2009/07/01. doi: 10.1016/j.cell.2009.06.001. PubMed PMID: 19563753; PubMed Central PMCID: PMCPMC3040116.
3. Bangham CRM, Matsuoka M. Human T-cell leukemia virus types 1 and 2. In: Howley PM, Knipe DM, editors. Fields Virology. RNA Viruses. 3. 7th ed: Wolters Kluwer; 2023. p. 527-57.
4. Laydon DJ, Sunkara V, Boelen L, Bangham CRM, Asquith B. The relative contributions of infectious and mitotic spread to HTLV-1 persistence. PLoS Comput Biol. 2020;16(9):e1007470. Epub 2020/09/18. doi: 10.1371/journal.pcbi.1007470. PubMed PMID: 32941445; PubMed Central PMCID: PMCPMC7524007.
5. Cook LB, Rowan AG, Melamed A, Taylor GP, Bangham CR. HTLV-1-infected T cells contain a single integrated provirus in natural infection. Blood. 2012;120(17):3488-90. Epub 2012/09/08. doi: 10.1182/blood-2012-07-445593. PubMed PMID: 22955925; PubMed Central PMCID: PMCPMC3482858.
6. Kulkarni A, Bangham CRM. HTLV-1: Regulating the Balance Between Proviral Latency and Reactivation. Front Microbiol. 2018;9:449. Epub 2018/04/05. doi: 10.3389/fmicb.2018.00449. PubMed PMID: 29615991; PubMed Central PMCID: PMCPMC5867303.
7. Billman MR, Rueda D, Bangham CRM. Single-cell heterogeneity and cell-cycle-related viral gene bursts in the human leukaemia virus HTLV-1. Wellcome Open Res. 2017;2:87. Epub 2017/10/25. doi: 10.12688/wellcomeopenres.12469.2. PubMed PMID: 29062917; PubMed Central PMCID: PMCPMC5645716.
8. Miura M, Miyazato P, Satou Y, Tanaka Y, Bangham CRM. Epigenetic changes around the pX region and spontaneous HTLV-1 transcription are CTCF-independent. Wellcome Open Res. 2018;3:105. Epub 2019/01/08. doi: 10.12688/wellcomeopenres.14741.2. PubMed PMID: 30607369; PubMed Central PMCID: PMCPMC6305241.
9. Mahgoub M, Yasunaga JI, Iwami S, Nakaoka S, Koizumi Y, Shimura K, et al. 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. 2018;115(6):E1269-E78. Epub 2018/01/24. doi: 10.1073/pnas.1715724115. PubMed PMID: 29358408; PubMed Central PMCID: PMCPMC5819419.
10. Melamed A, Yaguchi H, Miura M, Witkover A, Fitzgerald TW, Birney E, et al. The human leukemia virus HTLV-1 alters the structure and transcription of host chromatin in cis. Elife. 2018;7. Epub 2018/06/26. doi: 10.7554/eLife.36245. PubMed PMID: 29941091; PubMed Central PMCID: PMCPMC6019074.
11. Kiik H, Ramanayake S, Miura M, Tanaka Y, Melamed A, Bangham CRM. Time-course of host cell transcription during the HTLV-1 transcriptional burst. PLoS Pathog. 2022;18(5):e1010387. Epub 2022/05/17. doi: 10.1371/journal.ppat.1010387. PubMed PMID: 35576236.
12. Ramanayake S, Moulding DA, Tanaka Y, Singh A, Bangham CRM. Dynamics and consequences of the HTLV-1 proviral plus-strand burst. PLoS Pathog. 2022;18(11):e1010774. Epub 2022/11/29. doi: 10.1371/journal.ppat.1010774. PubMed PMID: 36441826; PubMed Central PMCID: PMCPMC9731428.

13. Anania C, Lupianez DG. Order and disorder: abnormal 3D chromatin organization in human disease. Brief Funct Genomics. 2020;19(2):128-38. Epub 2020/02/07. doi: 10.1093/bfgp/elz028. PubMed PMID: 32025693; PubMed Central PMCID: PMCPMC7115703.
14. Boyling A, Perez-Siles G, Kennerson ML. Structural Variation at a Disease Mutation Hotspot: Strategies to Investigate Gene Regulation and the 3D Genome. Front Genet. 2022;13:842860. Epub 2022/04/12. doi: 10.3389/fgene.2022.842860. PubMed PMID: 35401663; PubMed Central PMCID: PMCPMC8990796.
15. Spielmann M, Lupianez DG, Mundlos S. Structural variation in the 3D genome. Nat Rev Genet. 2018;19(7):453-67. Epub 2018/04/26. doi: 10.1038/s41576-018-0007-0. PubMed PMID: 29692413.
16. Norton HK, Phillips-Cremins JE. Crossed wires: 3D genome misfolding in human disease. J Cell Biol. 2017;216(11):3441-52. Epub 2017/09/01. doi: 10.1083/jcb.201611001. PubMed PMID: 28855250; PubMed Central PMCID: PMCPMC5674879.
17. Satou Y, Miyazato P, Ishihara K, Yaguchi H, Melamed A, Miura M, et al. The retrovirus HTLV-1 inserts an ectopic CTCF-binding site into the human genome. Proc Natl Acad Sci U S A. 2016;113(11):3054-9. Epub 2016/02/29. doi: 10.1073/pnas.1423199113. PubMed PMID: 26929370; PubMed Central PMCID: PMCPMC4801255.
18. Krijger PHL, Geeven G, Bianchi V, Hilvering CRE, de Laat W. 4C-seq from beginning to end: A detailed protocol for sample preparation and data analysis. Methods. 2020;170:17-32. Epub 2019/07/29. doi: 10.1016/j.ymeth.2019.07.014. PubMed PMID: 31351925.
19. Kataoka K, Nagata Y, Kitanaka A, Shiraishi Y, Shimamura T, Yasunaga J, et al. Integrated molecular analysis of adult T cell leukemia/lymphoma. Nat Genet. 2015;47(11):1304-15. Epub 2015/10/06. doi: 10.1038/ng.3415. PubMed PMID: 26437031.
20. Rosewick N, Durkin K, Artesi M, Marcais A, Hahaut V, Griebel P, et al. Cis-perturbation of cancer drivers by the HTLV-1/BLV proviruses is an early determinant of leukemogenesis. Nat Commun. 2017;8:15264. Epub 2017/05/24. doi: 10.1038/ncomms15264. PubMed PMID: 28534499; PubMed Central PMCID: PMCPMC5457497.
21. Takeda S, Maeda M, Morikawa S, Taniguchi Y, Yasunaga J, Nosaka K, et al. Genetic and epigenetic inactivation of tax gene in adult T-cell leukemia cells. Int J Cancer. 2004;109(4):559-67. Epub 2004/03/03. doi: 10.1002/ijc.20007. PubMed PMID: 14991578.
22. Heinz S, Texari L, Hayes MGB, Urbanowski M, Chang MW, Givarkes N, et al. Transcription Elongation Can Affect Genome 3D Structure. Cell. 2018;174(6):1522-36.e22. Epub 2018/08/23. doi: 10.1016/j.cell.2018.07.047. PubMed PMID: 30146161; PubMed Central PMCID: PMCPMC6130916.
23. Liu R, Yeh YJ, Varabyou A, Collora JA, Sherrill-Mix S, Talbot CC, Jr., et al. Single-cell transcriptional landscapes reveal HIV-1-driven aberrant host gene transcription as a potential therapeutic target. Sci Transl Med. 2020;12(543). Epub 2020/05/15. doi: 10.1126/scitranslmed.aaz0802. PubMed PMID: 32404504; PubMed Central PMCID: PMCPMC7453882.
24. Cheng X, Joseph A, Castro V, Chen-Liaw A, Skidmore Z, Ueno T, et al. Epigenomic regulation of human T-cell leukemia virus by chromatin-insulator CTCF. PLoS Pathog. 2021;17(5):e1009577. Epub 2021/05/22. doi: 10.1371/journal.ppat.1009577. PubMed PMID: 34019588; PubMed Central PMCID: PMCPMC8174705.
25. Matsuo M, Ueno T, Monde K, Sugata K, Tan BJY, Rahman A, et al. Identification and characterization of a novel enhancer in the HTLV-1 proviral genome. Nat Commun. 2022;13(1):2405. Epub 2022/05/04. doi: 10.1038/s41467-022-30029-9. PubMed PMID: 35504920; PubMed Central PMCID: PMCPMC9065021.

26. Melamed A, Fitzgerald TW, Wang Y, Ma J, Birney E, Bangham CRM. Selective clonal persistence of human retroviruses in vivo: Radial chromatin organization, integration site, and host transcription. Sci Adv. 2022;8(17):eabm6210. Epub 2022/04/30. doi: 10.1126/sciadv.abm6210. PubMed PMID: 35486737; PubMed Central PMCID: PMCPMC9054021.
27. Gillet NA, Malani N, Melamed A, Gormley N, Carter R, Bentley D, et al. The host genomic environment of the provirus determines the abundance of HTLV-1-infected T-cell clones. Blood. 2011;117(11):3113-22. Epub 2011/01/14. doi: 10.1182/blood-2010-10-312926. PubMed PMID: 21228324; PubMed Central PMCID: PMCPMC3062313.

**Figure legends**

**Fig 1. Structure and expression of the HTLV-1 provirus in the host genome**

(A) Definitions regarding classification of the direction of transcription. Here we use 'same sense' to denote transcription from the same strand of the host genome as the proviral plus strand (e.g. *tax* gene). We refer to flanking host transcription upstream and downstream of 5′LTR-driven plus-strand proviral expression respectively as (a) same (plus) sense, 5′ side of the provirus and (b) same sense, 3′ side of the provirus. Similarly, we refer to transcription upstream and downstream of 3′LTR-driven minus-strand proviral expression respectively as (c) antisense, 3′ side of the provirus and (d) antisense, 5′ side of the provirus. In this study we focus on the effects of plus-strand expression. (B) Diagram of HTLV-1 proviral genome and splicing pattern of the regulatory genes *tax* (encoded in plus strand) and *HBZ* (encoded in minus strand); red boxes represent exons. The 4C viewpoint containing the CTCF binding site (BS) is shown.

**Fig 2. HTLV-1 plus-strand expression results in fewer chromatin loops (number of q4C peaks) between provirus and host genome**

(A) q4C profiles of $Tax^{-}$ and $Tax^{+}$ cells of two different clones (3.60 and HA1). For each clone, the top panel depicts the q4C profile in the 5′ and 3′ host genome flanking the provirus (two biological duplicates), quantified as the normalized frequency of ligation events in overlapping windows (window width 10 kb, step 1 kb). On the horizontal axis, positive values denote positions extending from the 3′ LTR side of the provirus; negative values denote positions 5′ of the 5′LTR. VP – viewpoint in q4C (proviral integration site). Diamonds mark the positions of reproducible chromatin contact sites identified by the peak calling algorithm. CTCF track – open arrowheads denote positions of CTCF-binding sites (BS); the filled arrowhead denotes the CTCF-BS in the provirus. Genes track shows RefSeq protein-

coding genes in the flanking host genome. The q4C profiles of remaining clones are shown in S2 Fig and S3 Fig.

(B) Number of identified peaks in non-expressing (Tax- or GFP-) and expressing ($Tax^+$ or $GFP^+$) subsets isolated from 6 clones: total (all peaks) and (C) peaks with or without a CTCF binding site.

**Fig 3. HTLV-1 plus-strand expression results in a significant reduction in contact frequency (q4C peak height) with the host genome (A)** Definition of q4C peak regions, as determined by the position where a peak was detected in the non-expressing cell population. **(B)** Peak height, defined as the maximum number of ligation events per region (normalised to total ligation events in a sample) calculated for each peak region. $p = 0.027$ and $p = 9.7 * 10^{-13}$ for upstream and downstream regions, respectively (unpaired two-tailed Wilcoxon test. **(C)** Peak height was significantly greater in non-expressing cells, both in both peaks with a CTCF site and those without ($p = 0.0018$ and $p = 1.8 * 10^{-13}$, respectively, unpaired two-tailed Wilcoxon test).

**Fig 4. Proviral and host transcription and splicing in live-sorted T cell clones.** (A) RNA-seq analysis of HTLV-1 proviral expression in live-sorted d2EGFP clones (TBX4B, 11.50 and 11.63). (B) Tax expression measured by qPCR with primers specific for *tax* mRNA or *18S* ribosomal RNA (18S rRNA). qPCR plots are expression values normalized to 18S rRNA. Data represent a mean of two biological replicates; error bars are SEM. AU - arbitrary units. (C) Host RNA expression 30kb on either side of the proviral integration site. On the horizontal axis, positive values denote positions extending from the 3′ LTR side of the provirus; negative values denote positions 5′ of the 5′LTR.Each row shows the transcription density (normalized RNA-seq read count) flanking that genomic position in the clone

indicated at the right-hand side. In each case, transcription orientation and positions are shown relative to the integrated provirus. Read density shown in blue shows transcription in the same orientation as the proviral plus-strand (same sense); red shows transcription in the opposite sense to the proviral plus-strand (antisense).

(D) Identification of splice sites of viral-host fusion transcripts in d2EGFP-TBX4B clone cells. Coverage tracks of same sense transcription (blue) and antisense transcription (red) in Integrative Genomics Viewer (IGV). Exons of PNPLA3 in the 3′ side of the integration site are highlighted in yellow.

(E) Fusion transcripts between an HTLV-1 plus-strand major splice donor (red, proviral exon H1 or H2) and the canonical splice acceptor site in the host *PNPLA3* gene (blue, *PNPLA3* exon 3) were identified in $GFP^+$ (HTLV-1 plus-strand-expressing) cells. To identify splice sites of fusion transcripts, reads were aligned to a reference genome (hg19) containing the HTLV-1 provirus (AB513134) genome in the TBX4B clone integration site at chr22:44323198. Fusion transcripts are shown with fused sequences.

**Fig 5 Treatment of T cell clones with a transcriptional elongation inhibitor allows recovery of chromatin loop formation in the $Tax^+$ population. (**A) After treatment with 1 nM flavopiridol (FP) for 1.5 hrs, total RNA was extracted from clone 11.63 and subjected to RT-qPCR for *tax* and three regions in the 3′ flanking host genome, respectively a: +188 bp, b: +535 bp and c: +3,198 bp from the 3′ end of the provirus. (B) Relative expression intensity (normalized to 18s rRNA) of *tax* and the host genome at positions a, b and c. Data are mean ± SEM. (N = 3). * $P<0.05$ (paired t-test). (C) q4C profiles of mock-treated Tax- cells (top track) and $Tax^+$ cells (middle track), and flavopiridol (FP)-treated $Tax^+$ cells (bottom track). Diamonds mark the positions of reproducible chromatin contact sites identified by the peak calling algorithm. Open arrowheads denote positions of CTCF-BS; the filled arrowhead

denotes the CTCF-BS in the provirus. Gene track shows RefSeq protein-coding genes in the flanking host genome.

**S1 Fig Flow sorting of Tax$^-$ and Tax$^+$ cells**

HTLV-1-infected T cell clone 11.63 was stained for live cells, crosslinked in 1% formaldehyde, stained intracellularly for Tax protein, and flow-sorted to isolate Tax$^-$ and Tax$^+$ subsets (see Methods).

**S2 Fig Flow sorting of provirus-expressing and non-expressing cells by an independent technique produced similar q4C profiles.**

(A) q4C profiles of Tax$^-$ and Tax$^+$ cells sorted from clone TBX4B after intracellular staining of Tax. (B) q4C profile of provirus-expressing (GFP$^+$) and non-expressing (GFP$^-$) cells isolated from d2EGFP-TBX4B clones, selected by GFP signal (without Tax staining).

**S3 Fig** q4C profiles of Tax$^-$ and Tax$^+$ cells from clone 11.63 (A) and of provirus-expressing (GFP$^+$) and non-expressing (GFP$^-$) cells from clone d2EGFP-11.50 (B). Vertical axis shows the normalized frequency of chromatin contacts between the provirus and the host genome.

**S4 Fig Quantitative 3C (3C-qPCR ) analysis confirmed decreased frequencies of chromatin looping.**

**(A)** q4C profile of Tax$^-$ cells of clone 11.63. The technical peak seen in the q4C viewpoint (VP) and two of the main peaks (Peak 1) and (Peak 2) identified in the output of the Tax$^-$ fraction of clone 11.63. (B) As control, the frequency of chromatin interactions was quantified by 3C-qPCR on sorted Tax$^-$ and Tax$^+$ cells, using a primer pair and Taqman probe to detect the contacts between two regions: VP and another region in the provirus (Table 3).

There was no significant difference between $Tax^+$ and $Tax^-$ cells (combined p value = 0.932, Fisher's method of combining p values). (C and D) Primer pairs and probe were used to detect long-range chromatin contacts between the provirus and host genome region at Peak 1 (C) or Peak 2 (D). Results of 3C-qPCR of two biological replicates (rep) are shown. Peak1 contact frequency was significantly higher in $Tax^-$ cells than in $Tax^+$ cells (combined p value 0.012, Fisher's method). Peak2 contact frequency was significantly higher in Tax- cells than in $Tax^+$ cells (combined p value 0.000607, Fisher's method).

**S5 Fig Fusion transcripts between HTLV-1 provirus and host genome**

(A) Identification of splice sites of fusion transcripts in the plus strand expressing cells of clone d2EGFP-11.63 and (B) in timer protein reporter clone Timer-3.60. Plus strand fusion transcripts between HTLV-1 exon1 (H1) and same sense, 3′ side host gene exon (blue) or novel host exons (green) are shown with fused sequences.

**S6 Fig Treatment with flavopiridol for 1.5 hrs did not alter Tax protein expression.**

Clone 11.63 cells were treated with 1nM flavopiridol for 1.5 hrs and then stained for Live/Dead and then Tax protein.

**S7 Fig Expression of distant host genes correlates with expression of *tax***

(A) Normalized mRNA read counts of two genes that lie >1.4 Mb from the provirus in clone Timer-TBX4B, in the four successive phases of the HTLV-1 plus-strand transcriptional burst: DN – double negative (HTLV-1 silent); blue – early burst; DP double-positive (mid-burst); red – late burst. Results of two independent experiments are shown. Expression of both *SMC1B* and *RIBC2* closely followed the trajectory of the HTLV-1 burst in clone Timer-TBX4B, but not in the unrelated HTLV-1-infected clone Timer-3.60. Data from [11]. (B)

Knockout of the CTCF binding site in the provirus in clone Timer-TBX4B (middle panel) abolished the transcription of both *SMC1B* and *RIBC2* observed in the wild-type clone (lower panel). Results of two independent experiments are shown. Neither gene was expressed in an unrelated HTLV-1-infected clone ED. The results suggest that maintenance of a CTCF-dependent chromatin loop between the host genome and the provirus is required for the burst of transcription of these distant genes associated with the HTLV-1 plus-strand burst.

**S8 Fig RNA-seq analysis of HTLV-1 proviral expression in live-sorted cells of clone Timer-TBX4B.**

Cells were sorted into four populations based on the fluorescence of the Timer protein, DN – double negative (HTLV-1 silent); blue – early burst; DP double-positive (mid-burst); red – late burst. [11].

(A) Coverage tracks in IGV of plus strand HTLV-1 provirus transcription and (B) host gene PNPLA3 transcription. (C) Transcription in *PNPLA3* exons 1 and 2 (note range on vertical axis 0 to 500).