Intact HIV-1 proviruses accumulate at distinct chromosomal positions during prolonged antiretroviral therapy

Kevin B. Einkauf,[1,2] Guinevere Q. Lee,[1,2] Ce Gao,[2] Radwa Sharaf,[1] Xiaoming Sun,[2] Stephane Hua,[2] Samantha M.Y. Chen,[2] Chenyang Jiang,[1,2] Xiaodong Lian,[1,2] Fatema Z. Chowdhury,[2] Eric S. Rosenberg,[3] Tae-Wook Chun,[4] Jonathan Z. Li,[1] Xu G. Yu,[1,2,5] and Mathias Lichterfeld[1,2,5]

[1]Infectious Disease Division, Brigham and Women's Hospital, Boston, Massachusetts, USA. [2]Ragon Institute of MGH, MIT and Harvard, Cambridge, Massachusetts, USA. [3]Infectious Disease Division, Massachusetts General Hospital, Boston, Massachusetts, USA. [4]National Institute of Allergy and Infectious Diseases (NIAID), Bethesda, Maryland, USA. [5]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.

Chromosomal integration of genome-intact HIV-1 sequences into the host genome creates a reservoir of virally infected cells that persists throughout life, necessitating indefinite antiretroviral suppression therapy. During effective antiviral treatment, the majority of these proviruses remain transcriptionally silent, but mechanisms responsible for viral latency are insufficiently clear. Here, we used matched integration site and proviral sequencing (MIP-Seq), an experimental approach involving multiple displacement amplification of individual proviral species, followed by near-full-length HIV-1 next-generation sequencing and corresponding chromosomal integration site analysis to selectively map the chromosomal positions of intact and defective proviruses in 3 HIV-1–infected individuals undergoing long-term antiretroviral therapy. Simultaneously, chromatin accessibility and gene expression in autologous CD4$^+$ T cells were analyzed by assays for transposase-accessible chromatin using sequencing (ATAC-Seq) and RNA-Seq. We observed that in comparison to proviruses with defective sequences, intact HIV-1 proviruses were enriched for non-genic chromosomal positions and more frequently showed an opposite orientation relative to host genes. In addition, intact HIV-1 proviruses were preferentially integrated in either relative proximity to or increased distance from active transcriptional start sites and to accessible chromatin regions. These studies strongly suggest selection of intact proviruses with features of deeper viral latency during prolonged antiretroviral therapy, and may be informative for targeting the genome-intact viral reservoir.

Introduction

HIV-1 integration into host chromosomes, catalyzed by the viral enzyme integrase, allows the virus to persist in the human organism for the life span of the infected cell and represents the major obstacle against interventions to eradicate or cure HIV-1 infection (1–4). The vast majority of chromosomally integrated HIV-1 sequences in infected patients harbor lethal defects that are introduced during viral reverse transcription and preclude viral replication (5); these defective proviral sequences represent fossils of the replicative history of HIV-1 in a particular patient, rather than a functionally relevant viral reservoir. In contrast, genome-intact proviruses typically have an uncompromised capacity for viral gene transcription, translation, and virion production, but remain transcriptionally inactive or silent during suppressive antiretroviral therapy. This transcriptional latency protects against viral immune recognition by innate and adaptive effector cells, reduces cytopathic effects associated with HIV-1 infection in host cells, and likely represents the key factor enabling long-term persistence of replication-competent viral sequences despite highly effective antiretroviral therapy (6, 7). The mechanisms that maintain viral latency in vivo during treatment with antiretroviral therapy appear to be highly complex, may differ among individual patients, and have escaped clarification so far. Given that disruption of viral latency is frequently considered a prerequisite for targeting persisting HIV-1 reservoirs (8), a better understanding of viral latency in vivo would represent an important advance for developing curative strategies for HIV-1 infection. Exploring the mechanisms that maintain viral latency in HIV-1–infected patients will likely be facilitated by an interrogation of chromatin structure and host transcriptional activity at the chromosomal integration sites of individual intact HIV-1 proviruses, similar to the way identification of viral integration sites of human T lymphotropic virus–1 (HTLV-1) has recently allowed better understanding of HTLV-1 disease pathogenesis (9, 10). However, a simultaneous assessment of genome-intact proviral sequences and their corresponding chromosomal integration sites has so far only been possible in the rare instance of a cancer patient with a highly expanded HIV-1–infected CD4$^+$ T cell clone that continuously produced replication-competent HIV-1 detectable as ongoing plasma vire-

Authorship note: KBE, GQL, CG, and RS are co–first authors.
Conflict of interest: The authors have declared that no conflict of interest exists.

Submitted: August 16, 2018; **Accepted:** December 4, 2018.
Reference information: *J Clin Invest*. 2019;129(3):988–998.
https://doi.org/10.1172/JCI124291.

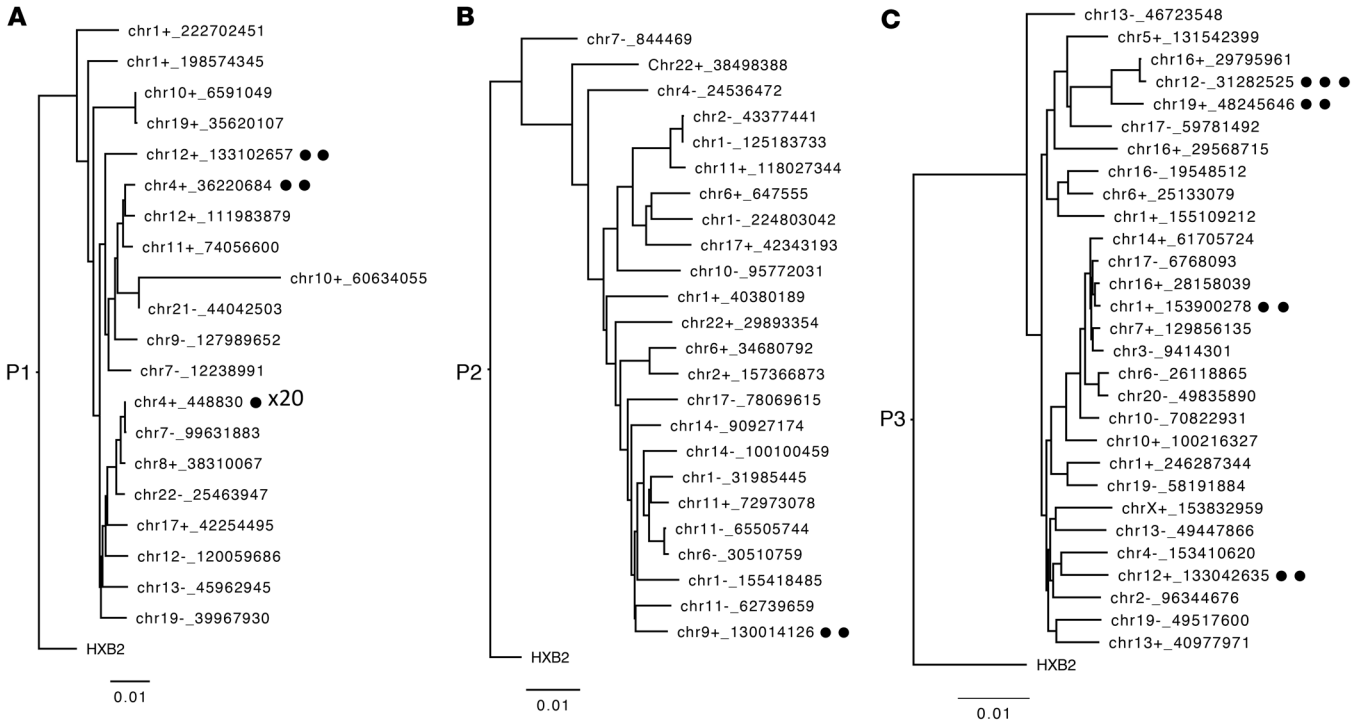

Figure 1. Simultaneous analysis of near-full-length HIV-1 proviral sequences and corresponding HIV-1 integration sites. (**A**–**C**) Horizontal phylogenetic trees of all intact, near-full-length HIV-1 sequences from 3 study participants (P1–P3). Clonal sequences are listed only once; the number of clones is indicated by circular symbols. Chromosomal integration site coordinates (3′-LTR border) for each sequence are indicated.

mia, despite treatment with suppressive antiretroviral therapy (11). To investigate mechanisms that contribute to viral latency in vivo, we here developed a technical approach, termed matched integration site and proviral sequencing (MIP-Seq), that allowed us to individually profile pairs of proviral species and their corresponding chromosomal integration sites in patient-derived material from a global and unbiased perspective. Using this method, we conducted a detailed and systematic analysis of the chromosomal positions of a large number of intact and defective proviral sequences in 3 individuals undergoing long-term antiretroviral therapy. Our results suggest that prolonged antiretroviral therapy is associated with an accumulation of intact proviruses with features of deeper latency, likely as a result of immune-mediated selection pressure.

Results

Simultaneous analysis of HIV-1 proviral sequences and integration sites. To investigate mechanisms of viral latency in HIV-1 patients treated with suppressive antiretroviral therapy, an analysis of proviral sequences in conjunction with corresponding chromosomal integration sites for each provirus would be highly informative; however, such an analysis has been precluded in the past by technical approaches that permit only isolated assessments of either proviral sequences or viral integration sites (12–14). To address this, we here developed an experimental approach to concurrently analyze pairs of proviral HIV-1 sequences and their respective chromosomal integration sites using a combined assay system, termed MIP-Seq. First, genomic DNA was isolated from $CD4^+$ T cells of 3 HIV-1–infected patients treated with suppressive antiretroviral therapy for approximately 10 years (Supplemental Table 1; supplemental material available online with this article; https://doi.org/10.1172/JCI124291DS1), subjected to quantification of viral *gag* copies by droplet digital PCR (ddPCR), and diluted to single proviral genomes based on ddPCR results and Poisson distribution statistics. Afterward, cells were exposed to multiple displacement amplification (MDA) mediated by phi29 polymerase; this whole-genome amplification (WGA) process generates 1000–10,000 identical copies of an individual cell's genome, including any proviral sequence possibly harbored by a given cell. Subsequently, material from each individual MDA reaction was split and separately subjected to viral sequence amplification with primers spanning near-full-length HIV-1 (15, 16) and to chromosomal integration site analysis based on integration site loop amplification (ISLA) (13), ligation-mediated PCR (LM-PCR) (17), or nonrestrictive linear amplification–mediated PCR (nrLAM-PCR) (18); frequently, a combination of these integration site assays was used, yielding identical results. Amplified near-full-length viral sequences and viral-host junctions were analyzed by Illumina MiSeq next-generation sequencing. Although intact proviruses constitute only a small minority of total HIV-1 DNA sequences, we sought to analyze roughly equal numbers of intact and defective sequences by prioritizing the investigation of proviral sequences that approximated the size of full HIV-1 genomes (>8 kb) based on gel electrophoresis analysis. Using this approach, we identified 100 intact proviral sequences and their corresponding integration sites from the 3 study patients; of these 100 intact sequences, we detected $n = 73$ distinct pairs of proviral sequences and integration sites. A total of 84 defective proviral sequences (with hypermutations, major deletions, or internal

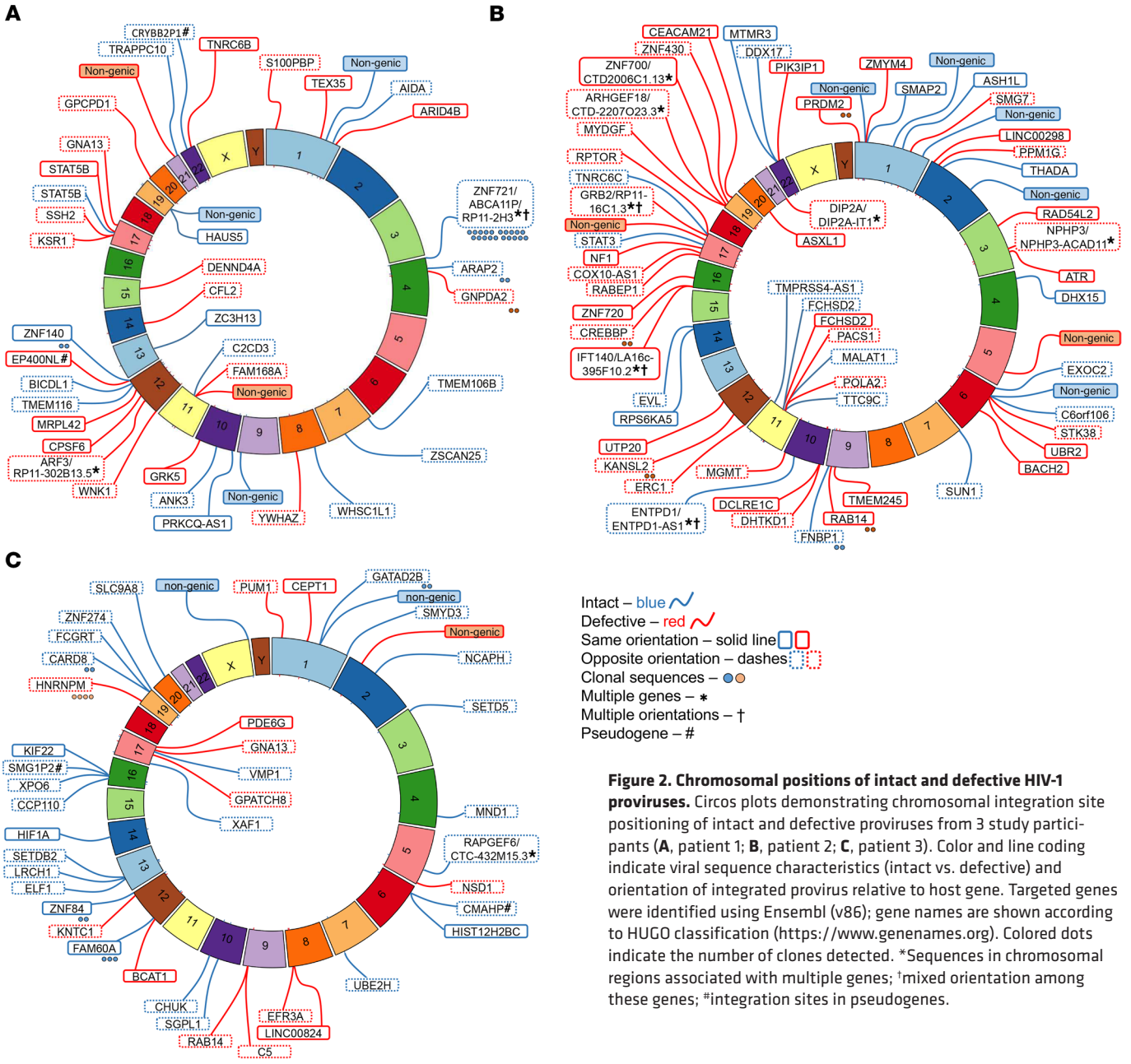

Figure 2. Chromosomal positions of intact and defective HIV-1 proviruses. Circos plots demonstrating chromosomal integration site positioning of intact and defective proviruses from 3 study participants (**A**, patient 1; **B**, patient 2; **C**, patient 3). Color and line coding indicate viral sequence characteristics (intact vs. defective) and orientation of integrated provirus relative to host gene. Targeted genes were identified using Ensembl (v86); gene names are shown according to HUGO classification (https://www.genenames.org). Colored dots indicate the number of clones detected. *Sequences in chromosomal regions associated with multiple genes; †mixed orientation among these genes; #integration sites in pseudogenes.

inversions) and their respective integration sites were also identified, of which $n = 76$ represented distinct combinations of proviral sequences and corresponding integration sites (Figure 1, Figure 2, Supplemental Tables 2 and 3, and Supplemental Figures 1 and 2). Notably, intact proviruses generated after MDA were phylogenetically intermingled with sequences identified without prior WGA, demonstrating that cell-free cloning of proviral sequences by MDA is not associated with a selection bias for individual proviruses (Supplemental Figure 1). Moreover, we observed intact proviral sequences after MDA that were highly similar or identical to near-full-length proviral sequences retrieved from viral outgrowth assays, indicating that genome-intact sequences can indeed be fully replication- and infection-competent (Supplemental Figure 1), as shown in our prior work (15). Within all amplified sequences, we detected 8 clusters of intact sequences, each consisting of multiple identical proviruses paired with identical chromosomal integration sites; one large cluster encompassed 20 individual identical intact sequences in study participant 1, all located at the same position in the zinc finger protein 721–encoding gene (*ZNF721*) on chromosome 4 (Figures 1 and 2). Together, these clusters of identical intact proviral sequences accounted for $n = 35$ (35%) of all $n = 100$ intact proviral sequences analyzed. The identification of such identical proviral sequences matched with identical viral integration sites strongly supports the role of clonal proliferation for maintaining and stabilizing a pool of viral reservoir cells encoding for intact HIV-1 (19–21). In addition to intact proviral sequences derived from such clonally expanded $CD4^+$ T cells, we also noted 6 clusters of defective proviruses exhibiting

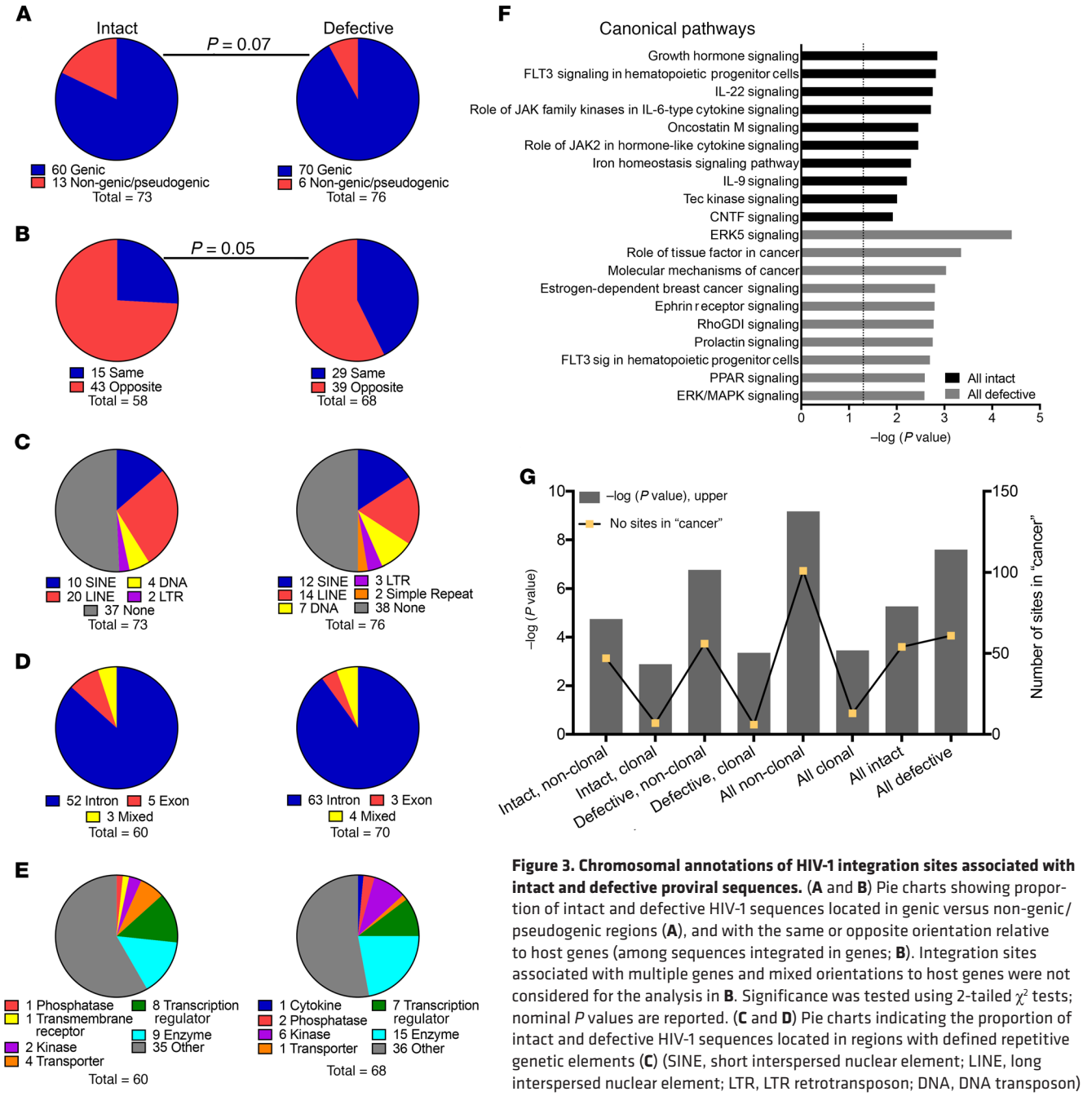

Figure 3. Chromosomal annotations of HIV-1 integration sites associated with intact and defective proviral sequences. (**A** and **B**) Pie charts showing proportion of intact and defective HIV-1 sequences located in genic versus non-genic/pseudogenic regions (**A**), and with the same or opposite orientation relative to host genes (among sequences integrated in genes; **B**). Integration sites associated with multiple genes and mixed orientations to host genes were not considered for the analysis in **B**. Significance was tested using 2-tailed χ^2 tests; nominal *P* values are reported. (**C** and **D**) Pie charts indicating the proportion of intact and defective HIV-1 sequences located in regions with defined repetitive genetic elements (**C**) (SINE, short interspersed nuclear element; LINE, long interspersed nuclear element; LTR, LTR retrotransposon; DNA, DNA transposon) and in exons or introns (**D**). (**E**–**G**) Ontology analysis of genes harboring defective and intact HIV-1 sequences. Data represent a categorization of genes harboring intact or defective HIV-1 sequences according to defined formal functional entities (**E**). (**F**) Top ten canonical pathways predicted by Ingenuity Pathway Analysis for genes containing intact or defective proviruses; *x* axis shows corresponding –log(*P* value) for each pathway using right-tailed Fisher's exact tests, with a threshold of –log(0.05) marked as a dotted line. RhoGDI, Rho GDP dissociation inhibitor. (**G**) Positioning of intact and defective HIV-1 proviruses in cancer-related genes. Left *y* axis shows upper limit of the –log(*P* value) for each indicated category (Ingenuity Pathway Analysis–based right-tailed Fisher's exact tests); right *y* axis depicts the number of sites identified in the "Cancer" category in the different gene groups. For **A**–**G**, clonal sequences were counted only once.

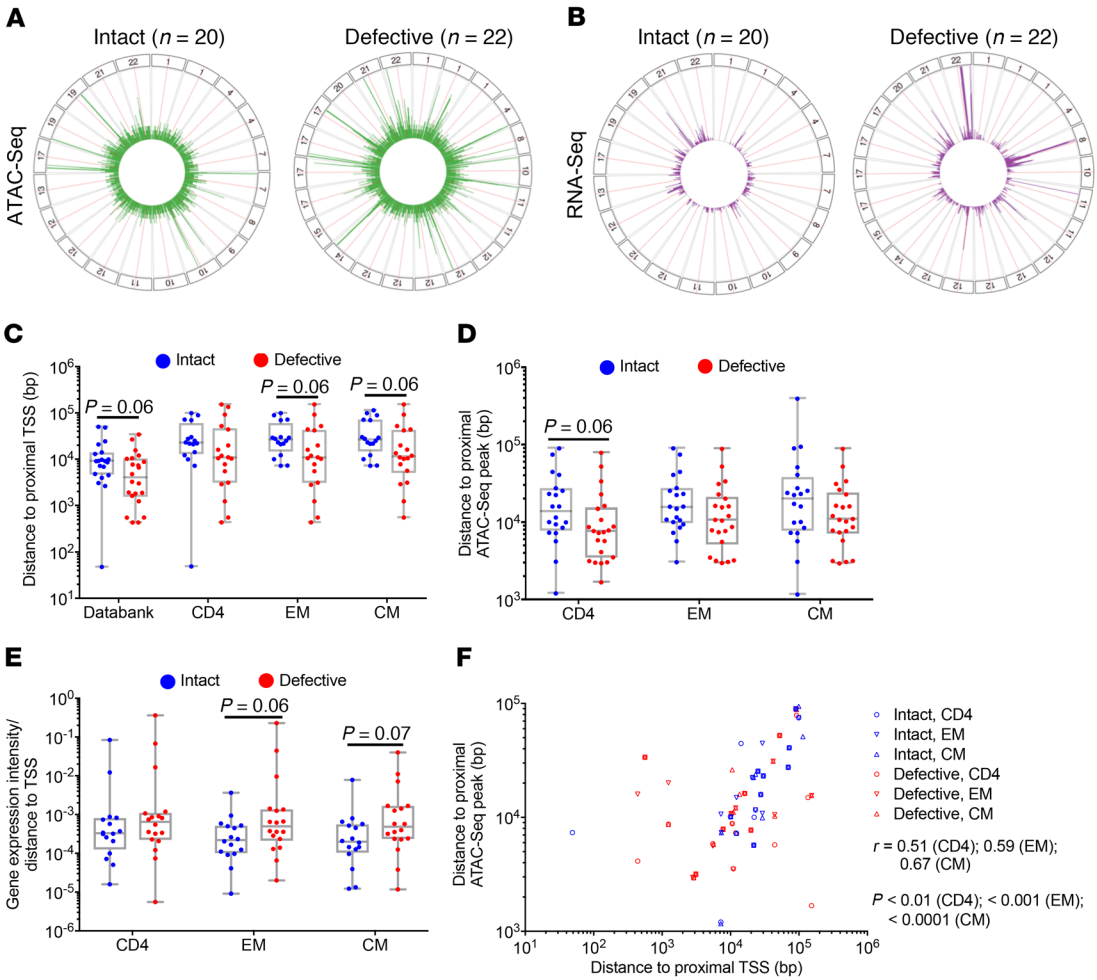

Figure 4. Distinct chromosomal locations of intact HIV-1 proviruses in study participant 1. (**A** and **B**) Circos plots highlighting ATAC-Seq and RNA-Seq reads in proximity (ATAC-Seq: ±8000 bp; RNA-Seq: ±5000 bp) to integration sites of intact and defective HIV-1 proviruses. (**C**) Combined individual-value/box-and-whisker plots indicating the chromosomal distance between HIV-1 integration sites and the most proximal TSS listed in Ensembl v86 (databank), or identified through analysis of expressed RNA species located within the boundaries of the host gene, using autologous RNA-Seq data from the indicated cell populations and limiting the analysis to proviruses integrated in expressed genes. (**D**) Combined individual-value/box-and-whisker plots showing the chromosomal distance between integration sites and the center of the most proximal ATAC-Seq peaks in indicated CD4$^+$ T cell populations. (**E**) Gene expression intensity of host genes harboring intact or defective proviral integration sites, normalized to the chromosomal distance between integration sites and the most proximal TSSs determined using autologous RNA-Seq data as described in **C**. In **C**–**E**, boxes and whiskers represent median, 25% and 75% percentiles, and minimum/maximum levels. Significance was calculated using 2-tailed Mann Whitney *U* tests; nominal *P* values are reported. (**F**) Distance between each integration site and most proximal TSS, plotted against corresponding distance between each integration site and the center of the nearest ATAC-Seq peak. Spearman's correlation coefficients are shown for each cell population. In **A**–**F**, clonal sequences were shown/counted only once. EM, effector memory; CM, central memory.

identical viral sequences with identical viral integration sites in each cluster; these clusters involved $n = 14$ (16.7%) sequences of the entire pool of $n = 84$ defective sequences analyzed. Although the amplification of identical viral sequences, coupled with identical corresponding integration sites from distinct single proviruses, supported the technical consistency of our experimental approach, we conducted additional experiments to further validate our method: for those intact proviral sequences from which sufficient material was available, we analyzed the viral-host junction sequence at both the 5′ long terminal repeat (5′-LTR) and the 3′-LTR border regions, which verified the identity of the respective chromosomal integration site (Supplemental Table 3 and Supplemental Figure 3). Moreover, our experimental approach allowed us to investigate viral sequence variations in the viral 5′-LTR and/or 3′-LTR promoter regions, which are not covered by the near-full-length sequencing assays used previously for identification of genome-intact proviruses (5, 15, 22). These additional studies demonstrated that relative to the functionally intact promoter regions in HXB2, patient-derived HIV-1 promoters were highly conserved and diversity was mostly attributable to single base substitution mutations (Supplemental Figure 4).

Chromosomal integration site features of intact proviruses. We subsequently focused on identifying distinguishing features of intact proviral sequences and their chromosomal locations. In order to avoid bias due to clonal expansion, integration sites for each cluster of clonally identical sequences were counted only

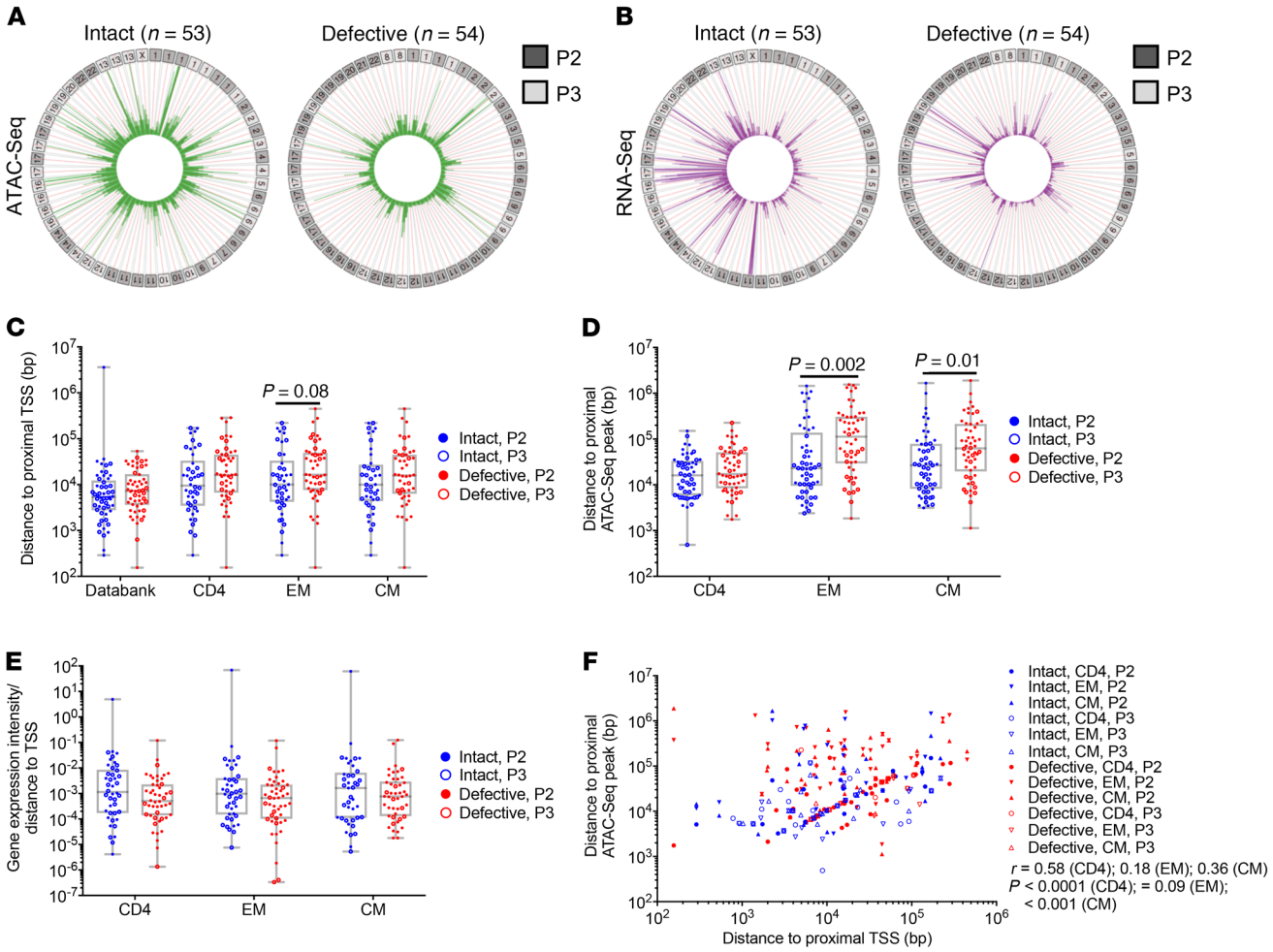

Figure 5. Chromosomal integration site features in study participants 2 and 3. (**A** and **B**) Circos plots highlighting ATAC-Seq and RNA-Seq reads in proximity (ATAC-Seq: ±20,000 base pairs; RNA-Seq: ±10,000 base pairs) to integration sites of intact and defective HIV-1 proviruses. (**C**) Combined individual-value/box-and-whisker plots summarizing the distance between integration sites and the most proximal TSS derived from Ensembl v86 (databank) or from autologous RNA-Seq data (as described for Figure 4C, among proviruses integrated in expressed genes). (**D**) Combined individual-value/box-and-whisker plots showing the distance between integration sites and the center of the most proximal ATAC-Seq peak. (**E**) Combined individual-value/box-and-whisker plots demonstrating the transcriptional activity of host genes harboring intact or defective proviral sequences, normalized to the distance between integration sites and the most proximal TSSs, as determined for Figure 4C and Figure 5C. In **C**–**E**, boxes and whiskers represent median, 25% and 75% percentiles, and minimum/maximum levels. Significance was calculated using 2-tailed Mann Whitney *U* tests; nominal *P* values are reported. (**F**) Distance between each integration site and most proximal TSS, plotted against corresponding distance between each integration site and the center of the nearest ATAC-Seq peak. Spearman's correlation coefficients are shown for each cell population. For **A**–**F**, clonal sequences were shown/counted only once. EM, effector memory; CM, central memory.

once for these studies. We observed that relative to defective proviruses, a larger fraction of intact proviruses was located in non-genic or pseudogenic genomic regions (Figure 3A), which were previously associated with a deeper level of viral latency in tissue culture models of HIV-1 infection (23–26). In addition, among proviruses integrated in genes, we found a higher proportion of intact proviruses integrated in opposite orientation to the host gene, which can support viral latency by increasing susceptibility to transcriptional interference (27) (Figure 3B). This trend for an enrichment of intact proviral sequences in non-genic/pseudogenic regions and in opposite orientation to host genes was also observable when the 3 study participants were analyzed individually (Supplemental Figure 5) and when clonal sequences were counted as individual sequences (Supplemental Figure 6). There were no differences between intact and defective proviruses with regard to their location in introns, exons, or repetitive genetic elements (Figure 3, C and D). Notably, we identified viral integration sites in a number of host genes that have frequently been associated with viral integration, including *BACH2* and *STAT5B* (13, 14, 28). In each of these 2 genes, we noted a defective proviral sequence integrated in the same direction as the host gene, in addition to an intact sequence in *STAT5B* integrated in the opposite orientation. Moreover, computational gene ontology analysis indicated that genes harboring proviral sequences were frequently encoding for T cell transcription factors or otherwise involved in the regulation of T cell behavior; however, there were no distinct differences in the predicted functional profile between genes hosting intact versus defective proviruses (Figure 3, E–G). Chromosomal locations

of proviral sequences derived from clonally expanded CD4+ T cells were genic for all analyzed intact and defective proviruses, consistent with prior results (13, 14), and slightly more frequently positioned in opposite orientation to host genes (intact proviruses: 4 opposite orientation, 3 same orientation, 1 mixed orientation; defective proviruses: 4 opposite orientation, 2 same orientation), but there was no evidence that genes harboring intact or defective sequences isolated from clonally expanded CD4+ T cells were enriched for cancer-associated functions (Figure 3G). These data suggest that during prolonged antiretroviral therapy, intact viral sequences located in non-genic regions and in opposite orientation to host genes are preferentially selected for, likely as a result of immune-mediated mechanisms.

Chromatin accessibility and transcriptional activity at integration sites of intact proviruses. For a closer analysis of chromosomal integration site features, we used RNA-Seq for genome-wide transcriptional profiling of autologous CD4+ T cells and of sorted central memory and effector memory CD4+ T cells, the two CD4+ T cell subpopulations most frequently harboring HIV-1 proviral sequences (29, 30). These studies allowed us to calculate the chromosomal distance between each proviral integration site and the most proximal transcriptional start site (TSS), and to determine the transcriptional activity of the respective host genes containing proviral sequences. Simultaneously, the chromatin accessibility of genomic DNA regions from our study participants was assessed in total, central memory, and effector memory CD4+ T cells, applying assays for transposase-accessible chromatin using sequencing (ATAC-Seq); these data were used to determine the chromosomal distance between each proviral integration site and the center of the most proximal ATAC-Seq peak. In study participant 1, these experiments showed that relative to defective proviruses, intact proviral sequences showed an increased distance to the nearest active TSSs, coupled with an increased distance to the most proximal accessible chromatin regions; this was true when clonal sequences were counted only once (Figure 4, A–E) or when included as individual sequences in this analysis (Supplemental Figure 7, A–C). The distances to the TSSs and the ATAC-Seq peaks were closely correlated with each other (Figure 4F), but there was no marked difference between the intact proviral sequences integrated in opposite orientation to the host genes and the few intact sequences integrated in the same direction. Consistent with prior data (1, 31), integration of intact and defective proviruses was biased toward highly expressed host genes (Supplemental Figure 8), with a small trend toward lower gene expression intensity in genes harboring intact proviruses compared with those containing defective viral sequences (Figure 4B). Moreover, a composite analysis of the transcriptional activity of host genes harboring integration sites, normalized to the distance between integration sites and the most proximal TSSs, was compatible with preferential persistence of intact proviruses in regions with more limited transcriptional activity (Figure 4E and Supplemental Figure 7C). Together, data from this patient suggest selection of intact proviruses located in less-accessible chromatin and with increased distance to active TSSs; these features were associated with deeper levels of HIV-1 latency in previous in vitro studies (23, 25, 32).

In study patients 2 and 3, our results demonstrated an opposite pattern for chromosomal integration site features of intact proviruses: in both of these participants, we noted that relative to defective proviral sequences, intact HIV-1 proviruses appeared to be located in closer proximity to active TSSs and to accessible chromatin (Figure 5, A–E); this trend was also observed when clonal sequences were considered as individual sequences (Supplemental Figure 7, D–F). Distances to TSSs or ATAC-Seq peaks, again closely correlated with each other (Figure 5F), did not differ notably between intact proviruses integrated in the same orientation and those integrated in the opposite orientation to their respective host genes, but our statistical power to detect such differences was low given the comparatively small number of intact proviruses with the same directional configuration as the host gene. There was no significant difference in expression intensity between genes harboring intact versus defective proviruses in these two patients (Supplemental Figure 8); however, the transcriptional activity of host genes, normalized to the chromosomal distance between integration sites and the most proximal TSSs, suggested preferential enrichment of intact proviruses in closer proximity to host transcriptional activity after long-term antiretroviral therapy (Figure 5E and Supplemental Figure 7F). This pattern supports the presence of transcriptional interference between host and proviral gene expression, which has been previously described in in vitro models of HIV-1 latency (26, 31, 33), as a predominant mechanism for maintaining HIV-1 latency in participants 2 and 3. Transcriptional interference can effectively inhibit gene expression of proviruses (31, 33) and may explain the otherwise paradoxical finding that HIV-1 can remain transcriptionally silent despite integration in actively transcribed and typically highly expressed host genes. Our data suggest that a greater susceptibility to transcriptional interference, due to closer proximity to active transcriptional units and accessible chromatin sites of the host, provided a selection advantage for intact proviruses during long-term antiretroviral therapy in participants 2 and 3.

Discussion

HIV-1 persistence during suppressive antiretroviral therapy is primarily related to the ability of chromosomally integrated proviruses to maintain a state of transcriptional silence, during which expression of viral gene products is inhibited. However, it is clear that for the vast majority of intact proviruses, viral latency is not permanent and that reactivation of previously latent, genome-intact proviruses can occur spontaneously or as a result of pharmacological interventions (34). Increasingly, it is recognized that this viral reactivation is the result of a complex interplay of viral and host factors that support active viral gene transcription, and that the susceptibility to reactivating stimuli can vary profoundly among different proviral species (32). Therefore, it is reasonable to assume that proviral latency entails a spectrum of degrees of transcriptional silence, ranging from very deep latent and difficult-to-reactivate proviruses to viral sequences with high responsiveness to reactivation signals (35). The positions of proviruses within host chromosomes, and the relative distance to accessible chromatin and active transcriptional units of the host, are likely to have a critical influence on the degree of viral latency and may strongly impact the ability to maintain deep viral latency for extended periods of time.

The present work, which to our knowledge is the first comprehensive study analyzing combinations of proviral species and

their corresponding integration sites, demonstrates that prolonged antiretroviral therapy may be associated with enrichment for proviral sequences exhibiting multiple discrete features suggestive of a deeper level of latency. Notably, these features involved a number of complementary characteristics, including proviral positioning in intergenic regions, in opposite orientation to host genes, and in either relative proximity to or increased distance from host TSSs and accessible chromatin. This seems to highlight the flexibility of HIV-1 in maintaining transcriptional silence through a variety of different mechanisms that may be individually adjusted to the host's immune environment. We propose that these findings may represent the result of active selection processes that favor persistence of proviruses with a higher resistance to viral reactivation, which protects against viral immune recognition and immune targeting by innate and adaptive immune responses, and, through reduction of cytopathic effects, may enable prolonged survival of host cells harboring genome-intact HIV-1 (36). Therefore, our data suggest that suppressive antiretroviral therapy for extended periods of time may be associated with a profound change in the structure and composition of the viral reservoir, potentially leading to an accumulation of intact proviral sequences with progressively increasing depths of viral latency, while intact proviruses more likely to respond to reactivation stimuli appear to be actively selected against. These data are reminiscent of endogenous human retroviruses, which during human evolution seem to have been progressively selected for genomic locations associated with increased latency, such as non-genic locations, locations with more pronounced distance to host TSSs, and positions with opposite orientation to host genes (37). Future longitudinal evaluations of changes in chromosomal positions of intact HIV-1 proviruses may delineate selection forces driving the composition and structure of the proviral reservoir landscape in greater detail, specifically when paired with a direct quantification of the transcriptional activity of each proviral species. In addition, such a future analysis could also clarify whether selection and persistence of virally infected $CD4^+$ T cell clones may be influenced by alterations in host gene function that result from retroviral integration events. In this context, it is remarkable that disruption of the methylcytosine dioxygenase *TET2* gene following chimeric antigen receptor–encoding lentiviral integration enabled massive expansion of the respective T cell clone (38); these and other results (28) suggest that host gene editing resulting from retroviral integration may indeed influence clonal survival and persistence. Notably, preferential clonal proliferation of cells harboring defective proviruses, possibly as a result of integration in genes actively promoting cell turnover, may also contribute to an imbalance between the proportion of intact and defective proviruses in genic versus non-genic regions, and deserves future detailed investigation.

Together, our results suggest that the genome-intact HIV-1 reservoir is dynamically evolving over time, susceptible to (likely immune-mediated) selection pressure, and, potentially, vulnerable to therapeutic interventions that may support and/or accelerate the natural selection of intact proviruses with deeper levels of latency. Eradication strategies that facilitate the development of a proviral reservoir with deeper latency and a more limited ability to fuel rebound viremia may represent one perspective for future efforts to delay viral recurrence after treatment interruptions, and to induce at least transient drug-free remission of HIV-1 infection. In addition, the simultaneous analysis of proviral sequences and corresponding chromosomal integration sites provided by the MIP-Seq assay may allow for a more detailed and precise assessment of qualitative and quantitative viral reservoir changes during interventional clinical trials aimed at destabilizing and reducing the persistence of latently HIV-1–infected cells.

Methods

Patients. HIV-1–infected study participants were recruited at Massachusetts General Hospital, Brigham and Women's Hospital, and the NIH Clinical Center. PBMC samples were used according to protocols approved by the respective Institutional Review Boards. Clinical and demographical characteristics of study participants are summarized in Supplemental Table 1.

ddPCR. $CD4^+$ T cells were enriched from total PBMCs using a $CD4^+$ T Cell Isolation Kit (Miltenyi Biotec, catalog 130-096-533) and subjected to DNA extraction using commercial kits (QIAGEN DNeasy, 69504). We amplified total HIV-1 DNA using ddPCR (Bio-Rad), with primers and probes described previously (127-bp 5′-LTR-*gag* amplicon; HXB2 coordinates 684–810). PCR was performed using the following program: 95°C for 10 minutes; 45 cycles (95°C for 30 seconds, 60°C for 1 minute); 98°C for 10 minutes. The droplets were subsequently read by a QX100 droplet reader, and data were analyzed using QuantaSoft software (Bio-Rad).

WGA. Extracted DNA was diluted to single viral genome levels according to ddPCR results, so that 1 provirus was present in approximately 20%–30% of wells. Subsequently, DNA in each well was subjected to MDA with phi29 polymerase (QIAGEN REPLI-g Single Cell Kit, catalog 150345), per the manufacturer's protocol. Following this unbiased WGA, DNA from each well was split and separately subjected to viral sequencing and integration site analysis, as described below. If necessary, a second-round MDA reaction was performed to increase the amount of available DNA.

HIV near-full-genome sequencing. DNA resulting from full-genome amplification reactions was subjected to HIV-1 near-full-genome amplification using a 1-amplicon (15) and/or non-multiplexed 5-amplicon approach (full list of primer sequences available in ref. 16, with modifications at the forward nested *gag* primers U5-623F [5′-AAATCTCTAGCAGTGGCGCCCGAACAG-3′] and U5-638F [5′-GCGCCCGAACAGGGACYTGAAARCGAAAG-3′]). One unit per 20-μl reaction of Invitrogen Platinum Taq (catalog 11304-029) was incubated with 1× reaction buffer, 2 mM $MgSO_4$, 0.2 mM dNTP, and 0.4 μM each of forward and reverse primers for 2 minutes at 94°C; 10 cycles (15 seconds at 94°C, 30 seconds at 55°C, 2 minutes 30 seconds at 72°C); 25 cycles (15 seconds at 94°C, 30 seconds at 55°C, 2 minutes 30 seconds at 72°C, adding 5 seconds/cycle); 72°C for 7 minutes and 4°C infinite hold. PCR products were visualized by agarose gel electrophoresis. All near-full-length and/or 5-amplicon positive amplicons, and selected sequences with major deletions (<8000 bp in gel size or <5 amplicons positive) were subjected to Illumina MiSeq sequencing at the MGH DNA Core facility. Resulting short reads were de novo assembled using Ultracycler v1.0 and aligned to HXB2 to identify large deleterious deletions (<8000 bp of the amplicon aligned to HXB2), out-of-frame indels, premature/lethal stop codons, internal inversions, or packaging signal deletions (≥15 bp insertions and/

or deletions relative to HXB2), using an automated in-house pipeline written in R scripting language (39). Presence/absence of APOBEC-3G/3F–associated hypermutations was determined using the Los Alamos HIV Sequence Database Hypermut 2.0 (40) program. Viral sequences that lacked all mutations listed above were classified as "genome-intact." If a near-full-length sequence showed a mapped 5′ deletion that removes the primer-binding site, but otherwise showed no lethal sequence defects, the missing 5′ sequence was inferred to be present, and this sequence was considered as an "inferred intact" HIV-1 sequence. Multiple sequence alignments were performed using MUSCLE (41). Phylogenetic distances between sequences were examined using a Clustal X–generated neighbor joining algorithm (42). Viral sequences were considered clonal if they had completely identical consensus sequences; single nucleotide variations in primer binding sites were not considered for clonality analysis.

Integration site analysis. Integration sites associated with each viral sequence were obtained using LM-PCR (Lenti-X Integration Site Analysis Kit; Clontech, catalog 631263), and/or nrLAM-PCR (18) and/or ISLA (13). No modifications were made to these protocols except in Lenti-X: along with the 5′-LTR–associated integration site amplification, we added a 3′-LTR–associated integration site amplification reaction using nested LTR1 (5′-CTTAAGCCTCAATAAAGCTTGCCTTGAG-3′, HXB2 9602 forward) and LTR2 (5′-AGACCCTTTTAGTCAGTGTGGAAAATC-3′, HXB2 9686 forward) primers. Resulting PCR products were subjected to next-generation sequencing using Illumina MiSeq. Control cells (8E5/LAV cell line; NIH AIDS Reagent Program, catalog 95) were included in most experimental conditions and consistently retrieved integration sites chr 13-67485907 (3′-LTR) and -67485903 (5′-LTR). MiSeq paired-end FASTQ files were demultiplexed and analyzed using 2 independent bioinformatics approaches. In the first method, small reads (142 bp) were aligned simultaneously to human reference genome GRCh38 and HIV-1 reference genome HXB2 using BWA-MEM (43). Chimeric reads containing both human and HIV-1 sequences were evaluated for mapping quality based on (i) absolute counts of chimeric reads, (ii) percentages of chimeric reads per total, (iii) depth of sequencing coverage in the host genome adjacent to the viral integration site, and (iv) HIV-1 coordinates mapping to the terminal nucleotides of the viral genome. In the second method, small reads were de novo assembled using Ultracycler v1.0, generating one consensus sequence from each contig, which were then mapped to HXB2 using R library Biostrings local pairwise alignment to identify terminal nucleotides of the viral genome. Then, the HIV-1 part of the contig consensus was trimmed, and the remaining portion (>20 bp) was submitted to Web-based BLAT (44). The final list of integration sites and corresponding chromosomal annotations was obtained using Ensembl (v86, http://www.ensembl.org). Pseudogenes were determined using UCSC Genome Browser (http://www.genome.ucsc.edu/) and GENCODE (v28, https://www.gencodegenes.org/). Repetitive genomic sequences harboring HIV-1 integration sites were identified using RepeatMasker (http://www.repeatmasker.org/).

Cell sorting and flow cytometry. PBMCs were stained with monoclonal antibodies to CD4 (clone RPA-T4, BioLegend, catalog 300518), CD3 (clone OKT3, BioLegend, 317332), CD45RO (clone UCHL1, BioLegend, 304236), and CCR7 (clone G043H7, BioLegend, 353216). Afterward, cells were washed, and $CD45RO^+CCR7^+$ (central memory), $CD45RO^+CCR7^-$ (effector memory), and $CD3^+CD4^+$ (total) $CD4^+$ T cells were sorted in a specifically designated biosafety cabinet (Baker Hood), using a FACSAria cell sorter (BD Biosciences) at 70 pounds per square inch. Cell sorting was performed by the Ragon Institute Imaging Core facility at Massachusetts General Hospital and resulted in isolation of lymphocytes with the defined phenotypic characteristics of >95% purity. Data were analyzed using FlowJo software (Tree Star).

RNA-Seq. Total RNA was extracted from sorted $CD4^+$ T cell populations using a PicoPure RNA Isolation Kit (Applied Biosystems, catalog KIT0204). RNA-Seq libraries were generated as previously described (45). Briefly, whole transcriptome amplification (WTA) and tagmentation-based library preparation were performed using SMART-seq2, followed by sequencing on a NextSeq 500 Instrument (Illumina). Quantification of transcript levels was conducted using RSEM software (v1.2.22) (46) supported by STAR aligner software (STAR 2.5.1b) and aligned to the Hg38 human genome. Transcripts per million (TPM) values were then normalized among all samples using the upper quantile normalization method.

ATAC-Seq. A previously described protocol with some modifications (47, 48) was used. Briefly, 20,000 sorted cells were centrifuged at 239 x *g* for 10 minutes at 4°C in a pre-cooled fixed-angle centrifuge. All supernatant was removed, and 50 μl transposase mixture (25 μl 2× TD buffer, 1.5 μl TDE1, 0.5 μl 1% digitonin, 16.5 μl PBS, 6.5 μl nuclease-free water) was added to the cells and incubated in a heat block at 37°C for 30 minutes. Transposed DNA was purified using a ChIP DNA Clean & Concentrator Kit (Zymo Research, catalog D5205), and eluted DNA fragments were used to amplify libraries. The libraries were quantified using an Agilent Bioanalyzer 2100 and the Q-Qubit dsDNA High Sensitivity Assay Kit. All Fast-ATAC libraries were sequenced using NextSeq with paired-end reads. The quality of reads was assessed using FastQC (https://www.bioinformatics.babraham.ac.uk). Low-quality DNA end fragments and sequencing adapters were trimmed using Trimmomatic (http://www.usadellab.org). Sequencing reads were then aligned to the human reference genome Hg38 using a short-read aligner (Bowtie2) with the non-default parameters "X2000," "non-mixed," and "non-discordant." Reads from mitochondrial DNA were removed using SAMtools (http://samtools.sourceforge.net). Peak calls were made using MACS2 with the callpeak command (https://pypi.python.org/pypi/MACS2), with a threshold for peak calling set to FDR-adjusted $P < 0.05$. Circos (https://cran.r-project.org/web/packages/circlize/) plots were used for visualization of ATAC-Seq peaks in individual $CD4^+$ T cell subpopulations.

Viral outgrowth assays. PBMCs were plated at approximately 20,000 cells/well and stimulated with 1 μg/ml phytohemagglutinin (PHA) for 48 hours. Subsequently, PHA was washed away, and 10,000 MOLT-4 $CCR5^+$ cells (NIH AIDS Reagent Program, catalog 4984) were added to each well on the first and seventh day to propagate infections. On the tenth day, culture supernatants from each well were individually incubated with 10,000 TZM-bl cells (NIH AIDS Reagent Program, 8129) to drive *Tat*-dependent luciferase production. On the twelfth day, TZM-bl cells were lysed, and luciferase activity was measured using britelite plus (PerkinElmer, 6066761). Luciferase-positive wells were defined as having signal levels >2.5-fold higher than negative controls. Cells from positive wells were then harvested and plated onto upper compartments of Transwell tissue culture inserts (Costar 6.5 mm Transwell, 0.4 μm Pore Polyester Membrane Inserts, STEMCELL Technologies, 38024), while 700,000 MOLT cells were placed at lower compart-

ments. After 7 additional days of culture, MOLT cells from the bottom well were harvested, subjected to near-full-length HIV-1 DNA amplification, and submitted for next-generation sequencing using Illumina MiSeq.

Statistics. Data are presented as Circos plots, pie charts, bar charts, scatter plots, and box-and-whisker plots overlaid with individual values. Differences were tested for statistical significance using Mann-Whitney *U* tests (2-tailed), Fisher's exact tests (right-tailed, using Ingenuity Pathway Analysis; QIAGEN), Wilcoxon's tests (2-tailed), or χ^2 tests (2-tailed) as appropriate. Correlations were determined by the Spearman's rank method. *P* values less than 0.05 were considered significant. All *P* values are presented as nominal (non-FDR-adjusted) *P* values. Analyses were performed using Prism (GraphPad), Ingenuity Pathway Analysis, and the "ggpubr" statistical package in R (R Project for Statistical Computing).

Study approval. Study participants gave written informed consent to participate in accordance with the Declaration of Helsinki. The study was approved by the institutional review board of Massachusetts General Hospital and Brigham and Women's Hospital.

Author contributions

KBE, GQL, CJ, and XL performed WGA and HIV-1 sequencing; RS, KBE, and GQL, integration site analysis; SH, RNA-Seq; XS, ATAC-Seq; CG and GQL, bioinformatics analysis; FZC, gene ontology analysis; SMYC and GQL, viral outgrowth assays. TWC contributed PBMC samples. KBE, GQL, CG, RS, JZL, XGY, and ML interpreted, analyzed, presented data. KBE, GQL, ESR, XGY, and ML conceived, designed, and discussed the study. XGY and ML developed the research idea and supervised the study.

Acknowledgments

ML is supported by NIH grants AI098487, AI130005, AI122377, AI114235, AI117841, AI120008, and AI124776. XGY is supported by NIH grants AI116228, AI078799, HL134539, AI125109, and DA047034. JZL is supported by NIH grant AI125109. ML is an associate member of the BEAT-HIV Martin Delaney Collaboratory (UM1AI126620). GQL and SH are supported by training grants from the Harvard University Center for AIDS Research (CFAR), an NIH-funded program (AI060354) that is supported by the following NIH Institutes and Centers: NIAID, National Cancer Institute (NCI), National Institute of Mental Health (NIMH), National Institute on Drug Abuse (NIDA), Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), National Heart, Lung, and Blood Institute (NHLBI), and National Center for Complementary and Integrative Health (NCCAM). The authors gratefully acknowledge the MGH DNA core facility, Summer Zheng (Center for Biostatistics in AIDS Research, Harvard T.H. Chan School of Public Health) for statistical support, and Kelvin Yang for help with graphical design of figures.

Address correspondence to: Mathias Lichterfeld, Infectious Disease Division, Brigham and Women's Hospital, 65 Landsdowne Street, Cambridge, Massachusetts 02139, USA. Phone: 617.768.8399; Email: mlichterfeld@partners.org.

1. Schröder AR, Shinn P, Chen H, Berry C, Ecker JR, Bushman F. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell*. 2002;110(4):521–529.
2. Debyser Z, Christ F, De Rijck J, Gijsbers R. Host factors for retroviral integration site selection. *Trends Biochem Sci*. 2015;40(2):108–116.
3. Crooks AM, et al. Precise quantitation of the latent HIV-1 reservoir: implications for eradication strategies. *J Infect Dis*. 2015;212(9):1361–1365.
4. Finzi D, et al. Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science*. 1997;278(5341):1295–1300.
5. Bruner KM, et al. Defective proviruses rapidly accumulate during acute HIV-1 infection. *Nat Med*. 2016;22(9):1043–1049.
6. Churchill MJ, Deeks SG, Margolis DM, Siliciano RF, Swanstrom R. HIV reservoirs: what, where and how to target them. *Nat Rev Microbiol*. 2016;14(1):55–60.
7. Perreau M, Banga R, Pantaleo G. Targeted immune interventions for an HIV-1 cure. *Trends Mol Med*. 2017;23(10):945–961.
8. Ruelas DS, Greene WC. An integrated overview of HIV-1 latency. *Cell*. 2013;155(3):519–529.
9. Melamed A, Laydon DJ, Gillet NA, Tanaka Y, Taylor GP, Bangham CR. Genome-wide determinants of proviral targeting, clonal abundance and expression in natural HTLV-1 infection. *PLoS Pathog*. 2013;9(3):e1003271.
10. Kulkarni A, Taylor GP, Klose RJ, Schofield CJ, Bangham CR. Histone H2A monoubiquitylation and p38-MAPKs regulate immediate-early gene-like reactivation of latent retrovirus HTLV-1. *JCI Insight*. 2018;3(20):123196.
11. Simonetti FR, et al. Clonally expanded CD4+ T cells can produce infectious HIV-1 in vivo. *Proc Natl Acad Sci U S A*. 2016;113(7):1883–1888.
12. Cohn LB, et al. HIV-1 integration landscape during latent and active infection. *Cell*. 2015;160(3):420–432.
13. Wagner TA, et al. HIV latency. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. *Science*. 2014;345(6196):570–573.
14. Maldarelli F, et al. HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science*. 2014;345(6193):179–183.
15. Lee GQ, et al. Clonal expansion of genome-intact HIV-1 in functionally polarized Th1 CD4+ T cells. *J Clin Invest*. 2017;127(7):2689–2696.
16. Lee GQ, et al. Prevalence and clinical impacts of HIV-1 intersubtype recombinants in Uganda revealed by near-full-genome population and deep sequencing approaches. *AIDS*. 2017;31(17):2345–2354.
17. Serrao E, Cherepanov P, Engelman AN. Amplification, next-generation sequencing, and genomic DNA mapping of retroviral integration sites. *J Vis Exp*. 2016;(109):53840.
18. Paruzynski A, et al. Genome-wide high-throughput integrome analyses by nrLAM-PCR and next-generation sequencing. *Nat Protoc*. 2010;5(8):1379–1395.
19. Hosmane NN, et al. Proliferation of latently infected CD4+ T cells carrying replication-competent HIV-1: potential role in latent reservoir dynamics. *J Exp Med*. 2017;214(4):959–972.
20. Bui JK, et al. Proviruses with identical sequences comprise a large fraction of the replication-competent HIV reservoir. *PLoS Pathog*. 2017;13(3):e1006283.
21. Lorenzi JC, et al. Paired quantitative and qualitative assessment of the replication-competent HIV-1 reservoir and comparison with integrated proviral DNA. *Proc Natl Acad Sci U S A*. 2016;113(49):E7908–E7916.
22. Hiener B, et al. Identification of genetically intact HIV-1 proviruses in specific CD4+ T cells from effectively treated participants. *Cell Rep*. 2017;21(3):813–822.
23. Sherrill-Mix S, et al. HIV latency and integration site placement in five cell-based models. *Retrovirology*. 2013;10:90.
24. Jordan A, Bisgrove D, Verdin E. HIV reproducibly establishes a latent infection after acute infection of T cells in vitro. *EMBO J*. 2003;22(8):1868–1877.
25. Battivelli E, et al. Distinct chromatin functional states correlate with HIV latency reactivation in infected primary CD4+ T cells. *Elife*. 2018;7:e34655.
26. Lewinski MK, et al. Genome-wide analysis of chromosomal features repressing human immunodeficiency virus transcription. *J Virol*. 2005;79(11):6610–6619.
27. Han Y, et al. Orientation-dependent regulation of integrated HIV-1 expression by host gene transcriptional readthrough. *Cell Host Microbe*. 2008;4(2):134–146.
28. Cesana D, et al. HIV-1-mediated insertional

activation of STAT5B and BACH2 trigger viral reservoir in T regulatory cells. *Nat Commun*. 2017;8(1):498.
29. Chomont N, et al. HIV reservoir size and persistence are driven by T cell survival and homeostatic proliferation. *Nat Med*. 2009;15(8):893–900.
30. Buzon MJ, et al. HIV-1 persistence in CD4+ T cells with stem cell-like properties. *Nat Med*. 2014;20(2):139–142.
31. Shan L, et al. Influence of host gene transcription level and orientation on HIV-1 latency in a primary-cell model. *J Virol*. 2011;85(11):5384–5393.
32. Chen HC, Martinez JP, Zorita E, Meyerhans A, Filion GJ. Position effects influence HIV latency reversal. *Nat Struct Mol Biol*. 2017;24(1):47–54.
33. Lenasi T, Contreras X, Peterlin BM. Transcriptional interference antagonizes proviral gene expression to promote HIV latency. *Cell Host Microbe*. 2008;4(2):123–133.
34. Archin NM, et al. Administration of vorinostat disrupts HIV-1 latency in patients on antiretroviral therapy. *Nature*. 2012;487(7408):482–485.
35. Vranckx LS, et al. LEDGIN-mediated inhibition of integrase-LEDGF/p75 interaction reduces reactivation of residual latent HIV. *EBioMedicine*. 2016;8:248–264.
36. Kuo HH, et al. Anti-apoptotic protein BIRC5 maintains survival of HIV-1-infected CD4+ T cells. *Immunity*. 2018;48(6):1183–1194.e5.
37. Brady T, et al. Integration target site selection by a resurrected human endogenous retrovirus. *Genes Dev*. 2009;23(5):633–642.
38. Fraietta JA, et al. Disruption of TET2 promotes the therapeutic efficacy of CD19-targeted T cells. *Nature*. 2018;558(7709):307–312.
39. R Core Team. R: a language and environment for statistical computing. R-project. http://www.R-project.org/. Accessed December 17, 2018.
40. Rose PP, Korber BT. Detecting hypermutations in viral sequences with an emphasis on G --> A hypermutation. *Bioinformatics*. 2000;16(4):400–401.
41. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32(5):1792–1797.
42. Larkin MA, et al. Clustal W and Clustal X version 2.0. *Bioinformatics*. 2007;23(21):2947–2948.
43. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754–1760.
44. Kent WJ. BLAT--the BLAST-like alignment tool. *Genome Res*. 2002;12(4):656–664.
45. Trombetta JJ, Gennert D, Lu D, Satija R, Shalek AK, Regev A. Preparation of single-cell RNA-Seq libraries for next generation sequencing. *Curr Protoc Mol Biol*. 2014;107:4.22.1–4.2217.
46. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12:323.
47. Corces MR, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat Genet*. 2016;48(10):1193–1203.
48. Corces MR, et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat Methods*. 2017;14(10):959–962.