Article

Distinct viral reservoirs in individuals with spontaneous control of HIV-1

https://doi.org/10.1038/s41586-020-2651-8

Received: 2 October 2019

Accepted: 15 July 2020

Published online: 26 August 2020

Check for updates

Chenyang Jiang[1,2,15], Xiaodong Lian[1,2,15], Ce Gao[1,15], Xiaoming Sun[1], Kevin B. Einkauf[1,2], Joshua M. Chevalier[1,2], Samantha M. Y. Chen[1], Stephane Hua[1], Ben Rhee[1,2], Kaylee Chang[1], Jane E. Blackmer[1], Matthew Osborn[1], Michael J. Peluso[3], Rebecca Hoh[3], Ma Somsouk[3], Jeffrey Milush[3], Lynn N. Bertagnolli[4], Sarah E. Sweet[4], Joseph A. Varriale[4], Peter D. Burbelo[5], Tae-Wook Chun[6], Gregory M. Laird[7], Erik Serrao[8,9], Alan N. Engelman[8,9], Mary Carrington[1,10], Robert F. Siliciano[4,11], Janet M. Siliciano[4,11], Steven G. Deeks[3], Bruce D. Walker[1,11,12,13], Mathias Lichterfeld[1,2,14] & Xu G. Yu[1,2] ✉

Sustained, drug-free control of HIV-1 replication is naturally achieved in less than 0.5% of infected individuals (here termed 'elite controllers'), despite the presence of a replication-competent viral reservoir[1]. Inducing such an ability to spontaneously maintain undetectable plasma viraemia is a major objective of HIV-1 cure research, but the characteristics of proviral reservoirs in elite controllers remain to be determined. Here, using next-generation sequencing of near-full-length single HIV-1 genomes and corresponding chromosomal integration sites, we show that the proviral reservoirs of elite controllers frequently consist of oligoclonal to near-monoclonal clusters of intact proviral sequences. In contrast to individuals treated with long-term antiretroviral therapy, intact proviral sequences from elite controllers were integrated at highly distinct sites in the human genome and were preferentially located in centromeric satellite DNA or in Krüppel-associated box domain-containing zinc finger genes on chromosome 19, both of which are associated with heterochromatin features. Moreover, the integration sites of intact proviral sequences from elite controllers showed an increased distance to transcriptional start sites and accessible chromatin of the host genome and were enriched in repressive chromatin marks. These data suggest that a distinct configuration of the proviral reservoir represents a structural correlate of natural viral control, and that the quality, rather than the quantity, of viral reservoirs can be an important distinguishing feature for a functional cure of HIV-1 infection. Moreover, in one elite controller, we were unable to detect intact proviral sequences despite analysing more than 1.5 billion peripheral blood mononuclear cells, which raises the possibility that a sterilizing cure of HIV-1 infection, which has previously been observed only following allogeneic haematopoietic stem cell transplantation[2,3], may be feasible in rare instances.

Individuals with untreated HIV-1 infections who durably control HIV-1 replication below the threshold of detection of commercial viral load assays (here termed 'elite controllers') may represent the closest possible approximation to a natural cure of HIV-1 infection[1]. Previous studies have linked elite HIV-1 control to specific variations in the human HLA class I gene locus[4], and to the presence of highly functional cellular immune responses[5] that have stronger abilities to kill virus-infected cells[5], target mutationally constrained epitopes[6] and limit viral escape[7]. Although the persistence of small, replication-competent proviral reservoirs has been documented in elite controllers[8,9], the characteristics and possible distinguishing features of reservoir cells in this specific group of individuals remain poorly defined.

We used full-length individual provirus sequencing (FLIP-seq)[10] to profile the proviral reservoir landscape at single-genome resolution of a large cohort of elite controllers who maintained undetectable HIV-1 plasma viral loads for a median of 9 years (range, 1–24 years) based on commercially available PCR assays. A reference cohort of individuals with HIV-1 infections who were treated with suppressive antiretroviral

[1]Ragon Institute of MGH, MIT and Harvard, Cambridge, MA, USA. [2]Infectious Disease Division, Brigham and Women's Hospital, Boston, MA, USA. [3]Department of Medicine, University of California at San Francisco, San Francisco, CA, USA. [4]Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [5]Dental Clinical Research Core, National Institute of Dental and Craniofacial Research, National Institutes of Health, Bethesda, MD, USA. [6]National Institute of Allergies and Infectious Diseases, Bethesda, MD, USA. [7]Accelevir Diagnostics, Baltimore, MD, USA. [8]Department of Cancer Immunology and Virology, Dana-Farber Cancer Institute, Boston, MA, USA. [9]Department of Medicine, Harvard Medical School, Boston, MA, USA. [10]Basic Science Program, Frederick National Laboratory for Cancer Research, Frederick, MD, USA. [11]Howard Hughes Medical Institute, Chevy Chase, MD, USA. [12]Institute for Medical Engineering and Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. [13]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. [14]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [15]These authors contributed equally: Chenyang Jiang, Xiaodong Lian, Ce Gao. ✉e-mail: xyu@mgh.harvard.edu

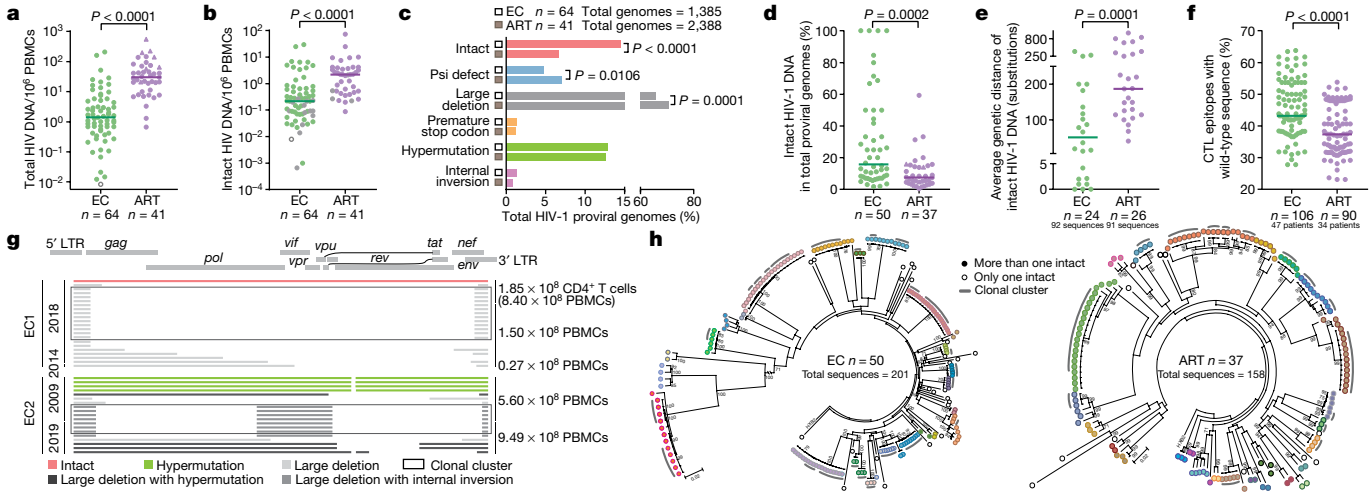


Fig. 1 | Proviral reservoir landscape in HIV-1 elite controllers. a, **b**, Relative frequencies of total (**a**) and near-full-length intact (**b**) HIV-1 DNA sequences in elite controllers (EC) and ART-treated individuals (ART). Grey symbols, limit of detection (expressed as 1 copy/total number of analysed cells without target identification). Circles, proviral sequences obtained from unfractionated PBMCs; triangles, proviral sequences retrieved from isolated CD4$^+$ T cells and normalized to the number of PBMCs. Open circles show the Berlin patient. **c**, Proportions of proviral sequences that have an intact genome or display defined structural defects among all proviral genomes. Psi, packaging signal. **d**, Proportion of genome-intact proviral sequences among all proviral genomes from each study participant. Only individuals for whom at least one genome-intact proviral sequence was detected are shown. **e**, Average genetic distance between distinct genome-intact proviral sequences obtained from each study participant. Participants with at least two detectable genome-intact proviral sequences are included. **f**, Proportion of optimal CTL epitopes (restricted by autologous HLA class I isotypes) with wild-type clade B consensus sequences within a given clade B genome-intact proviral sequence. Each dot represents data from one genome-intact proviral sequence. Clonal sequences are counted once. **g**, All proviral HIV-1 sequences isolated from EC1 and EC2. Dates of sample collection are indicated on the left; numbers of cells analysed are indicated on the right. Open boxes indicate clonal clusters. **h**, Circular maximum-likelihood phylogenetic trees for all genome-intact proviral sequences from elite controllers and ART-treated individuals. HXB2, reference HIV-1 sequence. Dots with the same colours indicate genome-intact proviral sequences that were detected in the same individuals. Clonal sequences, defined by complete sequence identity, are indicated by grey arches. Bootstrap analysis with 1,000 replicates was performed to assign confidence to tree nodes; bootstrap support values >70% are shown in the trees. Two-tailed Mann–Whitney *U*-tests were used for data shown in **a**, **b**, **d**–**f**; false-discovery rate (FDR)-adjusted two-tailed Fisher's exact tests were used for data shown in **c**.

therapy (ART) for a median of 9 years (range, 2–19 years) was recruited for comparative purposes (Extended Data Table 1). Collectively, our analysis of a large number of individual HIV-1 proviral genomes (n = 1,385 from 64 elite controllers and n = 2,388 from 41 ART-treated individuals) demonstrated that the median number of proviral amplification products (intact and defective) per person was significantly lower in elite controllers relative to ART-treated individuals (Fig. 1a). Frequencies of near-full-length proviral sequences with intact genomes that did not contain defined lethal sequence defects were also markedly reduced in elite controllers, although their quantitative spectrum varied considerably (Fig. 1b). Of note, genome-intact proviral sequences made up a significantly larger proportion of all proviral sequences in elite controllers at both the cohort level (Fig. 1c) and the per-study participant level (Fig. 1d) compared to ART-treated individuals; in four elite controllers, genome-intact proviral sequences accounted for 100% of the detected proviral species. Intra-individual diversity in the proviral sequences, determined by pair-wise comparisons of all genome-intact proviral sequences within a given study participant, was smaller in elite controllers (Fig. 1e and Extended Data Fig. 1a). Notably, within genome-intact proviral sequences from elite controllers, optimal epitope sequences of cytotoxic T lymphocytes (CTLs) restricted by autologous HLA class I isotypes displayed more limited evidence of mutational escape (Fig. 1f and Extended Data Fig. 1c–f). These data suggest that genome-intact proviral sequences from elite controllers were seeded early in the disease process and persisted long-term.

For a more in-depth analysis of the structure of the proviral reservoir, we initially focused on two elite controllers for whom no genome-intact proviral sequences were observed in our initial analysis. For EC1—an individual who had maintained drug-free HIV-1 control for a recorded time of 12 years with only one documented episode of viraemia of 56 HIV-1 RNA copies per ml out of 23 viral load tests that spanned this period (Extended Data Fig. 2)—we increased the number of analysed peripheral blood mononuclear cells (PBMCs) to the limit of available cells and found a single genome-intact proviral sequence in a total of 1.02 billion PBMCs analysed; 21 defective proviruses, many of which belonged to a sequence-identical cluster (Fig. 1g), were also detected. For EC2—who had a single documented episode of 93 HIV-1 RNA copies per ml in 39 viral load tests that spanned more than 24 years of follow-up without ART (Extended Data Fig. 2)—we did not detect a single genome-intact proviral sequence in more than 1.5 billion PBMCs, although 19 defective proviral species, including near-full-length sequences with lethal hypermutations, were observed, which clearly indicates that this individual had been infected with HIV-1 in the past (Fig. 1g). Members of a sequence-identical cluster of defective proviral sequences with large deletions were noted in samples that had been collected in 2009 and in 2019 from EC2, demonstrating the durability of a clonal cell population that contains this sequence.

Moreover, a subsequent quantitative viral outgrowth assay (qVOA) with 340 million resting CD4$^+$ T cells isolated from approximately 1 billion PBMCs (collected in 2019), and an additional qVOA that included 41 million total CD4$^+$ T cells isolated from 158.5 million PBMCs (collected in 2009) did not retrieve a single replication-competent viral species. The recently developed intact proviral DNA assay did not find evidence of genome-intact proviral sequences in 14 million resting CD4$^+$ T cells, but confirmed the presence of defective HIV-1 DNA sequences (Extended Data Fig. 1b). In addition, an analysis of 7.72 million gut cells collected by colonoscopy from the rectum (2.08 million CD45$^+$ mononuclear cells and 2.30 million CD45$^-$ cells) and terminal ileum (1.99 million CD45$^+$ mononuclear cells and 1.35 million CD45$^-$ cells) by FLIP-seq did not reveal any intact or defective proviruses in samples from EC2.

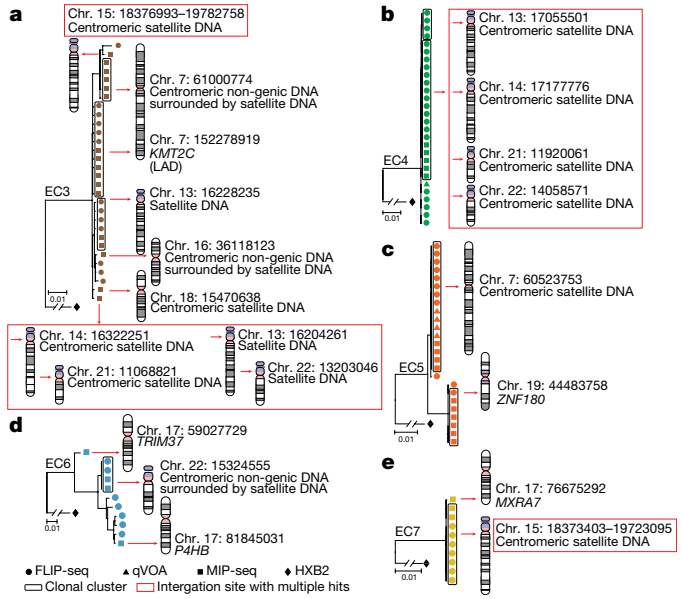


Fig. 2 | Increased frequency of genome-intact proviral sequences integrated in centromeric satellite DNA in elite controllers. a–e, Linear maximum-likelihood phylogenetic trees for genome-intact proviral sequences from five elite controllers are shown. Coordinates and relative positioning of integration sites are depicted; genes that contain integration sites are listed. Clonal genome-intact proviral sequences, defined by identical proviral sequences and identical corresponding integration sites, are highlighted in black boxes. Red boxes reflect multi-hit integration sites that cannot be definitively mapped to one particular genomic location owing to their position in repetitive centromeric satellite DNA that is present in multiple regions of the human genome. LAD, lamina-associated domain.

To our knowledge, the absence of genome-intact proviral sequences in such extremely large numbers of analysed cells has been documented only in the 'Berlin patient' who underwent an allogeneic haematopoietic stem cell transplantation from a donor who was homozygous for CCR5Δ32[2]; this resulted in what is widely considered a sterilizing cure of HIV-1 infection. Indeed, we did not retrieve any intact or defective proviral sequences using FLIP-seq in an analysis of 113 million PBMCs from the Berlin patient (collected in 2017 and 2018) (Fig. 1a, b). Although the logic of scientific discovery[11] does not allow us to confirm that EC2 has achieved a sterilizing cure of HIV-1 infection through natural immune-mediated mechanisms, it is notable that we have failed to falsify this hypothesis, despite analysing large amounts of cells with a range of complementary, highly sensitive detection techniques.

We next performed a phylogenetic analysis of all genome-intact proviral sequences obtained from 50 elite controllers and 37 ART-treated individuals. In both groups, we readily observed large clusters of sequences that were completely identical over entire analysed viral genomes (Fig. 1h), strongly suggesting that they originate from clonally expanded HIV-1-infected cells that passed on identical copies of genome-intact proviral sequences during cell divisions. The proportions of these genome-intact proviral sequences derived from clonally expanded cells were significantly higher in elite controllers than in ART-treated individuals (Extended Data Fig. 1g, h). A number of these sequences were also retrieved in qVOAs, indicating that these genome-intact proviral sequences are fully replication-competent (Figs. 2, 3).

For a detailed analysis of the viral reservoir landscape in elite controllers, we focused on eleven elite controllers (EC3–EC13), in whom large clusters of identical genome-intact proviral sequences were detected and from whom sufficient numbers of cells were available. We frequently observed oligoclonal, and sometimes almost monoclonal, compositions of the entire intact proviral reservoir landscape in cells from these individuals (Figs. 2, 3 and Extended Data Fig. 3). Notably, such a narrowly focused configuration of the viral reservoir that consists of few distinct genome-intact proviral sequences but displays relatively large expansions of identical clones of genome-intact proviral sequences is compatible with very low—if any—levels of ongoing viral replication in these elite controllers. This structure of the viral reservoir is atypical relative to the more-diverse spectrum of genome-intact proviral sequences that have previously been described for long-term ART-treated individuals[10,12]. Instead, the landscape of the viral reservoir of EC3–EC13 is more similar to the oligoclonal structure of the viral reservoir of genome-intact proviral sequences that are typically observed in individuals with chronic human T-cell leukaemia virus type 1 infection, a retroviral disease that is characterized by deep proviral latency that limits active viral transcription and replication, such that viral propagation occurs almost exclusively by mitotic spread during clonal proliferation of infected T cells[13]. On the basis of these considerations, we hypothesized that genome-intact proviral sequences from elite controllers maintain a state of deep, long-lasting latency, possibly owing to chromosomal integration into genomic regions that are not permissive to active viral transcription.

To investigate the chromosomal positions of genome-intact proviral sequences, we used matched integration site and proviral sequencing (MIP-seq)[14] to analyse integration sites together with the corresponding proviral sequences. In brief, proviral DNA was diluted to single-genome levels, amplified by Φ29-catalysed whole-genome amplification and analysed with near-full-length proviral sequencing[14] and integration site analysis using 'integration site loop amplification'[15] or ligation-mediated PCR[16]. These experiments, performed on samples from the eleven elite controllers (EC3–EC13), identified a total of 92 integration sites that corresponded to genome-intact proviral sequences, of which 33 were associated with unique chromosomal locations (Supplementary Table 1). These integration sites of genome-intact proviral sequences were preferentially located in chromosomes 7, 17 and 19, and to a lesser extent in chromosomes 16 and 18 (Fig. 4a and Extended Data Fig. 5a). Consistent with previous studies[14] in which a total of 100 pairs of genome-intact proviral sequences and corresponding integration sites ($n = 73$ genome-intact proviral sequences with unique integration sites) were analysed for long-term ART-treated individuals, proviral species that displayed complete sequence identity shared the same integration sites, which confirms their clonal origin. Notably, upstream HIV-1 long-terminal repeat regions, which are not included in typical FLIP-seq assays[10,12] but that were specifically amplified in sequences from these individuals, also displayed complete sequence identity within analysed clonal proviral sequences (Extended Data Fig. 4).

Notably, integration site analysis revealed that a significantly larger proportion of genome-intact proviral sequences from elite controllers were located in non-genic or pseudogenic regions, relative to genome-intact proviral sequences from long-term ART-treated individuals analysed using the same approach[14] (45% compared with 17.8% of distinct genome-intact proviral sequences, respectively, $P = 0.0051$; 40.2% compared with 13% of all genome-intact proviral sequences, respectively, $P < 0.0001$), and in comparison to previous studies in which integration sites of HIV-1 proviral sequences from ART-treated individuals[15,17] were analysed without distinguishing intact from defective proviruses (Fig. 4b and Extended Data Fig. 5b). Further investigation revealed that the non-genic integration sites of genome-intact proviral sequences from elite controllers were frequently positioned in or surrounded by centromeric satellite or microsatellite DNA (EC3–EC7; Fig. 2a–e), non-coding regions of the human genome that consist of dense heterochromatin 'gene deserts'[18] that are typically disfavoured for HIV-1 integration[19]. Localization of proviral sequences in such centromeric satellite DNA has been associated with deep viral latency in functional viral reactivation studies[20,21] and was extremely rare[22]

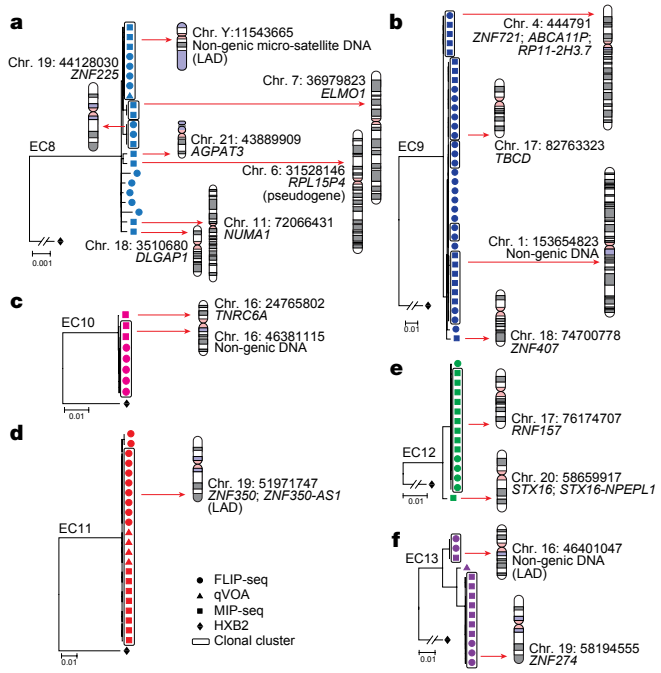


Fig. 3 | Preferential location of genome-intact proviral sequences from elite controllers in genes that encode KRAB-ZNF proteins. a–f, Linear maximum-likelihood phylogenetic trees of genome-intact proviral sequences from the indicated study participants are shown. Coordinates and relative positioning of integration sites are indicated. Other information is as described in the legend of Fig. 2.

or entirely undetectable in previous studies of ART-treated individuals[14]. In our study, the integration of genome-intact proviral sequences into centromeric satellite or microsatellite DNA was observed in a total of 8 unique genome-intact proviral sequences (24% of distinct genome-intact proviral sequences, 20.7% of all genome-intact proviral sequences) and occurred at least once in 5 (EC3–5, EC7 and EC8) (Figs. 2a–c, e, 3a) of the 11 elite controllers analysed. In addition, three integration sites of genome-intact proviral sequences were located in centromeric non-genic DNA surrounded by satellite DNA (EC3 and EC6) (Fig. 2a, d). Notably, as many as six different integration sites of genome-intact proviral sequences were located in or surrounded by centromeric satellite DNA in EC3 (Fig. 2a). In addition to this highly disproportionate overrepresentation of centromeric satellite DNA among integration sites of genome-intact proviral sequences from elite controllers, sequences from EC10 and EC13 contained integrations of clonal genome-intact proviral sequences in a large non-genic region in proximity to non-centromeric microsatellite DNA on chromosome 16 (Fig. 3c, f). Thus, in total, 39.4% of all 33 distinct genome-intact proviral sequences (32.6% of all 92 genome-intact proviral sequences) from elite controllers were located within or in proximity to satellite or microsatellite DNA.

Corresponding to the disproportionate enrichment of non-genic integration sites in elite controllers, we noted that the number of genic integration sites associated with genome-intact proviral sequences was significantly decreased in elite controllers, relative to ART-treated individuals[14]. These genic integration sites were almost exclusively located in introns of genes that, in comparison to long-term ART-treated individuals, showed weaker transcriptional activity (Extended Data Fig. 7a) and displayed an opposite orientation relative to the host gene, in which the proviral sequence was contained, in approximately 60% of all genic integration sites analysed (Extended Data Fig. 7b, c). Genes that encode members of the zinc-finger protein (ZNF) family and, in particular, Krüppel-associated box domain-containing ZNF (KRAB-ZNF) genes[23] accounted for 33% of all 18 genes that contained distinct genome-intact proviral sequences in elite controllers (corresponding to 49% of all 55 genic integration events of genome-intact proviral sequences), a notable enrichment relative to ART-treated individuals (Fig. 4c and Extended Data Fig. 5c). Clonal genome-intact proviral sequences were frequently integrated into KRAB-ZNF genes located in defined regions of chromosome 19[24] that display highly distinct chromatin features. In particular, these regions are extensively occupied by the heterochromatin proteins CBX1 and SUV39H1[25] and also show a strong enrichment for repressive chromatin marks that cover the lengths of ZNF genes but selectively spare the corresponding host transcriptional start sites[25]. Notably, a previous computational, genome-wide analysis of chromatin states based on the combinatorial evaluation of multiple different chromatin marks in their respective spatial context revealed that repetitive satellite DNA and ZNF genes share a common, highly distinct chromatin state (referred to as 'ZNF genes and repeats')[26]. When combined, genome-intact proviral sequences located either in satellite DNA or in ZNF genes represented more than 45% of all 33 independent genome-intact proviral sequences and more than 60% of all 92 genome-intact proviral sequences in elite controllers, proportions that were significantly increased relative to ART-treated individuals (Fig. 4d and Extended Data Fig. 5d).

To analyse the positioning of proviral integration sites relative to active transcription units in host DNA, we performed RNA-sequencing-based gene-expression profiling in autologous total $CD4^+$ T cells, as well as autologous central memory and effector memory $CD4^+$ T cell subsets, which contain the majority of the viral reservoir cells in peripheral blood[27]. These experiments showed a significantly increased chromosomal distance between the integration sites of genome-intact proviral sequences and the most proximal host transcriptional start sites in elite controllers, relative to long-term ART-treated individuals[14] (Fig. 4e). Simultaneously, we calculated the chromosomal distance between the coordinates of integration sites of genome-intact proviral sequences and accessible chromatin, as determined by genome-wide assays for transposase-accessible chromatin using sequencing (ATAC-seq) performed in autologous $CD4^+$ T cells. Although integration sites in satellite and microsatellite DNA were excluded from this analysis (and from the subsequent analysis using chromatin immunoprecipitation followed by sequencing (ChIP–seq), high-throughput chromatin conformation capture sequencing (Hi-C-seq) and methylation-sequencing data; see below) due to the reduced ability to map next-generation sequencing reads onto repetitive genomic DNA regions[28], we noted that integration sites of genome-intact proviruses from elite controllers were located at significantly increased distances from accessible chromatin, compared to those from ART-treated individuals[14] (Fig. 4f). These differences were observed when clonal sequences were counted only once (Fig. 4e, f) but were also notable when all clonal sequences were considered individually (Extended Data Fig. 5e, f).

In a subsequent analysis, we calculated the number of DNA reads associated with defined epigenetic histone marks in proximity to viral integration sites using ChIP–seq data from primary memory $CD4^+$ T cells available from the ROADMAP Epigenomics Project[26]. In comparison to ART-treated individuals[14], this analysis revealed a marked enrichment of the repressive histone feature H3K9me3 (on chromosomes 7 and 19) and/or a de-enrichment of the activating chromatin feature H3K4me1 (on chromosomes 17 and 19) at integration sites of genome-intact proviral sequences from elite controllers (Fig. 4g); a trend for differential expression of additional activating and inhibitory chromatin modifications in proximity to integration sites of genome-intact proviral sequences from elite controllers and ART-treated individuals was also noted (Extended Data Fig. 6a–d). Furthermore, an alignment of the coordinates of integration sites to three-dimensional chromosomal contact data generated by Hi-C-seq[29] demonstrated a significantly increased proportion of genome-intact proviral sequences from elite controllers located in compartment B, which mostly contains closed

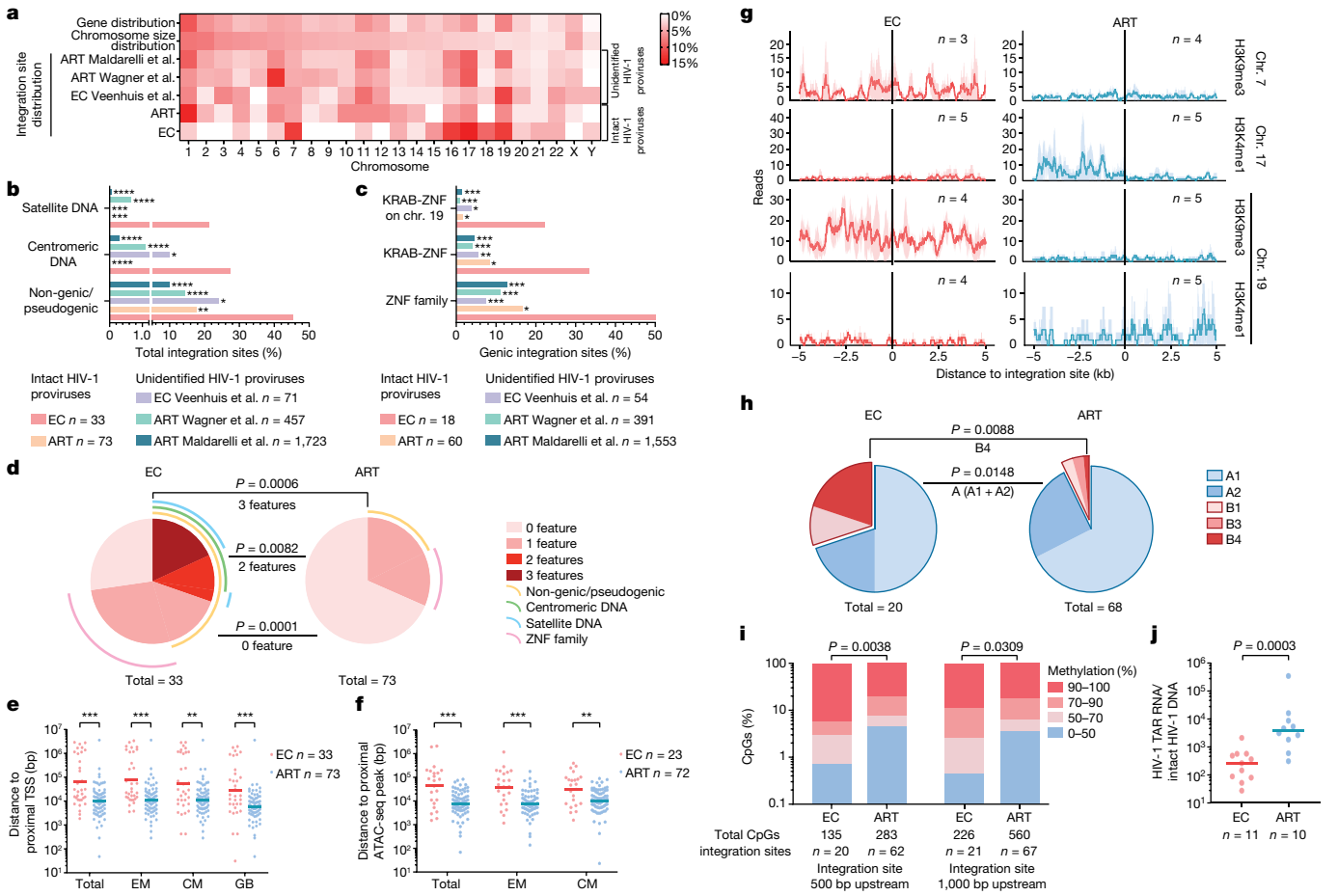


Fig. 4 | Distinct genomic and epigenetic features of integration sites of genome-intact proviral sequences from elite controllers. a, Relative proportion of proviral integration sites of genome-intact proviral sequences in each chromosome. Contributions of each chromosome to the total number of genes (first row) and to the total size of the human genome (second row) are included as references. **b**, **c**, Proportion of genome-intact proviral sequences located in the indicated genomic regions. **a**–**c**, Data from genome-intact proviral sequences in ART-treated individuals[14] and from unselected (intact and defective) proviral sequences from elite controllers (Veenhuis et al., ref. [9]) and ART-treated individuals (Wagner et al., ref. [15] and Maldarelli et al., ref. [17]) are shown as references. **d**, SPICE diagrams show the proportions of genome-intact proviral sequences with the indicated integration site features in elite controllers and ART-treated individuals. **e**, **f**, Chromosomal distance between integration sites of genome-intact proviral sequences and the most proximal transcriptional start sites (TSS) in autologous total, effector memory or central memory CD4$^+$ T cells or from the Genome Browser (GB) (**e**), or to the most proximal ATAC-seq peaks (**f**) in autologous total, effector memory and central memory CD4$^+$ T cells. Horizontal lines show the geometric mean. **g**, Numbers of DNA-sequencing reads associated with activating (H3K4me1) or repressive (H3K9me3) histone protein modifications in proximity to integration sites from elite controllers and long-term ART-treated individuals[14]. Median and confidence intervals (one standard deviation) of ChIP–seq data from primary memory CD4$^+$ T cells included in the ROADMAP repository[26] are shown. **h**, Proportions of genome-intact proviral sequences located in structural compartments A and B (and associated sub-compartments), as determined by Hi-C-seq data[29]. Integration sites in regions not covered in a previous study[29] were excluded. **i**, Numbers of cytosine residues with indicated levels of methylation (derived from CD4$^+$ T cells in the iMethyl database[31]) in proximity (500 or 1,000 bp upstream of the 5′ long-terminal repeat (LTR) host–viral junction) to integration sites from elite controllers and ART-treated individuals. **j**, Frequencies of HIV-1 RNA transcripts in PBMCs from elite controllers and ART-treated individuals, normalized to the corresponding number of genome-intact proviral sequences determined by FLIP-seq. **a**–**i**, Clonal sequences were only counted once. **f**–**i**, Sequences in genomic regions included in the ENCODE blacklist[28] were excluded. ****$P < 0.0001$, ***$P < 0.001$, **$P < 0.01$, *$P < 0.05$; data were analysed using two-sided Fisher's exact tests (**b**–**d**, **h**), two-sided Mann–Whitney *U*-tests (**e**, **f**, **j**) or two-tailed χ^2 test (**i**); **b**, **c**, **e**, **f**, **i**, FDR-adjusted *P* values are shown; **d**, **h**, **j**, nominal *P* values are shown. All comparisons were made between elite controllers and reference groups.

chromatin. This effect was particularly obvious for integration sites in KRAB-ZNF genes on chromosome 19 in elite controllers, which were all located in subcompartment B4 (Fig. 4h and Extended Data Fig. 5g). This very small compartment (which accounts for approximately 0.3% of the human genome) is known to contain dense heterochromatin marks[29] and represents a highly atypical location of a chromosomal integration site for HIV-1 in non-controller individuals[14]. A highly increased frequency of genome-intact proviral sequences from elite controllers in compartment B was also noted when Hi-C-seq data from Jurkat cells[30] were used for alignment (Extended Data Fig. 6e, f).

Taking advantage of previously published genome-wide bisulfite sequencing data of CD4$^+$ T cells[31], we observed that the frequency of hypermethylated (more than 90% methylation) cytosine residues was significantly higher in proximity to genome-intact proviral sequences from elite controllers, relative to integration sites of genome-intact proviral sequences from long-term ART-treated individuals[14] (Fig. 4i). These data suggest that chromosomal regions that are more susceptible to DNA methyltransferases represent preferential sites for the long-term persistence of genome-intact proviral sequences in elite controllers, arguably because the integration into hypermethylated genomic DNA might facilitate deep latency of genome-intact proviral sequences and protect against immune-cell targeting. Given that closely neighbouring cytosine residues are likely to share the same methylation status[32], these results raise the possibility that HIV-1

promoter methylation, which has previously been shown to induce proviral HIV-1 silencing in in vitro assays[33], may contribute to durable transcriptional repression of genome-intact proviral sequences from elite controllers. The frequencies of genome-intact proviral sequences located in lamina-associated domains—genomic regions that interact with the inner nuclear membrane, mostly contain closed chromatin and represent a rare target for HIV-1 integration[34]—were not significantly different between genome-intact proviral sequences from elite controllers and ART-treated individuals when clonal sequences were counted only once; however, a significant enrichment of genome-intact proviral sequences from elite controllers in lamina-associated domains was noted when clonal genome-intact proviral sequences were counted as independent proviruses (Extended Data Fig. 7d, e).

Given that non-coding centromeric satellite DNA is a highly disfavoured target site for HIV-1 integration[19], the disproportionately increased number of integration sites in satellite DNA described here is a remarkable feature of elite controllers. Notably, elite controllers expressed normal mRNA levels of *LEDGF* (also known as *PSIP1* or *p75*) and *CPSF6* (Extended Data Fig. 7f), host factors that interact directly with HIV-1 proteins to bias HIV-1 integration site selection to active transcription units[35,36]. Although protein levels of these molecules were not assessed, these results suggest that there is no increased susceptibility of centromeric satellite DNA to HIV-1 integration in elite controllers. To further address this, we infected $CD4^+$ T cells from $n = 12$ elite controllers from our study cohort and $n = 9$ HIV-1-negative healthy individuals with a GFP-encoding HIV-1 construct, followed by sorting of GFP^+ and GFP^- $CD4^+$ T cells and subsequent integration site analysis. These experiments, in which more than 120,000 independent HIV-1 integration coordinates were obtained, showed that integration sites in satellite DNA accounted for extremely low proportions of all integration events (0.04–0.06% in GFP^+ and 0.11–0.12% in GFP^- $CD4^+$ T cells), irrespective of the analysed study cohort (Extended Data Fig. 8a, b and Supplementary Table 2). Moreover, there was no evidence for preferential targeting of non-genic chromosomal regions or genes that encode KRAB-ZNF proteins in $CD4^+$ T cells from elite controllers that were infected in vitro (Extended Data Fig. 8b, c).

In conclusion, this work identifies a markedly distinct reservoir landscape of intact proviral sequences in PBMCs from individuals with durable natural control of HIV-1, characterized by features of integration sites that are highly suggestive of deep latency. For additional functional validation of this conclusion, we analysed the frequency of cell-associated HIV-1 RNA transcripts in elite controllers and ART-treated individuals; these additional experiments demonstrated that the number of cell-associated HIV-1 RNA copies, normalized to the corresponding number of genome-intact proviral sequences, was significantly lower in elite controllers (Fig. 4j). As such, elite controllers seem to exemplify attributes of a 'block and lock' mechanism[37] of viral control, which is defined by silencing of proviral gene expression through chromosomal integration into repressive chromatin locations[38]. We propose that the distinct reservoir configuration in elite controllers is not related to altered preferences for integration site locations during acute HIV-1 infection in elite controllers, but instead represents the result of cell-mediated immune selection forces that preferentially eliminate proviral sequences that are more permissive to viral transcription, in a process that we suggest referring to as the 'autologous shock and kill' mechanism. By contrast, less transcriptionally active proviral sequences with features of deep latency, leading to lower vulnerability to immune recognition, seem to persist long-term. In very rare cases, such as in EC1 and EC2, such selection forces may have accomplished near-complete clearance of all genome-intact proviral sequences, raising the possibility that a sterilizing cure of HIV-1 infection can, at least in principle, spontaneously occur through natural, immune-mediated mechanisms. Future studies will be necessary to determine whether signs of immune-mediated selection pressure on viral reservoir cells are also found in genome-intact proviral sequences from lymphoid tissues, which contain the majority of viral reservoir cells[39].

Although our data strongly suggest that deep latency has a role in maintaining spontaneous, drug-free control of HIV-1 in some elite controllers, deep viral latency is not completely permanent or irreversible, as reflected by our ability to retrieve replication-competent virus from elite controllers in in vitro qVOAs. However, in vitro qVOAs with maximum stimuli are unlikely to adequately reflect the susceptibility to viral reactivation in vivo; indeed, in vitro viral outgrowth may largely be a stochastic process[12,40], and may occur independently of molecular pathways that fine-tune the outgrowth behaviour of the virus in vivo. Nevertheless, it is likely that deep viral latency in elite controllers is a dynamic process, and that occasional bursts of viral transcription may occur despite genomic and epigenetic features of integration sites restricting viral gene expression. In fact, a proviral landscape with low permissiveness to viral reactivation stimuli may expose the immune system to a tailored viral antigen dose that can maintain a highly functional antiviral T cell response, a hallmark of antiviral immunity in elite controllers[5], without supporting high-level viral replication promoting cytotoxic T cell exhaustion. Therefore, a reciprocal equilibrium between a weakly inducible viral reservoir and an efficient HIV-1-specific $CD8^+$ T cell response may represent the cornerstone of natural HIV-1 immune control. Given that evidence for selection of genome-intact proviral sequences with features of deeper latency was also observed in long-term ART-treated individuals, albeit to a weaker degree[14], it is hoped that future longitudinal evaluations will be informative for designing strategies to induce long-term drug-free remission of HIV-1 infection in larger populations of individuals.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2651-8.

1. Sáez-Cirión, A. & Pancino, G. HIV controllers: a genetically determined or inducible phenotype? *Immunol. Rev.* **254**, 281–294 (2013).
2. Yukl, S. A. et al. Challenges in detecting HIV persistence during potentially curative interventions: a study of the Berlin patient. *PLoS Pathog.* **9**, e1003347 (2013).
3. Gupta, R. K. et al. HIV-1 remission following CCR5Δ32/Δ32 haematopoietic stem-cell transplantation. *Nature* **568**, 244–248 (2019).
4. McLaren, P. J. & Carrington, M. The impact of host genetic variation on infection with HIV-1. *Nat. Immunol.* **16**, 577–583 (2015).
5. Migueles, S. A. et al. Lytic granule loading of $CD8^+$ T cells is required for HIV-infected cell elimination associated with immune control. *Immunity* **29**, 1009–1021 (2008).
6. Gaiha, G. D. et al. Structural topology defines protective CD8+ T cell epitopes in the HIV proteome. *Science* **364**, 480–484 (2019).
7. Migueles, S. A. & Connors, M. Success and failure of the cellular immune response against HIV-1. *Nat. Immunol.* **16**, 563–570 (2015).
8. Blankson, J. N. et al. Isolation and characterization of replication-competent human immunodeficiency virus type 1 from a subset of elite suppressors. *J. Virol.* **81**, 2508–2518 (2007).
9. Veenhuis, R. T. et al. Long-term remission despite clonal expansion of replication-competent HIV-1 isolates. *JCI Insight* **3**, e122795 (2018).
10. Lee, G. Q. et al. Clonal expansion of genome-intact HIV-1 in functionally polarized Th1 $CD4^+$ T cells. *J. Clin. Invest.* **127**, 2689–2696 (2017).
11. Popper, K. *Die Logik der Forschung. Zur Erkenntnistheorie der modernen Naturwissenschaft* (Springer, 1935).
12. Ho, Y. C. et al. Replication-competent noninduced proviruses in the latent reservoir increase barrier to HIV-1 cure. *Cell* **155**, 540–551 (2013).
13. Melamed, A. et al. Genome-wide determinants of proviral targeting, clonal abundance and expression in natural HTLV-1 infection. *PLoS Pathog.* **9**, e1003271 (2013).
14. Einkauf, K. B. et al. Intact HIV-1 proviruses accumulate at distinct chromosomal positions during prolonged antiretroviral therapy. *J. Clin. Invest.* **129**, 988–998 (2019).
15. Wagner, T. A. et al. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. *Science* **345**, 570–573 (2014).
16. Cohn, L. B. et al. HIV-1 integration landscape during latent and active infection. *Cell* **160**, 420–432 (2015).
17. Maldarelli, F. et al. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science* **345**, 179–183 (2014).

18. McNulty, S. M. & Sullivan, B. A. Alpha satellite DNA biology: finding function in the recesses of the genome. *Chromosome Res*. **26**, 115–138 (2018).
19. Carteau, S., Hoffmann, C. & Bushman, F. Chromosome structure and human immunodeficiency virus type 1 cDNA integration: centromeric alphoid repeats are a disfavored target. *J. Virol*. **72**, 4005–4014 (1998).
20. Jordan, A., Bisgrove, D. & Verdin, E. HIV reproducibly establishes a latent infection after acute infection of T cells in vitro. *EMBO J*. **22**, 1868–1877 (2003).
21. Lewinski, M. K. et al. Genome-wide analysis of chromosomal features repressing human immunodeficiency virus transcription. *J. Virol*. **79**, 6610–6619 (2005).
22. Schröder, A. R. et al. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110**, 521–529 (2002).
23. Ecco, G., Imbeault, M. & Trono, D. KRAB zinc finger proteins. *Development* **144**, 2719–2729 (2017).
24. Lukic, S., Nicolas, J. C. & Levine, A. J. The diversity of zinc-finger genes on human chromosome 19 provides an evolutionary mechanism for defense against inherited endogenous retroviruses. *Cell Death Differ*. **21**, 381–387 (2014).
25. Vogel, M. J. et al. Human heterochromatin proteins form large domains containing KRAB-ZNF genes. *Genome Res*. **16**, 1493–1504 (2006).
26. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
27. Chomont, N. et al. HIV reservoir size and persistence are driven by T cell survival and homeostatic proliferation. *Nat. Med*. **15**, 893–900 (2009).
28. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: identification of problematic regions of the genome. *Sci. Rep*. **9**, 9354 (2019).
29. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
30. Lucic, B. et al. Spatially clustered loci with multiple enhancers are frequent targets of HIV-1 integration. *Nat. Commun*. **10**, 4059 (2019).
31. Komaki, S. et al. iMETHYL: an integrative database of human DNA methylation, gene expression, and genomic variation. *Hum. Genome Var*. **5**, 18008 (2018).
32. Affinito, O. et al. Nucleotide distance influences co-methylation between nearby CpG sites. *Genomics* **112**, 144–150 (2020).
33. Kauder, S. E., Bosque, A., Lindqvist, A., Planelles, V. & Verdin, E. Epigenetic regulation of HIV-1 latency by cytosine methylation. *PLoS Pathog*. **5**, e1000495 (2009).
34. Marini, B. et al. Nuclear architecture dictates HIV-1 integration site selection. *Nature* **521**, 227–231 (2015).
35. Ciuffi, A. et al. A role for LEDGF/p75 in targeting HIV DNA integration. *Nat. Med*. **11**, 1287–1289 (2005).
36. Achuthan, V. et al. Capsid–CPSF6 interaction licenses nuclear HIV-1 trafficking to sites of viral DNA integration. *Cell Host Microbe* **24**, 392–404 (2018).
37. Debyser, Z., Vansant, G., Bruggemans, A., Janssens, J. & Christ, F. Insight in HIV integration site selection provides a block-and-lock strategy for a functional cure of HIV infection. *Viruses* **11**, 12 (2018).
38. Battivelli, E. et al. Distinct chromatin functional states correlate with HIV latency reactivation in infected primary $CD4^+$ T cells. *eLife* **7**, e34655 (2018).
39. Estes, J. D. et al. Defining total-body AIDS-virus burden with implications for curative strategies. *Nat. Med*. **23**, 1271–1276 (2017).
40. Weinberger, L. S., Burnett, J. C., Toettcher, J. E., Arkin, A. P. & Schaffer, D. V. Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic diversity. *Cell* **122**, 169–182 (2005).

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Study participants

HIV-1-infected study participants were recruited at the Massachusetts General Hospital (MGH), the Brigham and Women's Hospital (BWH) and at the University of California, San Francisco (UCSF) at the Zuckerberg San Francisco General Hospital. PBMCs and tissue samples were obtained according to protocols approved by the respective Institutional Review Boards. Clinical and demographical characteristics of study participants are summarized in Extended Data Table 1.

Droplet digital PCR

DNA was extracted from PBMCs or $CD4^+$ T cells isolated from total PBMCs (CD4 T Cell Isolation Kit, Miltenyi Biotec, 130-096-533) using commercial kits (Qiagen, DNeasy, 69504). We amplified total HIV-1 DNA using droplet digital PCR (ddPCR; Bio-Rad), using primers and probes that have previously been described[10] (127-bp 5′LTR–*gag* amplicon; HXB2 coordinates 684–810). PCR was performed using the following program: 95 °C for 10 min, 45 cycles of 94 °C for 30 s and 60 °C for 1 min, 72 °C for 1 min. The droplets were subsequently read by the QX200 droplet reader and data were analysed using QuantaSoft software (Bio-Rad).

Whole-genome amplification

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 multiple displacement amplification (MDA) with Φ29 polymerase (Qiagen, REPLI-g Single Cell Kit, 150345), as per the manufacturer's protocol. Following this unbiased whole-genome amplification step[41], DNA from each well was split and separately subjected to viral sequencing and integration site analysis, as described below. If necessary, a second-round multiple displacement amplification reaction was performed to increase the amount of available DNA.

HIV-1 near-full-genome sequencing

DNA resulting from whole-genome amplification reactions was subjected to near-full-length HIV-1 genome amplification using a one-amplicon and/or non-multiplexed five-amplicon approach, as previously described[14]. PCR products were visualized by agarose gel electrophoresis (Quantify One and ChemiDoc MP Image Lab, Bio-Rad). All near-full-length and/or five-amplicon-positive amplicons were subjected to Illumina MiSeq sequencing at the MGH DNA Core Facility. Resulting short reads were de novo assembled using Ultracycler v.1.0 and aligned to HXB2 to identify large deleterious deletions (<8,000 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 Python programming language (https://github.com/BWH-Lichterfeld-Lab/Intactness-Pipeline), consistent with previous studies[10,42,43,44]. The presence or absence of APOBEC-3G/3F-associated hypermutations was determined using the Los Alamos National Laboratory (LANL) HIV Sequence Database Hypermut 2.0[45] program. Viral sequences that lacked all mutations listed above were classified as 'genome-intact' sequences. Sequence alignments were performed using MUSCLE[46]. Phylogenetic distances between sequences were examined using maximum-likelihood trees in MEGA (https://www.megasoftware.net/) and MAFFT (https://mafft.cbrc.jp/alignment/software), and visualized using Highlighter plots (https://www.hiv.lanl.gov/content/sequence/HIGHLIGHT/highlighter_top.html). 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. Clades of intact HIV-1 proviral sequences were determined using the LANL HIV-1 Sequence Database Recombinant Identification Program[47]. Within intact HIV-1 clade B sequences, the proportions of optimal CTL epitopes (restricted by autologous HLA class I alleles) that match the clade B consensus sequence and CTL escape variants restricted by selected HLA class I alleles and supertypes described in the LANL HIV Immunology Database (https://www.hiv.lanl.gov/content/index) were determined.

Integration site analysis

Integration sites associated with each proviral sequence were obtained using integration site loop amplification, as previously described[15], or by ligation-mediated PCR[16] (Lenti-X Integration Site Analysis Kit (Takara Bio, 631263)); DNA produced by whole-genome amplification was used as template. For selected clonal sequences, viral–host junction regions were also amplified using primers that anneal upstream of the integration site in host DNA and downstream of the integration site in viral DNA. Resulting PCR products were subjected to next-generation sequencing using Illumina MiSeq. MiSeq paired-end FASTQ files were demultiplexed; small reads (142 bp) were then aligned simultaneously to the human reference genome GRCh38 and HIV-1 reference genome HXB2 using bwa-mem[48]. Biocomputational identification of integration sites was performed according to previously described procedures[15,49]. In brief, chimeric reads containing both human and HIV-1 sequences were evaluated for mapping quality based on (1) HIV-1 coordinates mapping to the terminal nucleotides of the viral genome; (2) absolute counts of chimeric reads; and (3) depth of sequencing coverage in the host genome adjacent to the viral integration site. The final list of integration sites and corresponding chromosomal annotations was obtained using Ensembl (v.86, http://www.ensembl.org/index.html), the UCSC Genome Browser (http://www.genome.ucsc.edu/) and GENCODE (v.29, https://www.gencodegenes.org/). Repetitive genomic sequences containing HIV-1 integration sites were identified using RepeatMasker (http://www.repeatmasker.org/).

Cell sorting and flow cytometry

PBMCs were stained with monoclonal antibodies against CD4 (1:50, clone RPA-T4, Biolegend, 300518), CD3 (1:50, clone OKT3, Biolegend, 317332), CD45RO (1:40, clone UCHL1, Biolegend, 304236) and CCR7 (1:40, clone G043H7, Biolegend, 353216). Afterwards, 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 FACS Aria cell sorter (BD Biosciences) at 70 pounds per square inch. Cell sorting was performed by the Ragon Institute Imaging Core Facility at MGH and resulted in isolation of lymphocytes with the defined phenotypic characteristics of >95% purity. Data were analysed using FlowJo software (Treestar).

RNA-seq

Total RNA was extracted from sorted $CD4^+$ T cell populations using a PicoPure RNA Isolation Kit (Applied Biosystems, KIT0204). RNA-seq libraries were generated as previously described[50]. In brief, whole-transcriptome amplification and tagmentation-based library preparation was performed using SMART-seq2, followed by sequencing on a NextSeq 500 Instrument (Illumina). The quantification of transcript abundance was conducted using RSEM software (v1.2.22) supported by STAR aligner software (STAR 2.5.1b) and aligned to the GRCh38 human genome. Transcripts per million values were then normalized among all samples using the upper-quantile-normalization method.

ATAC-seq

A previously described protocol with some modifications[51,52] was used. In brief, 20,000 sorted cells were centrifuged at 1,500 rpm for 10 min

at 4 °C in a pre-cooled fixed-angle centrifuge. All of the supernatant was removed and a modified transposase mixture (including 25 µl of 2× TD buffer, 1.5 µl of TDE1, 0.5 µl of 1% digitonin, 16.5 µl of PBS, 6.5 µl of nuclease-free water) was added to the cells and incubated in a heat block at 37 °C for 30 min. Transposed DNA was purified using a ChIP DNA Clean & Concentrator Kit (Zymo Research, D5205) and eluted DNA fragments were used to amplify libraries. The libraries were quantified using an Agilent Bioanalyzer 2100 and the Qubit dsDNA High Sensitivity Assay Kit. All Fast-ATAC libraries were sequenced using paired-end, single-index sequencing on a NextSeq 500/550 instrument with v.2.5 Kits (75 cycles). 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 GRCh38 using a short-read aligner (Bowtie2, http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) with the non-default parameters 'X2000', 'non-mixed' and 'non-discordant'. Reads from mitochondrial DNA were removed using Samtools (http://www.htslib.org). 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 $P < 0.05$.

qVOAs

$CD4^+$ cells were isolated from PBMCs using the EasySep Human CD4 Positive Selection Kit II (STEMCELL Technologies 17852). Cells were plated in limiting dilutions based on the intact provirus reservoir size determined through FLIP-seq. Irradiated feeder PBMCs were added at 1×10^5 cells per well. Cells were activated with 1 µg/ml PHA for 4 days, which was subsequently washed away and 10,000 MOLT-4 $CCR5^+$ cells (NIH AIDS Reagent Program, 4984) were added to propagate infection. On the 13th and 20th days, 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 15th and 22nd days, TZM-bl cells were lysed, and luciferase activity was measured using Britelite Plus (PerkinElmer, 6066761). Luciferase-positive wells were defined as having signal levels that were >3-fold higher than negative controls. Cells from positive wells were then collected and plated into bottom compartments of Transwell tissue-culture inserts (Costar 6.5 mm Transwells, 0.4-µm pore polyester membrane inserts, STEMCELL, 38024), while 1×10^6 MOLT-4 cells were placed in top compartments. After five additional days of culture, MOLT-4 cells from the upper wells were collected and subjected to FLIP-seq. Large-scale quantitative viral outgrowth measurements on cells from patient EC2 were performed by a similar standard method[53] with a p24 ELISA assay used to detect outgrowth.

Intact proviral DNA assays

The intact proviral DNA assay (IPDA) uses ddPCR to quantify proviruses that lack overt fatal defects, especially large deletions and hypermutations, and was performed as previously described[54].

In vitro-infection assays

$CD4^+$ T cells were stimulated in RPMI medium supplemented with 10% fetal calf serum, recombinant IL-2 (50 U/ml), and an anti-CD3/CD8 bispecific antibody (0.5 µg/µl, NIH AIDS Reagent Program, 12277). Cells were infected on day 5 with a GFP-encoding NL4-3 construct with a BAL-derived R5-tropic envelope[55] at a multiplicity of infection (MOI) of 0.1 for 4 h at 37 °C. After two washes, cells were resuspended in medium and plated at 5×10^5 cells per well in a 24-well plate. On day 5, GFP^+ and GFP^- $CD4^+$ T cells were sorted. Cells were processed for DNA extraction and integration site analysis using ligation-mediated PCR according to a previously described protocol[49].

Analysis of cell-associated HIV-1 RNA

Total cell-associated RNA and DNA was extracted in parallel from the same PBMC sample, using the GenElute RNA/DNA/Protein Purification Plus Kit (Sigma RDP300) according to the manufacturer's protocol. RNA was reverse-transcribed into cDNA using a polyadenylation-RT reaction[56] to efficiently detect HIV-1 RNA transcripts, followed by ddPCR-based amplification with primers and probes that span the HIV-1 trans-activation response (tar) region, as described previously[56]. Simultaneously, cell-associated DNA was subjected to ddPCR-based amplification of the *RPP30* gene to determine cell counts in PBMC samples, using probes and primers described previously[57]. Cell-associated HIV-1 RNA copies per million PBMCs were normalized to the corresponding number of intact proviruses per million PBMCs (determined by FLIP-seq).

Statistics

Data are shown as pie charts, bar charts, scatter plots with individual values or heat maps. Differences were tested for statistical significance using Mann–Whitney *U*-tests (two-tailed), Fisher's exact tests (two-tailed) or χ^2 tests (two-tailed), as appropriate. $P < 0.05$ was considered significant, FDR correction was performed using the Benjamini–Hochberg method[58]. Analyses were performed using Prism (GraphPad Software), SPICE[59] and R (R Foundation for Statistical Computing)[60].

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 Boards of MGH, BWH and UCSF.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

RNA-seq and ATAC-seq data have been deposited in the NCBI GEO (accession number GSE144334). Owing to study participant confidentiality concerns, full-length viral sequencing data cannot be publicly released, but will be made available to investigators upon reasonable request and after signing a coded tissue agreement. The Los Alamos HIV Sequence Database Hypermut 2.0 and the Los Alamos HIV Immunology Database 2.0 are available at https://www.hiv.lanl.gov/content/index. The iMethyl database is available at http://imethyl.iwate-megabank.org. ROADMAP epigenomic data are available at http://www.roadmapepigenomics.org.

41. Burtt, N. P. Whole-genome amplification using Φ29 DNA polymerase. *Cold Spring Harb. Protoc.* **2011**, pdb.prot5552 (2011).
42. Lee, G. Q. et al. HIV-1 DNA sequence diversity and evolution during acute subtype C infection. *Nat. Commun.* **10**, 2737 (2019).
43. Hiener, B. et al. Identification of genetically intact HIV-1 proviruses in specific $CD4^+$ T cells from effectively treated participants. *Cell Rep.* **21**, 813–822 (2017).
44. Pinzone, M. R. et al. Longitudinal HIV sequencing reveals reservoir expression leading to decay which is obscured by clonal expansion. *Nat. Commun.* **10**, 728 (2019).
45. Rose, P. P. & Korber, B. T. Detecting hypermutations in viral sequences with an emphasis on G→A hypermutation. *Bioinformatics* **16**, 400–401 (2000).
46. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
47. Siepel, A. C., Halpern, A. L., Macken, C. & Korber, B. T. A computer program designed to screen rapidly for HIV type 1 intersubtype recombinant sequences. *AIDS Res. Hum. Retroviruses* **11**, 1413–1416 (1995).
48. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
49. Serrao, E., Cherepanov, P. & Engelman, A. N. Amplification, next-generation sequencing, and genomic DNA mapping of retroviral integration sites. *J. Vis. Exp.* **109**, e53840 (2016).
50. Trombetta, J. J. et al. Preparation of single-cell RNA-seq libraries for next generation sequencing. *Curr. Protoc. Mol. Biol.* **107**, 4.22.1–4.22.17 (2014).
51. Corces, M. R. et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
52. Corces, M. R. et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017).
53. Laird, G. M., Rosenbloom, D. I., Lai, J., Siliciano, R. F. & Siliciano, J. D. Measuring the frequency of latent HIV-1 in resting $CD4^+$ T cells using a limiting dilution coculture assay. *Methods Mol. Biol.* **1354**, 239–253 (2016).

54. Bruner, K. M. et al. A quantitative approach for measuring the reservoir of latent HIV-1 proviruses. *Nature* **566**, 120–125 (2019).
55. Chen, H. et al. CD4$^+$ T cells from elite controllers resist HIV-1 infection by selective upregulation of p21. *J. Clin. Invest.* **121**, 1549–1560 (2011).
56. Yukl, S. A. et al. HIV latency in isolated patient CD4$^+$ T cells may be due to blocks in HIV transcriptional elongation, completion, and splicing. *Sci. Transl. Med.* **10**, eaap9927 (2018).
57. Kuo, H. H. et al. Anti-apoptotic protein BIRC5 maintains survival of HIV-1-infected CD4$^+$ T cells. *Immunity* **48**, 1183–1194 (2018).
58. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
59. Roederer, M., Nozzi, J. L. & Nason, M. C. SPICE: exploration and analysis of post-cytometric complex multivariate datasets. *Cytometry* **79A**, 167–174 (2011).
60. R Core Team. *R: A Language and Environment for Statistical Computing*. http://www.R-project.org/ (R Foundation for Statistical Computing, 2019).
61. Robson, M. I. et al. Constrained release of lamina-associated enhancers and genes from the nuclear envelope during T-cell activation facilitates their association in chromosome compartments. *Genome Res.* **27**, 1126–1138 (2017).

Acknowledgements X.G.Y. is supported by NIH grants HL134539, AI116228, AI078799, DA047034 and the Bill and Melinda Gates Foundation (INV-002703). M.L. is supported by NIH grants AI098487, AI135940, AI114235, AI117841, AI120008 and DK120387. M.L. and X.G.Y. are Associated Members of the BEAT-HIV Martin Delaney Collaboratory (UM1 AI126620). A.N.E. is supported by NIH grant AI052014. Support was also provided by the Harvard University (HU) and University of California at San Francisco (UCSF)/Gladstone Institute for HIV Cure Research Centers for AIDS Research (P30 AI060354 and P30 AI027763, respectively), which are supported by the following institutes and centers that are co-funded by and associated with the US National Institutes of Health: NIAID, NCI, NICHD, NHLBI, NIDA, NIMH, NIA, FIC and OAR, and by HU CFAR Developmental Awards (S.H.). We thank the MGH DNA core facility. R.F.S. and J.M.S. are supported by the NIH Martin Delaney I4C (UM1 AI126603), BEAT-HIV (UM1 AI126620) and the Delaney AIDS Research Enterprise (DARE; UM1 AI126611) Collaboratories and by the Howard Hughes Medical Institute and the Bill and Melinda Gates Foundation (OPP1115715). Additional support for the SCOPE cohort was provided by DARE (AI096109 and AI127966) and the amfAR Institute for HIV Cure Research (amfAR 109301). G.M.L. is supported by NSF grant 1738428 and NIH grant R44AI124996. The International HIV Controller Cohort is supported by the Bill and Melinda Gates Foundation (OPP1066973), the Ragon Institute of MGH, MIT and Harvard, the NIH (R37 AI067073 to B.D.W.) and the Mark and Lisa Schwartz Family Foundation. This project has been funded in whole or in part with federal funds from the Frederick National Laboratory for Cancer Research, under contract no. HHSN261200800001E. This research was supported in part by the Intramural Research Program of the NIH, Frederick National Lab, Center for Cancer Research and Intramural Programs of NIDCR, NIH. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government.

Author contributions C.J., X.L., K.B.E., J.M.C., B.R., K.C. and J.E.B. performed whole-genome amplification and HIV-1 sequencing; C.J. and K.B.E. analysed integration sites in cells infected in vivo; E.S. and A.N.E. analysed integration sites in cells infected in vitro; S.H. and X.S. performed RNA-seq; X.S. carried out ATAC-seq; C.J., X.L., K.B.E., J.E.B. and M.O. analysed HIV-1 RNA transcripts; C.G. performed bioinformatics analysis; J.M.C., S.M.Y.C., L.N.B., S.E.S., J.A.V., R.F.S. and J.M.S. carried out qVOAs; M.J.P., R.H., M.S., J.M., P.D.B., T.W.C., S.G.D. and B.D.W. contributed PBMCs and tissue samples; M.C. carried out HLA class I typing; G.M.L., R.F.S. and J.M.S. performed IPDA; C.J., X.L., C.G., M.L. and X.G.Y. carried out data interpretation, analysis and presentation; C.J., X.L., C.G., M.L. and X.G.Y. prepared and wrote the manuscript; C.G., X.S., K.B.E., R.H., A.N.E., M.C., S.G.D., R.F.S. and B.D.W. critically reviewed and edited the manuscript; M.L. and X.G.Y. conceived the research idea and concept and supervised the study.

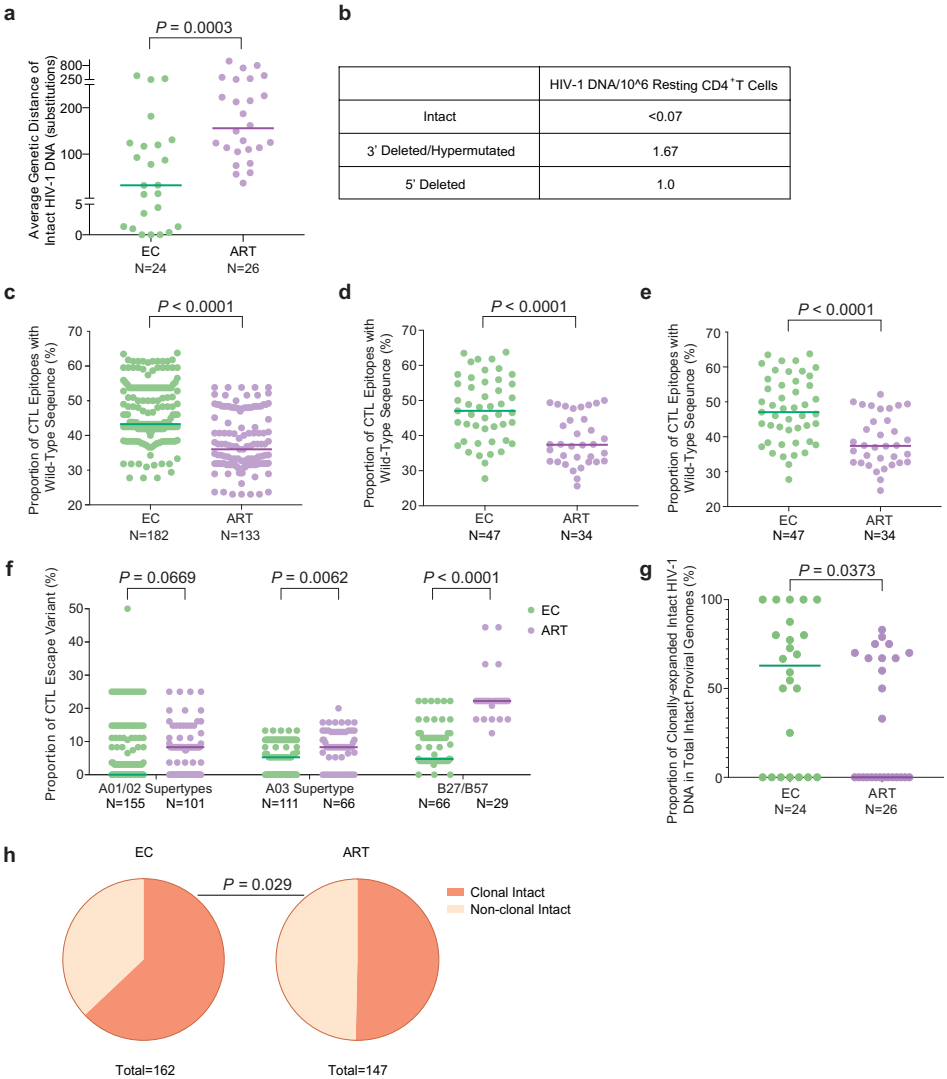
Competing interests A.N.E. has received fees from ViiV Healthcare within the past year for work unrelated to this project. All other authors declare no competing interests.

Additional information

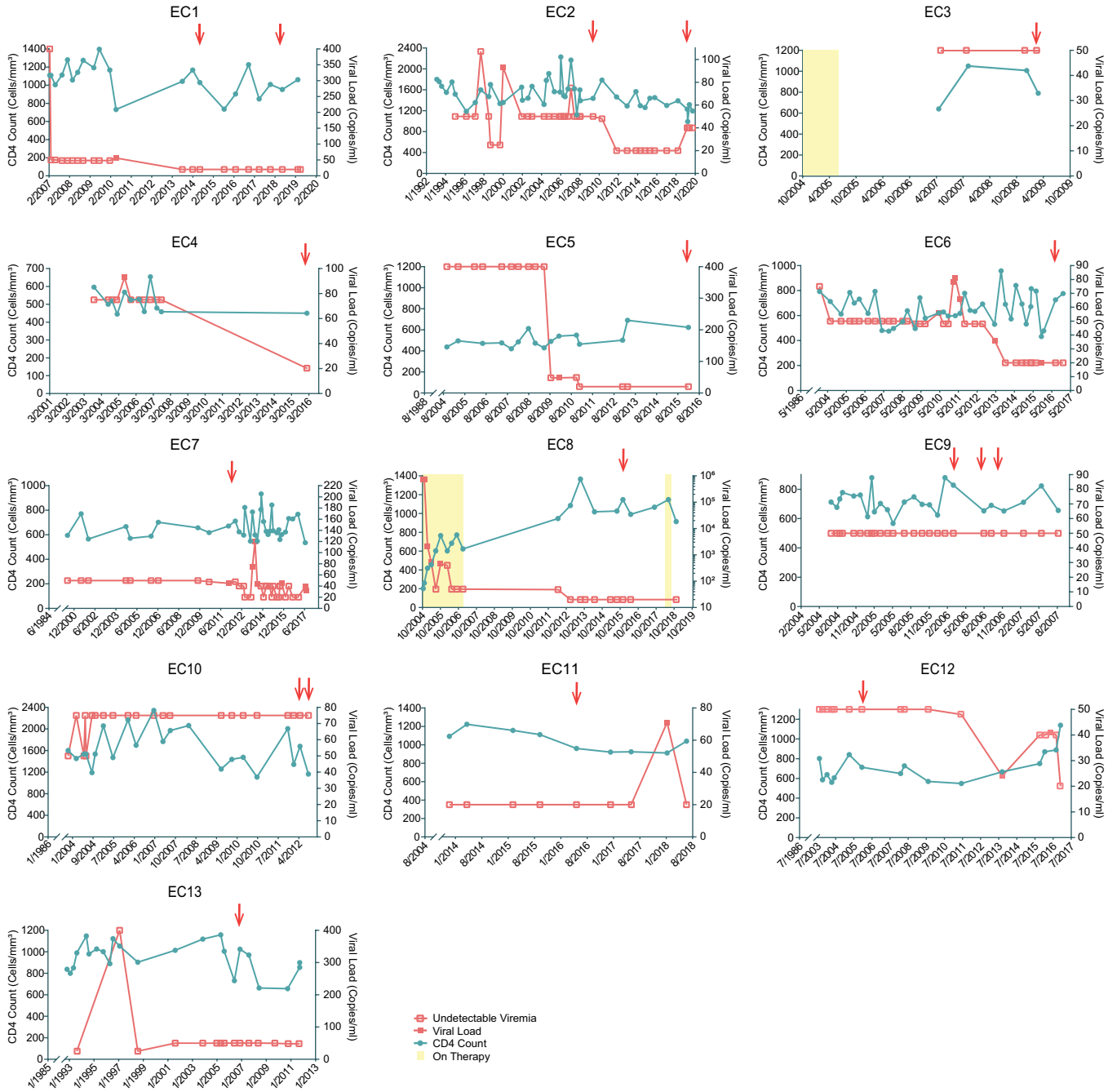
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2651-8.

Correspondence and requests for materials should be addressed to X.G.Y.

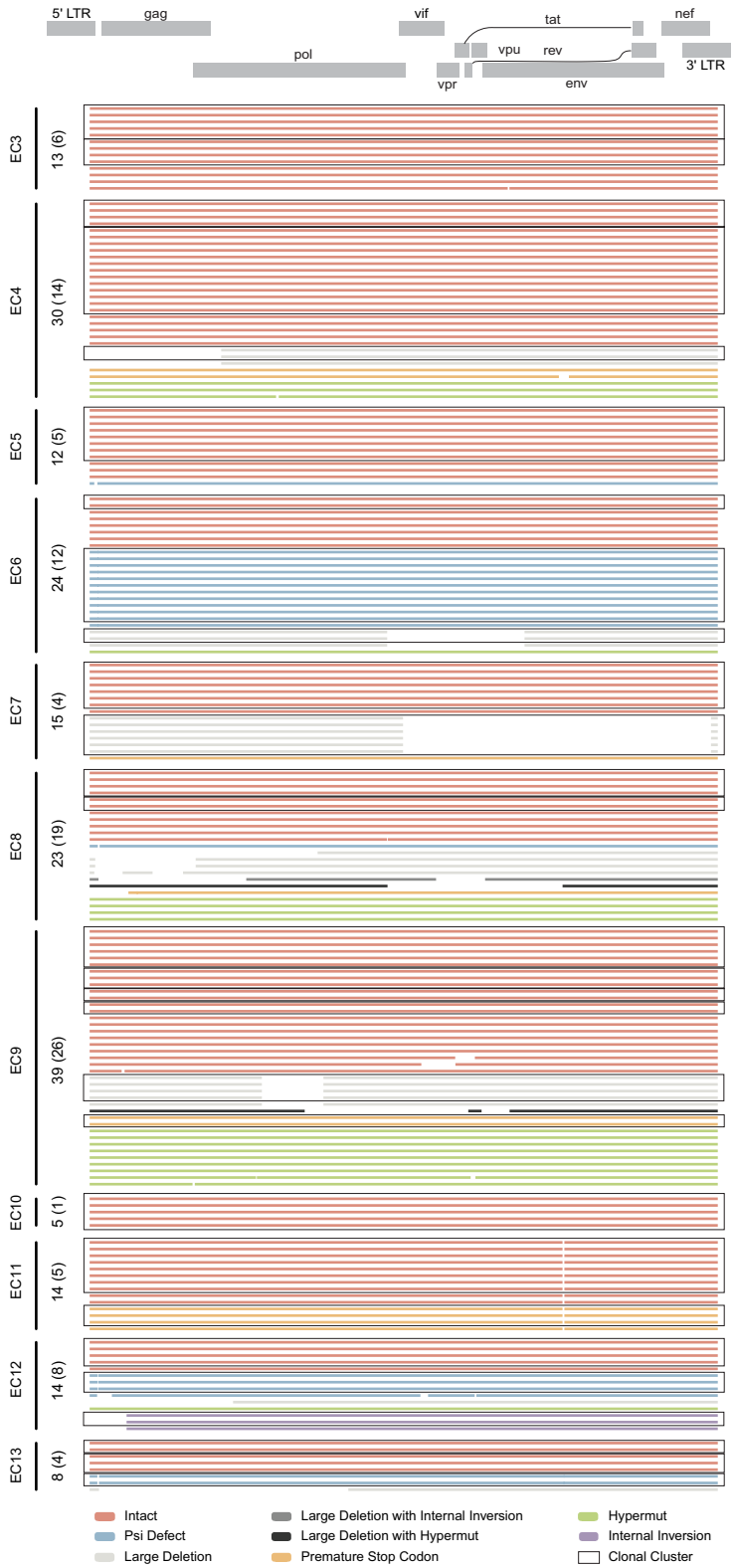
Peer review information *Nature* thanks Nicolas Chomont, Philippe Lemey and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



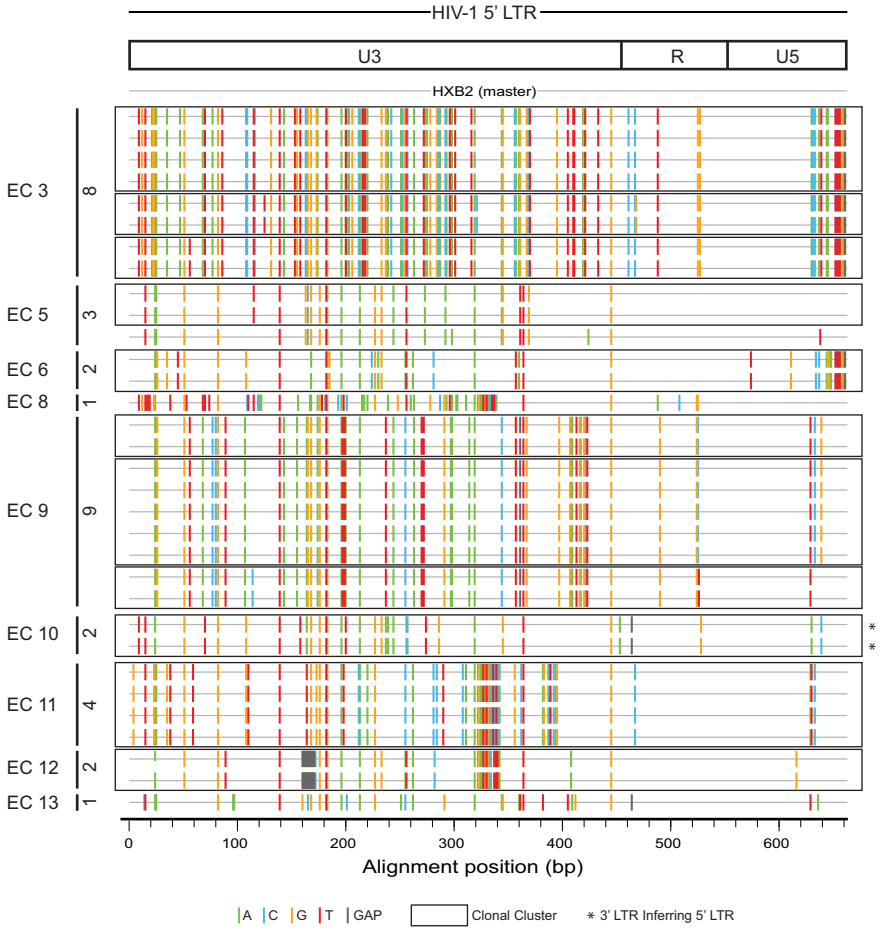
Extended Data Fig. 1 | Viral sequence analysis of intact HIV-1 proviruses from elite controllers. a, Genetic distance (expressed as the average number of base pair substitutions) among all intact near-full-length proviral sequences obtained from each study participant. Clonal sequences were considered to be individual sequences; participants with at least two intact proviruses are included ($n = 175$ intact proviral sequences from 24 elite controllers and $n = 147$ intact proviral sequences from 26 ART-treated individuals). **b**, Frequencies of proviral species (copies per million resting $CD4^+$ T cells) detected by IPDA from EC2. **c**, Proportion of optimal CTL epitopes (restricted by autologous HLA class I isotypes) with wild-type sequences within intact HIV-1 clade B sequences. Each dot represents one intact proviral sequence. $n = 182$ and $n = 133$ HIV-1 clade B intact sequences from 47 elite controllers and 34 ART-treated individuals are included, respectively. Optimal CTL epitopes matching the clade B consensus sequences were considered to be wild-type sequences. Clonal sequences were considered to be individual sequences. **d**, **e**, Average proportions of autologous HLA-class I restricted optimal CTL epitopes with wild-type sequences calculated from intact proviruses in each study participant. Clonal sequences were counted either once (**d**) or as individual sequences (**e**). Each dot represents one study participant. **f**, Proportions of optimal CTL epitopes containing escape variants (restricted by HLA-A01/A02 supertypes, HLA-A03 supertype or HLA-B*27/B*57) within intact proviruses from elite controllers and ART-treated individuals. Each dot represents one intact proviral sequence. Clonal sequences were counted individually. **g**, **h**, Proportion of clonal intact proviruses among all intact proviruses within each study participant (**g**) or within all intact proviruses from elite controllers and ART-treated individuals (**h**). Study participants for whom at least two intact proviruses were detected are included in **g** and **h**. Two-tailed Mann–Whitney *U*-tests were used for data shown in **a**, **c**–**g**; two-sided Fisher's exact test was used for data shown in **h**.



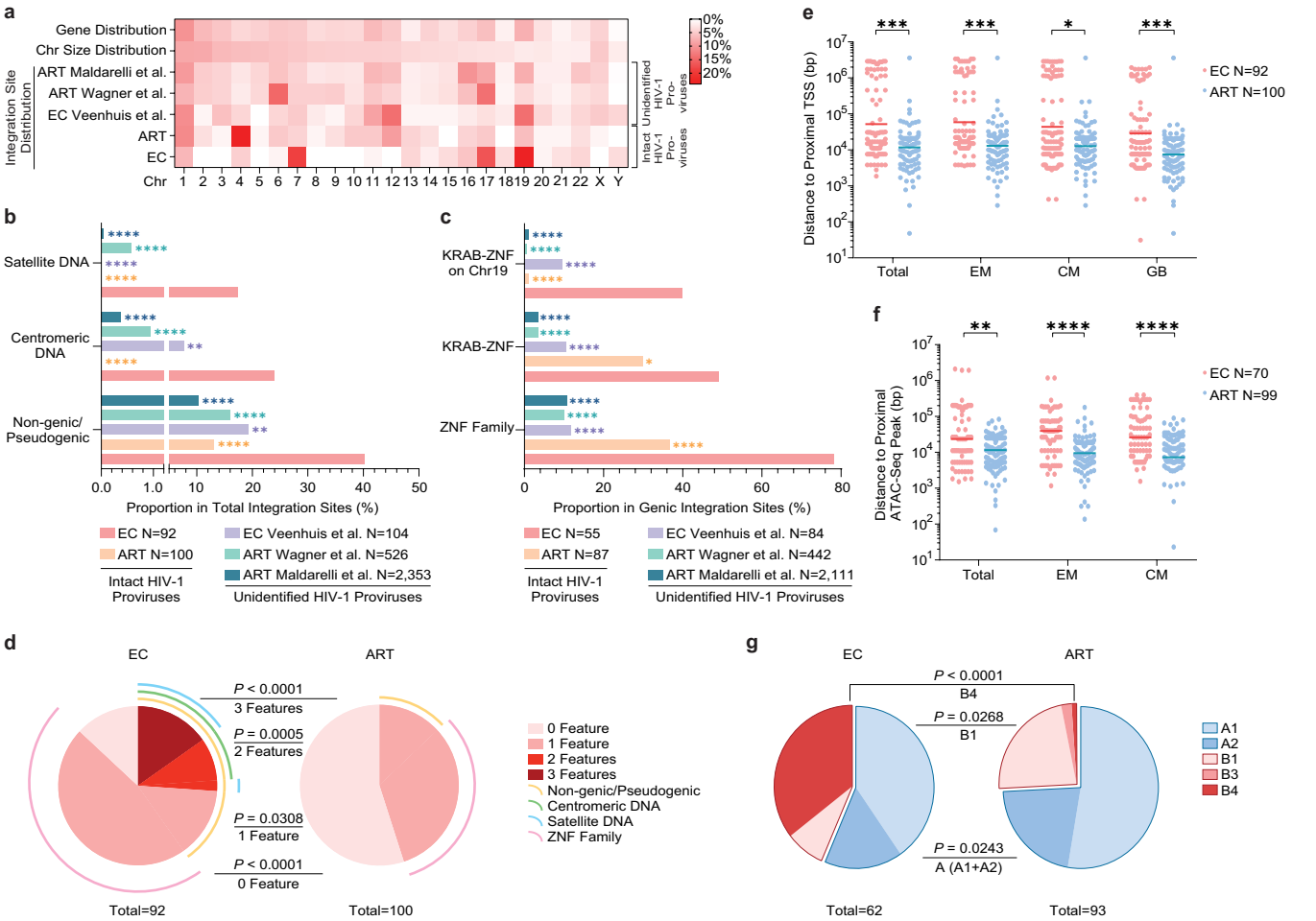
Extended Data Fig. 2 | Longitudinal evolution of CD4⁺ T cell counts and HIV-1 viral loads in EC1–EC13. The recorded diagnosis date of HIV-1 infection for each study participant is shown as the first date on the *x* axis. PBMC sampling time points are indicated by red arrows.



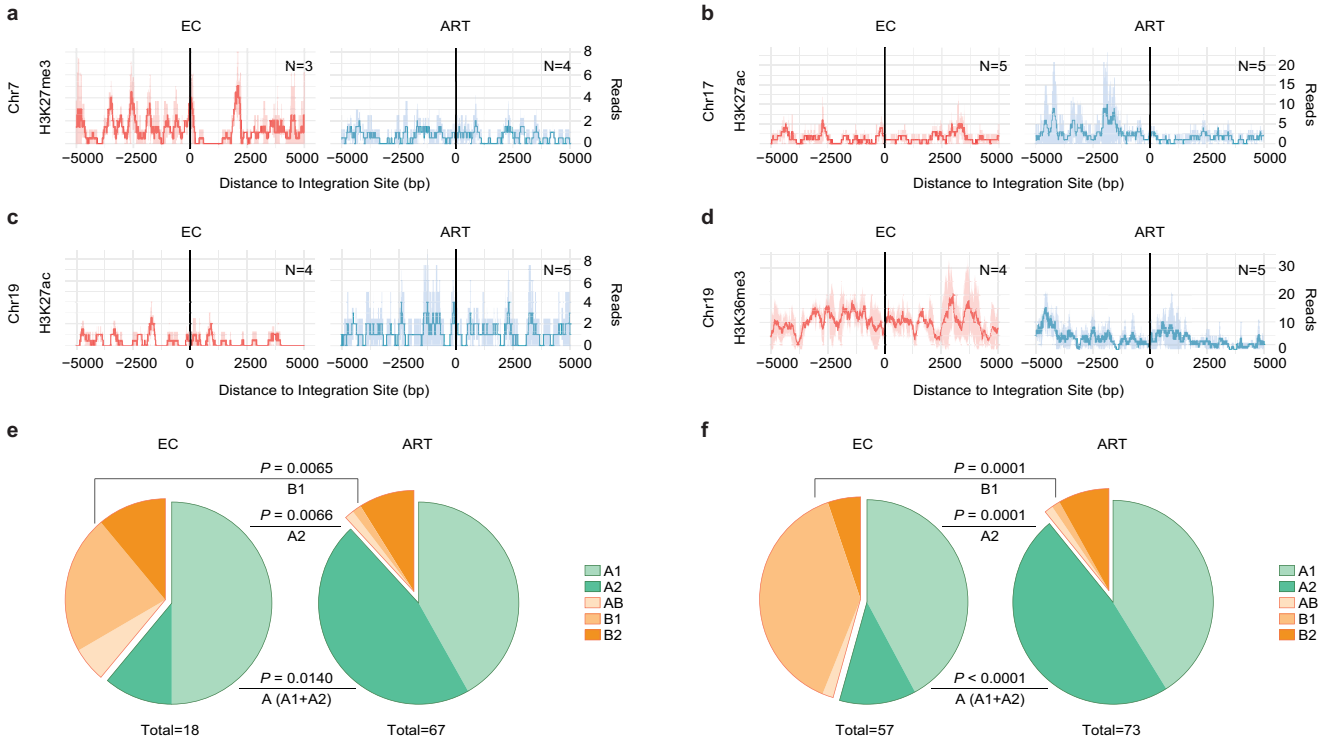
Extended Data Fig. 3 | The structural composition of proviral reservoirs in elite controllers. Virograms reflect the genetic coverage of individual sequences of proviral genomes analysed in EC3–EC13. Numbers of total near-full-length proviral sequences obtained from each individual are shown on the *y* axis; numbers of independent sequences are indicated in brackets. Open boxes indicate clonal clusters.



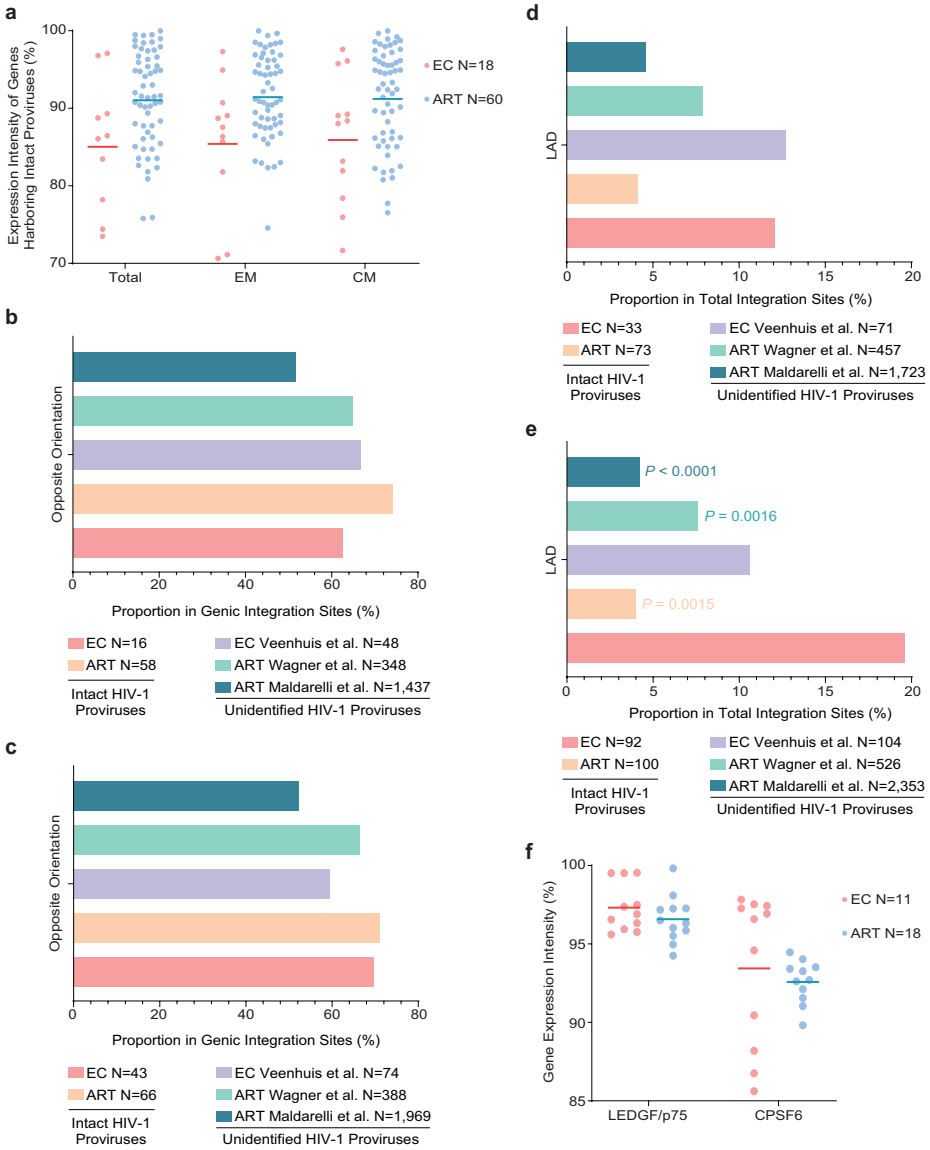
Extended Data Fig. 4 | The variations in HIV-1 DNA sequences in 5′ LTR regions from intact proviruses isolated from the indicated elite controllers, relative to HXB2. Numbers of 5′ LTR sequences of intact proviruses obtained from each individual are shown on the vertical axis. Open boxes indicate clonal clusters.



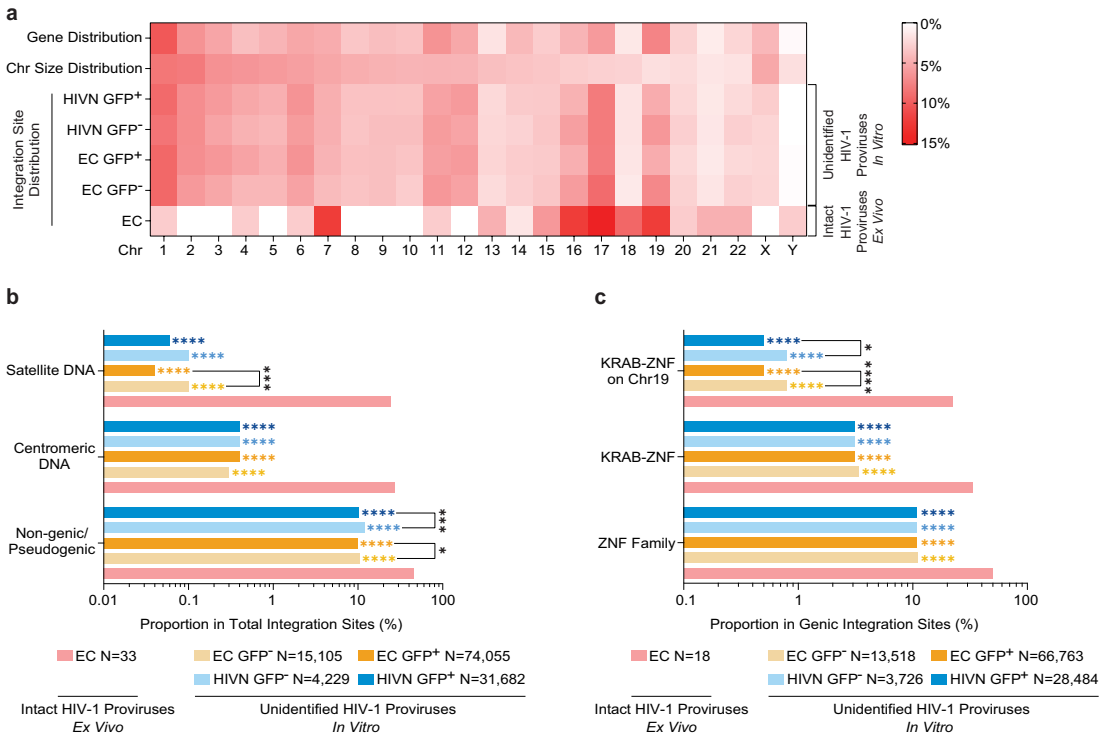
Extended Data Fig. 5 | Features of the chromosomal integration sites of intact proviruses from elite controllers after counting clonal sequences individually. a, Heat map indicating the relative proportion of proviral integration sites of intact proviruses in each chromosome in elite controllers, relative to corresponding data from long-term ART-treated individuals[14]. Proviral integration site data from previous publications[9,15,17] are shown for comparison; integration sites from intact and defective proviruses were not distinguished in these studies. Contributions of each chromosome to the total number of genes (first row) and to the total size of the human genome (second row) are included as references. **b**, **c**, Proportion of near-full-length intact proviruses located in the indicated genomic regions. Data from near-full-length intact proviral sequences in long-term ART-treated individuals are shown as a reference[14]; chromosomal integration sites from unselected (intact and defective) proviral sequences in elite controllers[9] and in ART-treated individuals[15,17] are also shown for comparison. **d**, SPICE diagrams[59] showing the proportion of intact proviruses with the indicated chromosomal integration site features in elite controllers and ART-treated individuals. **e**, **f**, Chromosomal distance between integration sites of intact proviruses and the most proximal transcriptional start sites (determined by RNA-seq) (**e**) or to the most proximal ATAC-seq peak (**f**) in autologous total, central memory and effector memory $CD4^+$ T cells and in the Genome Browser (GB). Horizontal lines show the geometric mean. **g**, Proportions of proviral sequences located in structural compartments A and B, as determined using previously published Hi-C-seq data[29]. Chromosomal integration regions not covered in the previous study[29] were excluded from the analysis. **f**, **g**, Sequences in genomic regions included in the blacklist for functional genomics analysis identified by the ENCODE and modENCODE consortia[28] were excluded owing to the absence of reliable ATAC-seq and Hi-C-seq reads in such repetitive regions. **a**–**g**, All members of clonal clusters were included as individual sequences. $^{****}P < 0.0001$, $^{***}P < 0.001$, $^{**}P < 0.01$, $^{*}P < 0.05$; FDR-adjusted two-sided Fisher's exact tests were used for data shown in **b** and **c**; two-sided Fisher's exact tests were used for data shown in **d** and **g**; FDR-adjusted two-tailed Mann–Whitney *U*-tests were used for data shown in **e** and **f**; all comparisons were made between elite controllers and reference groups.



Extended Data Fig. 6 | Epigenetic features of the chromosomal integration sites of intact proviruses from elite controllers. a–**d**, Numbers of DNA-sequencing reads associated with activating (H3K27ac) or repressive (H3K27me3) histone protein modifications in proximity to integration sites from elite controllers and long-term ART-treated individuals; median and confidence intervals (defined by one standard deviation) of ChIP–seq data from primary memory CD4$^+$ T cells included in the ROADMAP repository[26] are shown. Negative distances indicate genomic regions upstream of the HIV-1 5′ LTR host–viral junction; positive distances indicate regions downstream of the 3′ LTR viral–host junction. DNA-sequencing reads associated with H3K36me3, a chromatin mark that is atypically enriched in *KRAB-ZNF* genes on chromosome 19, are also shown[29]. **e**, **f**, Proportions of intact proviral sequences located in structural compartments A and B (and associated sub-compartments) by counting clonal sequences once (**e**) or by counting clonal sequences individually (**f**), as determined based on the alignment of chromosomal integration sites of intact proviruses to Hi-C-seq data from Jurkat cells[30]. Chromosomal integration regions not covered in the Jurkat cell study[30] were excluded from the analysis. Compartment B4 was not assessed in the source data[30] for this analysis. Two-sided Fisher's exact tests were used for statistical comparisons; nominal *P* values are reported. **a**–**f**, Sequences in genomic regions included in the blacklist for functional genomics analysis identified by the ENCODE and modENCODE consortia[28] were excluded owing to the absence of reliable ChIP–seq and Hi-C-seq reads in such repetitive regions.



Extended Data Fig. 7 | Accessory features of chromosomal integration sites of intact proviral sequences from elite controllers. a, Expression of host genes that contain intact proviral sequences in elite controllers and long-term ART-treated individuals, as determined by autologous RNA-seq data in total, central memory and effector memory CD4$^+$ T cells. Gene expression percentiles are indicated. **b**, **c**, Orientation of intact proviruses relative to host genes in elite controllers and long-term ART-treated individuals. All data for genic integration sites are included, except for integration sites in genic regions associated with multiple genes in opposing orientations. Integration site data from previous studies of elite controllers[9] and ART-treated individuals[15,17] are shown for comparative purposes. **d**, **e**, Proportion of intact proviruses from elite controllers and long-term ART-treated individuals in lamina-associated domains, determined using Lamin B1–DNA adenine methyltransferase identification (DamID)[61] for resting Jurkat cells. Integration site data from previous studies of elite controllers[9] and ART-treated individuals[15,17] are shown for comparative purposes. **b**, **d**, Clonal proviruses were counted once. **c**, **e**, Clonal proviruses were counted as individual sequences (FDR-adjusted two-sided Fisher's exact tests). **f**, Expression of *LEDGF* (also known as *PSIP1* or *p75*) and *CPSF6* mRNA in autologous total CD4$^+$ T cells from elite controllers and long-term ART-treated individuals, as determined by RNA-seq. Gene expression percentiles are indicated. **a**, **f**, Horizontal lines show the geometric mean. All comparisons were made between elite controllers and reference groups.



Extended Data Fig. 8 | Features of chromosomal integration sites of in vitro-infected CD4$^+$ T cells from elite controllers and HIV-1-negative study participants. a, Heat map showing the relative proportion of proviral integration sites in sorted GFP$^+$ or GFP$^-$ in vitro-infected CD4$^+$ T cells (determined by ligation-mediated PCR[49]) from elite controllers and HIV-1-negative study participants (HIVNs), relative to proviral integration sites of intact proviruses in each chromosome in elite controllers; integration sites from intact and defective proviruses were not distinguished in in vitro-infection studies. Data from GFP$^+$ (n = 74,055) and GFP$^-$ (n = 15,105) CD4$^+$ T cell populations from elite controllers and from GFP$^+$ (n = 31,682) and GFP$^-$ (n = 4,229) CD4$^+$ T cell populations from HIV-1-negative study participants were included. Contributions of each chromosome to the total number of genes (first row) and to the total size of the human genome (second row) are included as references. **b**, **c**, Proportion of proviral integration sites located in indicated genomic regions (**b**) or defined genes (**c**). Data from near-full-length intact proviral sequences in elite controllers are indicated for reference. ****P < 0.0001, ***P < 0.001, *P < 0.05; FDR-adjusted two-sided Fisher's exact tests or two-tailed χ^2 tests were used as appropriate; P values indicating comparisons made between intact proviruses from elite controllers (determined ex vivo) and each in vitro-infection group are shown in corresponding colours.

Extended Data Table 1 | Demographical and clinical characteristics of all study participants

	Elite Controllers (EC)	ART-treated Participants (ART)
Number of participants	64	41
Age in years*	57 (31 - 75)	55 (34 - 73)
Female (%)	18.75%	21.95%
CD4 counts*	908† (450 - 2,282)	726 (316 - 1,649)
Viral loads	Under limit of detection	Under limit of detection
Number of viral load tests*	18 (3 - 91)	32.5 (4 - 73)
HLA-B*27/B*57 (%)	27.34%‡	8.75%
Time since diagnosis (year)*	17 (1 - 34)	17 (5 - 35)
Recorded duration of undetectable viremia (year)*	9 (1 - 24)	9 (2 - 19)

*Median with range.

†$P = 0.0006$, tested using a two-tailed Mann–Whitney *U*-test.

‡$P = 0.0012$, tested using two-sided Fisher's exact test.

nature research

Corresponding author(s): Xu G. Yu

Last updated by author(s): Jun 18, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Quantify One (version 4.4.1), ChemiDoc MP Image Lab software (BioRad, version 6.0.1), BD FACSDiva software (version 8.0.1), QuantaSoft (version 1.7.4.0917)
Data analysis	Los Alamos HIV Sequence Database Hypermut 2.0 (https://www.hiv.lanl.gov/content/sequence/HYPERMUT/hypermut.html), MEGA (https://www.megasoftware.net, version 7.0.26), MUSCLE (http://www.drive5.com/muscle, version 3.8.1551), Graphpad prism (https://www.graphpad.com/scientific-software/prism/, version 8.2.1), UltraCycler v1.0 (Brian Seed and Huajun Wang from MGH CCIB DNACore, unpublished), R (https://www.r-project.org, version 3.5.3), UCSC Genome Browser (https://genome.ucsc.edu), GENCODE (https://www.gencodegenes.org, version 29), Ensembl (https://ensembl.org, version 86), RepeatMasker (www.repeatmasker.org), RSEM (https://deweylab.github.io/RSEM/, version 1.2.22), STAR (https://github.com/alexdobin/STAR,version 2.5.1b), FastQC (https://www.bioinformatics.babraham.ac.uk, version 0.11.9), Trimmomatic (http://www.usadellab.org, version 0.39), Samtools (http://www.htslib.org/, version 1.3.1), MACS2 (https://pypi.python.org/pypi/MACS2, version 2.1.1.20160309), iMethyl (http://imethyl.iwate-megabank.org), ROADMAP (http://www.roadmapepigenomics.org/), MAFFT (https://mafft.cbrc.jp/alignment/software, version 7), Highlighter (https://www.hiv.lanl.gov/content/sequence/HIGHLIGHT/highlighter_top.html), SPICE (https://niaid.github.io/spice/), Recombinant Identification Program (https://www.hiv.lanl.gov/content/sequence/RIP/RIP.html), FlowJo software (version 10.6), Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml, version 2.2.9), in-house intactness pipeline (https://github.com/BWH-Lichterfeld-Lab/Intactness-Pipeline)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

RNA-Seq and ATAC-Seq data have been deposited in a public repository (NCBI GEO, accession number GSE144334). Due to study participant confidentiality concerns, full-length viral sequencing data cannot be publicly released, but will be made available to investigators upon reasonable request and after signing a coded tissue agreement. The Los Alamos HIV Sequence Database Hypermut 2.0 and the Los Alamos HIV Immunology Database 2.0 are available at www.hiv.lanl.gov. The iMethyl database is available at http://imethyl.iwate-megabank.org. ROADMAP epigenomic data are available at http://www.roadmapepigenomics.org.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	A total of n=64 EC and n=41 ART-treated individuals were analyzed in data described in Figure 1. In Figure 2-4, n=11 EC are described in detail. No computational approach was used to determine these sample sizes, testing was based on availability of more than 50 million PBMC per study participant.
Data exclusions	No data from the described individuals were excluded.
Replication	Viral and integration site sequencing was performed once for each individual proviral sequence. To test the accuracy of our sequencing approach, we repeated sequencing of near full-length HIV-1 DNA from the 8E5 cell line 50 consecutive times, which resulted in 100% identical sequences in all runs.
Randomization	No randomization was performed, because we performed a cross-sectional analysis of study participants enrolled in an observational study.
Blinding	Coded samples from study participants were used throughout the study; laboratory personnel was not blinded with regard to the respective study cohorts, since this was a non-interventional, observational study. All sequencing reactions were performed at a local core facilities; core facility employees were fully blinded.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	CD3 (clone OKT3, BioLegend, catalog 317332) CD4 (clone RPA-T4, BioLegend, catalog 300518) CCR7 (clone G043H7, BioLegend, catalog 353216) CCD45RO (clone UCHL1, BioLegend, catalog 304236) CD3/CD8 bi-specific antibody (NIH AIDS Reagent Program #12277)

Validation	CD3 (clone OKT3, BioLegend, catalog 317332): Reactivity: Human Host Species: Mouse Application: FC - Quality tested Application Notes: The OKT3 monoclonal antibody reacts with an epitope on the epsilon-subunit within the human CD3 complex. Clone OKT3 can block the binding of clones SK7 and UCHT1.4 The OKT3 antibody is able to induce T cell activation. Additional reported applications (for the relevant formats) include: immunohistochemical staining of acetone-fixed frozen sections and activation of T cells. The LEAF™ purified antibody (Endotoxin <0.1 EU/µg, Azide-Free, 0.2 µm filtered) is recommended for functional assays (Cat. No. 317304). For highly sensitive assays, we recommend Ultra-LEAF™ purified antibody (Cat. No. 317326) with a lower endotoxin limit than standard LEAF™ purified antibodies (Endotoxin <0.01 EU/µg). Application References: Schlossman S, et al. Eds. 1995. Leucocyte Typing V. Oxford University Press. New York. Knapp W. 1989. Leucocyte Typing IV. Oxford University Press New York. Barclay N, et al. 1997. The Leucocyte Antigen Facts Book. Academic Press Inc. San Diego. Li B, et al. 2005. Immunology 116:487. CD4 (clone RPA-T4, BioLegend, catalog 300518): Reactivity: Human, Chimpanzee Host Species: Mouse Application: FC - Quality tested Application Notes: The RPA-T4 antibody binds to the D1 domain of CD4 (CDR1 and CDR3 epitopes) and can block HIV gp120 binding and inhibit syncytia formation. Additional reported applications (for the relevant formats) include: immunohistochemistry of acetone-fixed frozen sections3,4,5, and blocking of T cell activation1,2. This clone was tested in-house and does not work on formalin fixed paraffin-embedded (FFPE) tissue. The LEAF™ purified antibody (Endotoxin <0.1 EU/µg, Azide-Free, 0.2 µm filtered) is recommended for functional assays (Cat. No. 300516). Application References: Knapp W, et al. 1989. Leucocyte Typing IV. Oxford University Press. New York. (Activ) Moir S, et al. 1999. J. Virol. 73:7972. (Activ) Deng MC, et al. 1995. Circulation 91:1647. (IHC) Friedman T, et al. 1999. J. Immunol. 162:5256. (IHC) Mack CL, et al. 2004. Pediatr. Res. 56:79. (IHC) Lan RY, et al. 2006. Hepatology 43:729. Zenaro E, et al. 2009. J. Leukoc. Biol. 86:1393. (FC) PubMed Yoshino N, et al. 2000. Exp. Anim. (Tokyo) 49:97. (FC) Stoeckius M, et al. 2017. Nat. Methods. 14:865. (PG) CCR7 (clone G043H7, BioLegend, catalog 353216): Reactivity: Human, African Green, Baboon, Cynomolgus, Rhesus Host Species: Mouse Application: FC - Quality tested CCD45RO (clone UCHL1, BioLegend, catalog 304236): Reactivity: Human, Chimpanzee, Cynomolgus, Common Marmoset Host Species: Mouse Application: FC - Quality tested Application Notes: The UCHL1 antibody is commonly used in combination with antibodies against CD45RA to discern memory and naïve T cells. Additional reported applications (for the relevant formats) include: immunohistochemical staining of acetone-fixed frozen tissue sections5 and formalin-fixed paraffin-embedded tissue sections4, Western blotting2, and immunoprecipitation3. Application References: Knapp W, et al. Eds. 1989. Leucocyte Typing IV. Oxford University Press. New York. (FC) Ishii T, et al. 2001. P. Natl. Acad. Sci. USA 98:12138. (WB) Ponsford M, et al. 2001. Clin. Exp. Immunol. 124:315. (IP) Yamada M, et al. 1996. Stroke 27:1155. (IHC) Sakkas LI, et al. 1998. Clin. Diagn. Lab. Immunol. 5:430. (IHC) CD3/CD8 bi-specific antibody (NIH AIDS Reagent Program #12277) The bi-specific CD3/8 (CD3.8) monoclonal antibody was generated by fusing the anti-CD3 mAb producing hybridoma (12F6) with the anti-CD8 mAb producing hymbridoma (OKT8). The resulting anti-CD3/8 antibody, when added to long term peripheral blood co-cultures results in the potent elimination of CD8+ T cells. The remaining cells are highly activated and serve as a reliable source of purified activated cells of interest.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Please see Extended Data Table 1.
Recruitment	EC and ART-treated individuals were recruited based on referral by HIV clinicians and infectious disease physicians. The enrollment protocols allowed recruited of men and women >18 years old, of any race or ethnicity.
Ethics oversight	The Partners Human Research Committee approved all sample collection at MGH and BWH; the IRB of UCSF supervised sample collection at UCSF.

Note that full information on the approval of the study protocol must also be provided in the manuscript.