Droplet-microfluidics-assisted sequencing of HIV proviruses and their integration sites in cells from people on antiretroviral therapy

Chen Sun, Leqian Liu, Liliana Pérez, Xiangpeng Li, Yifan Liu, Peng Xu, Eli A. Boritz, James I. Mullins and Adam R. Abate

The human immunodeficiency virus (HIV) integrates its genome into that of infected cells and may enter an inactive state of reversible latency that cannot be targeted using antiretroviral therapy. Sequencing such a provirus and the adjacent host junctions in individual cells may elucidate the mechanisms of the persistence of infected cells, but this is difficult owing to the 150-million-fold higher amount of background human DNA. Here we show that full-length proviruses connected to their contiguous HIV–host DNA junctions can be assembled via a high-throughput microfluidic assay where droplet-based whole-genome amplification of HIV DNA in its native context is followed by a polymerase chain reaction (PCR) to tag droplets containing proviruses for sequencing. We assayed infected cells from people with HIV receiving suppressive antiretroviral therapy, resulting in the detection and sequencing of paired proviral genomes and integration sites, 90% of which were not recovered by commonly used nested-PCR methods. The sequencing of individual proviral genomes with their integration sites could improve the genetic analysis of persistent HIV-infected cell reservoirs.

Nearly 37 million people are infected with human immunodeficiency virus (HIV) worldwide and, despite huge research and clinical investments, these infections remain incurable. A central obstacle to curing HIV lies in its mechanism of infection, because HIV integrates its genome into the genome of infected cells. Although many infected cells actively express virus genes, some infected CD4+ T cells enter a state of reversibly non-productive latency. Although antiretroviral therapy (ART) can suppress virus replication to undetectable levels and prevent progression to acquired immunodeficiency syndrome (AIDS), it lacks activity against latent cellular reservoirs. Therefore, ART must be continued indefinitely to prevent virus rebound and recrudescence of disease progression. Lifelong ART is costly, has the potential for toxicity, can be difficult to access and fails to address HIV-associated stigma. Thus, eliminating or suppressing HIV-infected cellular reservoirs to achieve a functional cure is an important goal of HIV research.

To elucidate how infected cellular reservoirs persist, paired analysis of integration sites and full-length provirus sequences is important. Identifying proviruses with matching integration sites has demonstrated that much of the latent reservoir consists of expanded cellular clones. In ART-treated people living with HIV, lethally mutated proviruses far outnumber the genetically intact ones capable of encoding replication-competent viruses, the latter of which are difficult to identify and assess. Sequencing the entire provirus is thus critical to determining whether it could give rise to rebound viremia after ART interruption. In addition, host genetic context may determine the expression level of a given provirus, thus influencing the stability of the host cell. Importantly, understanding the relationships among provirus sequence integrity, clonality and integration site effects requires that these multiple measurements be made for enough proviruses to faithfully represent complex cellular reservoirs in vivo.

Simultaneous assessment of genomes and integration sites has been hindered by the challenge of amplifying these components from individual DNA molecules. Current strategies distribute replicate aliquots of thousands of cell DNA-equivalents in microtitre plate wells at the limiting dilution of the HIV proviruses and then perform multiple displacement amplification (MDA). Wells containing proviruses are detected by subgenomic polymerase chain reaction (PCR), and additional MDA product is amplified again to recover near-full-length provirus genomes and products spanning virus–host integration sites, all of which are then sequenced. However, because the target provirus exists as a single copy within thousands of human genome equivalents, many reactions are required. In addition, because of the large background of human DNA during MDA and the complexity of the PCRs to amplify the HIV genome and virus–host junctions, the reactions are prone to artefact generation, yielding products that contain spurious deletions, inversions and HIV–human junctions that can confound analyses. Moreover, because multiple primer sets are required to amplify the genetically diverse HIV genome, the approach is biased towards variants that best match the primers and the regions targeted. Most importantly, such methods are poorly suited to the even more extreme rarity of infected cells during ART, requiring massive expense in terms of time and resources to yield sequences.
for just a handful of proviruses. Consequently, although performing such analyses on many patients could elucidate important information about the cellular reservoirs of HIV, doing so with current technologies is impractical. To enable comprehensive HIV reservoir genomic characterization, a method that reliably and cost-effectively isolates and sequences rare proviruses ex vivo would be beneficial.

We describe an approach to characterize proviruses and their cellular genomic context within HIV reservoirs—simultaneous integration site and provirus sequencing (SIP-seq). SIP-seq uses whole-genome amplification in microfluidic droplets to amplify the HIV genome in its native context, and TaqMan PCR to tag droplets containing proviruses for sequencing. The result is a technology providing the full-length virus genome connected to its host-cell junctions in a single contiguous assembly. The speed and efficient reagent usage of droplet microfluidics allow recovery of single provirus genomes in a 150-million-fold higher background of DNA fragments. Using SIP-seq, we comprehensively profile the provirus population in multiple ART-treated people to expand our understanding of the latent HIV reservoir. Although we focus on HIV, our approach is applicable to viruses that insert into their host's genome and thus provides a general technology for characterizing the genetics of diverse infections.

**Results**

Under effective ART, HIV replication can be suppressed to undetectable levels in the blood, but persist in rare infected cells in a reversibly latent form. Accurate quantitation of replication-competent proviruses is challenging, but recent estimates suggest values as low as 1 in 100,000 or more CD4+ T cells22. However, if therapy is stopped, viraemia almost always rebounds to pre-therapy levels. CD4+ T cells represent a critical reservoir for HIV1 and thus are the cells we interrogate for HIV persistence (Fig. 1a). The goal of our method is to recover, from a population of millions of CD4+ T cells, all DNA fragments containing HIV genomes and to individually sequence these molecules and their immediately adjacent host junctions (Fig. 2a). To isolate integrated HIV genomes, large DNA fragments are extracted from negatively selected or cultured CD4+ T cells and encapsulated in microdroplets30. Each genomic fragment is then non-specifically amplified via MDA23, yielding enough product of each single provirus for sequencing (Fig. 1b, left). Next, multiplexed TaqMan PCR is used to identify, and microfluidic sorting is used to isolate, individual droplets containing HIV proviruses (Fig. 1b, right). Each positively sorted droplet is barcoded and sequenced. Finally, reads from each droplet are mapped to an HIV reference genome and chimaeras having host and virus sequences are detected to identify integration sites (Fig. 1c).

Individual provirus encapsulation is necessary to obtain single provirus data. Our approach minimizes multiple virus encapsulation by partitioning cellular DNA to the level of one provirus per >10,000 drops, yielding a doublet rate of below 0.0001%. By contrast, brute-force dilution approaches typically target one provirus per three wells, yielding an ~4.5% doublet rate15,16,24. Thus, by leveraging the small volumes of droplet microfluidics, SIP-seq decreases reagent use by 100-fold while also reducing doublets. Moreover, a unique and valuable property of performing the reactions in droplets is that this allows sequencing of picograms of DNA, which allows direct sequencing of the whole-genome amplification products without the need for additional complex multi-primer amplification, yielding large, gapless contiguous assemblies that encapsulate the entire provirus genome connected to the integration site. This provides unambiguous information about virus genome completeness and integration sites that is crucial to assessing the cellular reservoir.

**Provirus detection with SIP-seq.** Random DNA cleavage and common deletions may result in partial genomes within a given droplet. To increase the specificity for full-length proviruses, we employ a dually specific multiplexed TaqMan PCR targeting conserved regions of HIV pol and env spaced >5 kbp apart (Fig. 2a). Dual positive droplets in a representative HIV-infected cell sample were present at ~1 in 13,000, and each recovered droplet yielded ~3 pg of DNA. This amount of DNA was sufficient for sequencing. To confirm the recovery of HIV proviruses, a small portion of the sorted DNA from each droplet was subjected to a TaqMan reaction targeting a different locus of the HIV genome (Supplementary Fig. 1).
Microfluidics of SIP-seq. The microfluidics of SIP-seq consist of three devices: a droplet encapsulator, a merger and a sorter. The encapsulator loads human genomic DNA fragments with MDA reagents in ~20-μm droplets (Fig. 2b(i)). It runs at ~20 kHz, allowing encapsulation of ~10 billion DNA fragments ~75 kbp in length in ~20 million separate droplets over ~15 min. Each droplet thus contains ~500 distinct fragments. Encapsulating the sample into more droplets will deliver fewer co-encapsulated off-target molecules per sorted positive. This will probably reduce the number of reads required to cover the virus genome, but increase the number of droplets that must be sorted. The droplets are incubated at 30 °C to enable non-specific amplification before combining with TaqMan PCR reagents in the merger device. This device merges each MDA droplet with an ~40-μm droplet containing TaqMan reagents, running at ~2.5 kHz, and thus takes a few hours to process all droplets (Fig. 2b(ii)).
Articles

NATuRE BIoMEDICAL enGineerinG

Analysis of HIV in CD4+ T cells from an ART-treated person after in vitro expansion. We performed in vitro infected cell expansion (ICE) to allow rare infected CD4+ T cells to be cultured at limiting dilution. This reduces infected cell rarity and enables downstream studies. Moreover, ICE cultures provide an excellent test of SIP-seq, because they contain proviruses from HIV-positive people in their original host genome contexts. The cells are prepared by plating resting CD4+ T cells from an ART-treated person at one infected cell per approximately five wells (100–3,000 total cells per well), followed by stimulation and in vitro culture to allow proliferation (Fig. 3a). We extracted and analysed DNA from several wells of an ICE culture plate using all three methods (shotgun, nested PCR and SIP-seq using TaqMan probes targeting LTR and tat (Fig. 3b) or one probe targeting pol (Fig. 3c)). As expected, due to the rarity of infected cells even with ICE enrichment, shotgun sequencing recovered little of the HIV genomes (Fig. 3b, top row). By contrast, both SIP-seq and nested PCR yielded excellent coverage of the virus genomes (Fig. 3b, rows 2 and 3), with SIP-seq providing the insertion sites (Fig. 3b, row 3). Notably, a second well containing single virus genomes and integration sites. Moreover, the detection rates of the cell lines matched the known input ratio, demonstrating no bias in recovery.

Validation of SIP-seq on HIV cell lines. To validate SIP-seq we used J-Lat 5A8 cells, a Jurkat-derived cell line containing full-length HIV with a frameshift in env and green fluorescent protein (GFP) in place of nef. As a comparator, we used conventional nested PCR to amplify the ~9-kbp near-full-length HIV genome. The SNPs detected by SIP-seq agreed with those detected by nested PCR with 0.02% mismatch (probably due to MDA or PCR error; Supplementary Fig. 2). We also extracted virus–human chimaeric reads from the SIP-seq data and confirmed the known insertion site

for this cell line. This illustrates that SIP-seq can detect HIV SNPs and insertion sites.

To further assess the ability of SIP-seq to recover complete viral and flanking sequences, we performed a two-cell population experiment on DNA from J-Lat lines mixed at an equal ratio (Fig. 2c, left), with analysis carried out using conventional shotgun and PCR methods, and SIP-seq (Fig. 2c, middle and right). These cell lines (J-Lat 15.4 and 5A8) have identical virus genomes but different integration sites (Fig. 2c, second row). By contrast, SIP-seq provided a connected contig comprising the full-length proviruses and insertion sites (Fig. 2c, third and fourth rows). As expected, negative droplets yielded no HIV provirus, and single positive droplets (pol+/env+ or pol−/env−) recovered partial HIV genomes and integration sites (Supplementary Fig. 3). The peaks at pol and env derive from the abundant TaqMan PCR amplicons used for droplet detection. Reads from two representative drops mapped to either of the two cell lines with minimal cross-contamination. To illustrate the multiplexing ability of the approach for targeting distinct regions of the HIV genome, we also tested a different TaqMan set targeting long terminal repeats (LTRs) and tat and were able to recover HIV proviruses (Supplementary Fig. 4). In total, we sequenced 27 DNA fragments containing single virus genomes and integration sites. Moreover, the detection rates of the cell lines matched the known input ratio, demonstrating no bias in recovery.

Genomic landscape of HIV infection in ART-treated individuals. We next characterized the genomic landscape of HIV proviruses isolated directly from the CD4+ T cells of infected persons (Fig. 4 and Supplementary Table 1). We added TaqMan sets targeting pol and env to the detection PCR to demonstrate multiplexing (Fig. 4a), but sorted based on pol solely to capture the widest swath of proviruses.

Fig. 3 | SIP-seq of HIV in ART-treated participant CD4+ T cells after clonal expansion. a, ICE was prepared by seeding participant CD4+ T cells to less than one infected cell per five wells and culturing for clonal expansion. Two ICE clones were separately processed for sequencing with different technologies. b, SIP-seq detected an intact HIV genome integrated into a human genome. c, SIP-seq identified an integrated HIV genome containing an inverted sequence and a large deletion including 3’ LTR. Nested PCRs from unprocessed genomic DNA using primers designed on the basis of the SIP-seq results confirmed the HIV genomes (bottom panels of b and c).
Overall, we recovered ~10% of the pol-containing HIV proviruses estimated to be in the sample by bulk quantitative PCR (qPCR; Supplementary Table 2). Individuals 1 and 2 were receiving effective ART with virus RNA loads in blood plasma below 20 copies ml⁻¹ at the time of study. In these two individuals we recovered 29 provirus genomes and integration sites (Fig. 4a,b). None of these proviruses were intact. Of the recovered genomes, 14 yielded integration sites, with four having junctions on the 3′ and 5′ ends. Two had intact gag and pol and were integrated into host genes STAT5B and HIVEP1 of participant 1 in the opposite orientation relative to the gene.
We also observed higher numbers of insertions in genic over intergenic (12 versus 2) regions, with no preference in orientation relative to host genes (7 the same versus 7 opposite) (Fig. 4c(i)). We verified representative mutations using specific primers based on SIP-seq data (Supplementary Fig. 6), and representative integrations using customized integrant-specific PCRs, including several integrations missing LTRs (Supplementary Fig. 7). Although these findings are intriguing, we cannot exclude the possibility they are the result of an amplification artefact, because no further unprocessed sample from this individual was available for validation. Thus, these unique genomic structures should be accepted with reservation, as only some have been confirmed (Supplementary Table 3). The P values of the associated deletions in several representative sequences are shown in Supplementary Fig. 8. Sixteen proviruses had large deletions that rule out replication-competency, although not necessarily excluding virus protein expression and the resulting cellular activation and immune recognition. Proliferation of latently infected cells is thought to play an important role in HIV persistence. Our analysis reveals three clonal lineages (Fig. 4c(ii)) and several proviruses that matched sequences obtained from previous studies of the same individuals (Supplementary Fig. 9). Our results thus support the clonal origin of these lineages and provide additional integration site information. Individual sequences generated for a clone are identical in the regions for which we obtain coverage. However, they exhibit coverage variation due to MDA bias and reads cross-contamination.

Importantly, >90% of the obtained proviruses contained deletions in the primer positions used for most near-full-length nested PCRs, and would thus not have been characterized effectively by those methods. Furthermore, we recovered genome variants with mutations in LTR regions not captured by nested PCR. Four of 22 sequences for participant 1 contained an intact 5’ LTR through the major splice donor and thus have the potential to drive transcription of LTR transcripts into cellular sequences (Supplementary Fig. 10).

At the time of sampling, individual 3 had just started ART and individual 4 had received ART for a year before going off therapy for five, and both had viral loads of >100,000 copies ml$^{-1}$ (Fig. 4d). As expected for the setting of active HIV replication, SIP-seq revealed abundant unintegrated and intact virus compared to the two individuals receiving effective ART. Of the 18 sequences characterized, 16 were unintegrated, of which two were intact (Fig. 4e(i)). Only one provirus from each individual (one intact, one defective) was integrated, yielding integration sites. In participant 3, one provirus had a large internal deletion flanked by a direct repeat of ~124 bp of host sequences derived from the gene TNFSF10 (indicated by arrows in Fig. 4d). In participant 4, an intact provirus was integrated into the VMP1 gene. We also found two groups of three sequences that probably originate from clonally proliferating cells, showing a higher ratio of expanded clones compared to that in ART-suppressed individuals (Fig. 4e(ii)). Proviruses from participant 3 had duplications of the transcription factor TCF-1$\alpha$ binding site in several LTRs (Supplementary Fig. 11).

**Discussion**

SIP-seq allows the study of HIV proviruses that cannot be recovered by current PCR methods alone, and provides a more comprehensive analysis of the HIV genetic landscape in vivo. Capturing full-length proviral HIV genomes with their associated integration sites is essential for characterizing the latent reservoir and its contribution to HIV persistence. However, the rarity and lack of distinct surface markers for latently infected cells pose major obstacles to current techniques. The best option is a brute-force-based approach that aliquots thousands of cells in hundreds of wells and recovers copied amplicons using a multi-primer PCR spanning the HIV genome. However, in addition to often biasing against certain variants and generating artefacts that can confound analysis, the approach requires that samples be split into hundreds of aliquots to minimize doublets, which is slow, labour-intensive and reagent-consuming. Another strategy is to reactivate the reservoir and sort based on viral messenger RNA or surface markers. However, this changes the host-cell physiology away from latency and misses cells infected with virus that remain quiescent. Although microfluidic enrichment and pooled sequencing methods allow the characterization of rare genome sequences, SIP-seq is a technology to provide robust sequencing of individual proviruses in a format that is fast, cost-effective and scalable. These advantages will increase as samples with more genomes are analysed. For example, with plates, the amount of reagent used is fixed to the number of cells that must be processed, such that a tenfold increase in cells will require a comparable increase in reagent usage to ensure a high probability of single virus isolation by Poisson loading. By contrast, in SIP-seq, the frequency of viruses in the droplets is so low (<0.01%) that a tenfold increase in virus concentration increases the probability of co-encapsulation from $10^{-6}$ to $10^{-2}$, even with the same number of droplets. This means that more viruses can be recovered without increasing the number of droplets processed or reagent consumed, at the cost of an insignificant increase in doublet rate. This affords exciting opportunities for labelling and batching patient samples together to markedly increase the number of full-length viral genomes that can be recovered.

A limitation of SIP-seq is that it requires short-ranged TaqMan PCR amplification to detect DNA fragments with integrated proviruses. If the targeted regions are not present in a provirus, it will not be detected. For this reason, we designed primers to target conserved regions of HIV genomes, ensuring that proviruses with the potential for infectivity would be enriched for detection. Moreover, SIP-seq is flexible with respect to the TaqMan primers used, allowing their optimization to capture unexpected variants. The objective of our study was to capture a complete picture of HIV genomics in the patient, irrespective of genome intactness. We therefore used primers targeting a region in pol that is highly conserved. The result, as expected, is that most recovered genomes are inviable mutants, although we do observe intact proviruses, both integrated and unintegrated. So far, the lower estimates indicate one intact replication-competent genome in 100,000 cells. We can use additional primers, including targeting env, or the established intact proviral DNA assay (IPDA) primers to enrich for proviruses likely to be intact. Additionally, this minimum requirement is less constraining than nested PCR, which requires multiple primer pairs to amplify an incomplete portion of the HIV genome. As we have shown in a case example, viruses in which these regions are absent or divergent can easily be missed. Moreover, the multiple rounds of PCR used to separately amplify the provirus genome and its integration site after microwell MDA have the potential to generate artefacts. Because SIP-seq yields contigs containing the full-length provirus physically connected to its integration site on both sides, the resultant sequences are of high confidence.

By applying SIP-seq to people receiving ART, we confirm that proviruses in people with undetectable viraemia are most often defective and found in clonally expanded cells, and that ongoing viraemia is associated with higher levels of unintegrated and intact virus genomes. We also made several observations that have not been reported before. Most importantly, >90% of the proviruses we observed would not have been recovered by the nested-PCR methods commonly in use. These proviruses were defective but can still be immunogenic and influence cell function. These findings demonstrate that SIP-seq is a valuable tool for rigorous assessment of the latent reservoir, to confirm hypothesized attributes and to discover different ones. Although we focused on HIV-1 infection, SIP-seq should be applicable to other viral infections that include an integrated state as part of their life cycle.
Methods
Microfluidic device fabrication. The microfluidic devices were fabricated by soft lithography. Photomasks, designed in AutoCAD, were printed on transparencies. The features on the photomask were transferred to a negative photopolymer (MicroChem, SU-8 2025) on a silicon wafer (University Wafer) by UV photolithography. Polydimethylsiloxane (PDMS, Dow Corning, Sygslide 184) prepolymer mixture was poured over the patterned silicon wafer and cured in a 65 °C oven for 2 h. The PDMS replica was peeled off and punched for inlets and outlets by a 0.75-mm biopsy core (World Precision Instruments). The PDMS slice was bonded to a clear, circular, 5-mm silicon wafer (Harrick Plastics) by baking at 65 °C for 1 h to ensure strong bonding. The microfluidic channels were treated with Aquapel (PPG Industries) and baked at 65 °C for 30 min for hydrophobicity.

J-Lat cell culture and genomic DNA extraction. J-Lat HIV latency clones 5A8 and 15.4 were provided by M. Montano at the Gladstone Institute at UCSF. J-Lat 5A8 and J-Lat 15.4 cells were grown in RPMI 1640 medium (Gibco, cat. no. 11879030) with 10% fetal bovine serum (FBS, Gibco, cat. no. 26100079), and penicillin-streptomycin (Gibco, cat. no. 15104122). Cells were incubated with 5% CO2 at 37 °C. The genomic DNA of J-Lat cells and CD4+ T cells from individuals 3 and 4 were extracted with a Quick-DNA Miniprep Plus kit (Zymo Research, cat. no. D4068) according to the product protocol. The cell culture and DNA extraction from ice cultures (1737H3 and TC1288/B9) were conducted in the Gladstone biosafety level 3 (BSL3) facility.

Isolation of CD4+ T cells. Participant recruitment and informed consent were performed under protocols approved by the Institutional Review Board (IRB) at the US National Institutes of Health (NIH) and University of Washington. Peripheral blood mononuclear cells were isolated from whole blood by density gradient centrifugation, then incubated with Fc-receptor-blocking reagent for 10 min and stained with LIVE/DEAD Aqua stain, CD3-APC-H7 (BD Biosciences, cat. no. 641406), CD4-BV785 (BioLegend, cat. no. 317442), CD8-PerCP-Cy5.5 (BioLegend, cat. no. 301836), CD16-PerCP/Cy5.5 (BioLegend, cat. no. 302028), CD19-BV605 (BD Biosciences, cat. no. 562653), CD20-BV570 (BioLegend, cat. no. 302332), CD27-Allo700 (BioLegend, cat. no. 302814), CD32-PE (BioLegend, cat. no. 303206); CD45RO-ECD (Beckman Coulter, cat. no. IM47212), CD3-PE/Cy5 (BD Biosciences, cat. no. 551065) and TCRCy5-APC (BD Biosciences, cat. no. 555718). CD4+ T cells were isolated by fluorescence-activated cell sorting on a FACSAria system (Becton Dickinson) using previously described protocols18.

Extraction of cell-associated DNA. Sorted cells were sedimented by centrifugation at 400g for 7 min at 4 °C, lysed in RNAzol RT at <5 × 106 cells ml−1, homogenized by pipetting, and stored at −80 °C until extraction. For DNA extraction, 0.4 vol. of RNAzol RT in step 1.4 was added to each lysate to allow aqueous and organic phase separation. The organic phase of each lysate was solubilized in DNAzol (Molecular Research Centers) to allow DNA extraction according to the manufacturer's instructions. The average fragment size of this DNA was >75 kb. Approximately 100,000–250,000 cells equivalent DNA was processed to generate the probes for each donor. The HIV frequencies in these samples were approximately 1/500–1/2,000, measured by bulk qPCR.

ICE. CD4+ T cells were purified by negative selection and cultured in the presence of antiretroviral drugs to block horizontal transmission of virus in the culture, interleukin-2 and phytohemagglutinin at limiting dilution, such that 1/10 to 1/1,000 cells corresponded to an infected clone post expansion. A portion of each culture was screened for the presence of viral genomes using qPCR after 10 days of culture, and positive cultures were frozen for SIP-seq analysis after 14 days of culture.

Nested PCR. Nested PCR was performed as described previously19. Post SIP-seq nested PCR amplified the full-length HIV genomes and integration sites in two ICE clones. The primers used in post SIP-seq nested PCR were designed based on the adjacent human sequences and are listed in Supplementary Table 4. Amplified DNA was prepared for next-generation sequencing using the Illumina Nextera XT DNA Library Preparation Kit according to the manufacturer's instructions.

MDA in microfluidic emulation droplets. The MDA reaction mixture was prepared with phi29 DNA Polymerase (New England BioLabs, cat. no. M0269L) following the manufacturer's protocol. Purified T-cell DNA was added into a 100-µl reaction to confirm by qPCR the DNA polymerase reaction. 200 ng ml−1 bovine serum albumin, 200 µM deoxynucleotide triphosphates, 25µM random hexamers and 60 µM of phi29 DNA polymerase. The reaction mixture was loaded immediately into a 1-ml syringe backfilled with HFE-7500 fluorocarbon oil (3M, cat. no. 98-0212-2928-5) and injected into a flow-focus droplet maker (Supplementary Fig. 12a) at a constant rate of 400 µl h−1 by a syringe pump (New Era). HFE-7500 oil with 2% (wt/wt) poly(ethylene glycol) 6000 (Sigma-Aldrich) and 0.8 M 1,2-propanediol (Sigma-Aldrich) were used for droplet stability during thermal cycling. 1.2-Propanediol was used as a PCR enhancer for low-temperature denaturation. The following two sets of conditions were used: 1. Another stream of HFE-7500 oil at 1,300 µl h−1 was used as spacer. PCR reagent (flow rate 200 µl h−1) was introduced at the ~2-kV a.c. potential at the electrode from a 2-V voltage. The merged droplets containing MDA products and now 1X TaqMan PCR mixture were collected into PCR tubes. The bottom oil layer was removed and replaced with FC-40 fluorinated oil (Sigma-Aldrich, cat. no. 51142-49-5) with 5% (wt/wt) PEG-PFPE surfactant to maintain emulsion stability during thermal cycling. The cycling was performed on a T100 thermal cycler (Bio-Rad) with the following conditions: 2 min at 86 °C; 35 cycles of 30 s at 86 °C, 90 s at 60 °C and 30 s at 72 °C; and, finally, 5 min at 72 °C. A low-temperature incubation temperature of 86 °C was used to minimize DNA fragmentation (Supplementary Fig. 13).

Dielectrophoretic sorting. After thermal cycling, the droplets were transferred to a 1-ml syringe and reinfected into a microfluidic dielectrophoresis (DEP) sorter (as shown in Supplementary Fig. 12c) at 80 µl h−1. HFE-7500 oil was loaded into a 5-ml syringe and injected into the DEP device as spacer oil at a flow rate of 0.6 µl h−1. Another stream of HFE-7500 oil at 1.300 µl h−1 was introduced at the sorting junction to drive the droplets to waste collection when the electrode was deactivated. A negative pressure from a syringe at ~1,300 µl h−1 was applied to the waste collection to make sure the drops flowed into waste when deactivated. The coating was performed with 2 M NaCl solution. The photomultipliers (Thorlabs, PMM01 model) were controlled by a LabVIEW program (National Instruments) to measure droplet fluorescence when passing the laser. The sorting electrode was activated when the fluorescence intensity was higher than a preset threshold. A high-voltage amplifier (Trek) was used to amplify the electrode pulse to 1 kV for DEP sorting. The sorted drops were collected into individual PCR tubes with one droplet per PCR tube for single viral sequencing.

Library preparation and sequencing of single droplets. The sorted individual droplet with its carrier oil in a PCR tube was dried out in a vacuum chamber for ~1.5 h, then 1.5 µl of deionized H2O was added to dissolve the sorted DNA, of which 0.5 µl was used to confirm sorting of the target by qPCR on a different region of the virus genome than pol and env. The qPCR was performed using 1X Kapa probe fast qPCR 2X master mix (KAPA Biosystems, cat. no. KK4702), 140 nM TaqMan probe (IDT), 300 nM of each primer (IDT) and 0.5 µl of sorted DNA, under the following conditions: 3 min at 95 °C; 40 cycles of 5 s at 95 °C, 15 s at 58 °C and 30 s at 60 °C in a QuantStudio 5 Real-Time PCR System (Thermo Fisher Scientific). The sequences of the probes and primers are listed in Supplementary Table 5. The shifts in the curves (for one or more of the five loci) confirm sorting of HIV sequence-containing drops. To construct Illumina libraries from picogram amounts of DNA recovered from single droplets, the standard library reaction was significantly modified to increase the reaction efficiency by reducing the reaction volume and optimizing the enzymes. The remaining 1 µl of redissolved contents was tagged using 0.6 µl of TD Tagmentation buffer and 0.3 µl of ATM Tagmentation enzyme from the Nextera XT DNA Library Prep Kit (Illumina, cat. no. FC-121-1030) for 5 min at 55 °C. A 1-µl volume of Neutralize Tagment buffer was added to stop the reaction and then 50 µl of 0.5X PEG-800 (New England BioLabs) and 1X PEG-800 PEG-PFPE surfactant solution containing 1.5 µl of NPM PCR master mix, 0.5 µl of each of the index primers i5 and i7 from the Nextera Index Kit (Illumina, cat. no. FC-121-1011) and
1.5 μl of H₂O, and placed on a thermal cycler with the following program: 3 min at 72 °C, 30 at 95 °C; 20 cycles of 10 at 95 °C, 30 at 55 °C and 30 at 72 °C; and finally 5 min at 72 °C. The DNA library was purified and size-selected for 200–600 bp fragments using Agencourt AMPure XP beads (Beckman Coulter), then quantified using a Qubit dsDNA HS assay kit (Thermo Fisher Scientific) and a high-sensitivity DNA bioanalyzer chip (Agilent). The library was sequenced using Illumina MiSeq PE300 or HiSeq PE300. Post SIP-seq Sanger sequencing verified detected mutations in the 5′ LTR of a provirus. Post SIP-seq qPCR confirmed the integration sites in four proviruses. The primers used in post SIP-seq qPCR and Sanger sequencing were designed based on the specific integration sites and are listed in Supplementary Table 6. The integrate-specific qPCR was performed in Kapa probe fast qPCR master mix using the protocol described above. Sanger sequencing was performed by Quintara Biosciences.

Bioinformatic analysis. Sequencing reads passing quality control were mapped to the HIV reference (HXB2) using Bowtie 2®. Genomic coverage as a function of genome position was generated using SAMTools®. Approximately one million paired-end reads of 150 bp were used for enriched or unenriched cell line samples, ~2 million paired-end reads of 150 bp each were analysed for patient-derived ICR samples, and ~10 million reads of 300 bp were used to study participant samples to achieve an average coverage of >50x. This is consistent with the read requirement calculation that ~1/500 to 1/2,000 cells contain an HIV genome and one cellular DNA was partitioned into ~100,000 drops for primary samples. The overall recovery was calculated as the number of virus genome copies recovered from SIP-selected and total number of virus genome estimated by bulk qPCR. BCFtools, within SAMtools, was used for variant calling. Chimaeric reads and their soft-clipped regions from alignments of non-HIV regions were extracted using extractSofclipped (https://github.com/dpryan79/SE-MEI). Soft-clipped reads were analysed by a web base tool for integration site localization (https://indra.mullins.microbiol.washington.edu/integrationsites/). Reads were de novo assembled using SPAdes® in Python, and contigs were evaluated by QUAST® to confirm the integration sites and provide additional information on proviral structures. The LTR specific reads were aligned with the guidance of de novo assembled contigs. The reads (including those reads only aligned to LTRs) were assigned to 5’ or 3’ LTRs that had a contig to connect to other regions. If both LTRs were connected to other regions (including those reads only aligned to LTRs) were assigned to 5

Gaps in the coverage map were considered genome deletions if there were sequencing reads spanning over the gap. To differentiate random dropouts from actual deletions in the regions with no reads, we developed a process to determine the P value that a gap is a dropout. Because there are few deletions in the region 800–3,000 bp, we used this region to measure the expected dropout size distribution. We then calculated the P value for gaps in other regions to be dropout-free. We also defined gaps of <200 bp as dropouts from MDA bias®, and thus smoothed the coverage maps with a bin size of 200 bp.

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

Data availability
The sequence data have been deposited in the Sequencing Read Archive under BioProject accession number PRJCA006195. All other data supporting the findings of this study are available within the paper and its Supplementary Information. Source data are provided with this paper.

Code availability
Custom scripts and functions are available at https://github.com/abateLab.

Received: 3 October 2020; Accepted: 28 January 2022; Published online: 28 March 2022

References
1. Siliciano, R. F. & Greene, W. C. HIV latency. Cold Spring Harb. Perspect. Med. 1, a007096 (2011).
2. Finzi, D. et al. Latent infection of CD4+ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. Nat. Med. 5, 512–517 (1999).
3. Martinson, G. C., Fauci, A. S. &ashes. S. R. HIV infection: epidemiology, pathogenesis, treatment and prevention. Lancet 384, 258–271 (2014).
4. Deeks, S. G. Towards an HIV cure: a global scientific strategy. Nat. Rev. Immunol. 12, 607–614 (2012).
5. Buell, K. G. et al. Lifelong antiretroviral therapy or HIV cure: the benefits for the individual patient. AIDS Care Psychol. Socio Med. Asp. AIDS/HIV 28, 242–246 (2016).
6. Dassenbach, C. W. & Fauci, A. S. Thirty years of HIV and AIDS: future challenges and opportunities. Ann. Intern. Med. 154, 766–771 (2011).
7. Rui, J. K. et al. Proviruses with identical sequences comprise a large fraction of the replication-competent HIV reservoir. PLoS Pathog. 13, e1006283 (2017).
8. Symons, J., Cameron, P. U. & Levin, S. R. HIV integration sites and implications for maintenance of the reservoir. Curr. Opin. HIV AIDS 13, 152–159 (2018).
9. Wiegand, A. et al. Single-cell analysis of HIV-1 transcriptional activity reveals expression of proviruses in expanded clones during ART. Proc. Natl Acad. Sci. USA 114, E3659–E3668 (2017).
10. Wagner, T. A. et al. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. Science 345, 570–573 (2014).
11. Wagner, T. A. et al. An increasing proportion of monotypic HIV-1 DNA sequences during antiretroviral treatment suggests proliferation of HIV-infected cells. J. Virol. 87, 1770–1778 (2013).
12. Maldarelli, F. et al. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. Science 345, 179–183 (2014).
13. Brunner, K. M. et al. A quantitative approach for measuring the reservoir of latent HIV-1 proviruses. Nature 566, 120–125 (2019).
14. Brunner, K. M. et al. Defective proviruses rapidly accumulate during acute HIV-1 infection. Nat. Med. 22, 1043–1049 (2016).
15. Linkauf, K. et al. Infect HIV-1 proviruses accumulate at distinct chromosomal positions during prolonged antiretroviral therapy. J. Clin. Invest. 129, 988–998 (2019).
16. Patro, S. C. et al. Combined HIV-1 sequence and integration site analysis informs viral dynamics and allows reconstruction of replicating viral ancestors. Proc. Natl Acad. Sci. USA 116, 25891–25899 (2019).
17. Menzel, 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).
18. Gall, A. et al. Universal amplification, next-generation sequencing, and assembly of HIV-1 genomes. J. Clin. Microbiol. 50, 3838–3844 (2012).
19. Lan, F., Demaree, B., Ahmed, N. & Abate, A. Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding. Nat. Biotechnol. 35, 640–646 (2017).
20. Clark, I. C. & Abate, A. R. Finding a helix in a haystack: nucleic acid cytometry with droplet microfluidics. Lab Chip 17, 2032–2045 (2017).
21. Lim, S. W., Tran, T. M. & Abate, A. R. PCR-activated cell sorting for cultivation-free enrichment and sequencing of rare microbes. PLoS ONE 10, e0113549 (2015).
22. Massanella, M. & Richman, D. D. Measuring the latent reservoir in vivo. J. Clin. Invest. 126, 644–647 (2016).
23. Dean, F. B. et al. Comprehensive human genome amplification using multiple displacement amplification. Proc. Natl Acad. Sci. USA 99, 5261–5266 (2002).
24. Linkauf, K. et al. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection of latent HIV. Cell Rep. 7, e7749 (2014).
25. Lenasi, T., Contreras, X. & Peterlin, B. M. Transcriptional interference antagonizes proviral gene expression to promote HIV latency. Curr. Opin. Microbiol. 5, 123–133 (2008).
26. Mullins, J. I. & Frenkel, L. M. Clonal expansion of human immunodeficiency virus-infected cells and human immunodeficiency virus persistence during antiretroviral therapy. J. Infect. Dis. 215, SI19–SI27 (2017).
27. Pollack, R. A. et al. Defective HIV-1 proviruses are expressed and can be recognized by cytotoxic T lymphocytes, which shape the proviral landscape. Cell Host Microbe 21, 494–506 (2017).
28. Lorenzo-Redondo, R. et al. Persistent HIV-1 replication maintains the tissue reservoir during therapy. Nature 530, 51–56 (2016).
29. Perez, L. et al. Conflicting evidence for HIV enrichment in CD32+ CD4 T cells. Nature 561, E9–E16 (2018).
30. Rouseau, C. M. et al. Large-scale amplification, cloning and sequencing of near full-length HIV-1 subtype C genomes. J. Virol. Methods 136, 118–125 (2006).
31. Hamid, F., Kim, J. & Shin, C. G. Distribution and fate of HIV-1 unintegrated DNA species: a comprehensive update. AIDS Res. Ther. 14, 9 (2017).
32. Cohn, L. B. et al. HIV-1 integration landscape during latent and active infection. Science 360, 420–423 (2015).
33. Liu, R. X. Single-cell transcriptional landscapes reveal HIV-1-driven aberrant host gene transcription as a potential therapeutic target. Sci. Transl. Med. 12, eaaz8082 (2020).
Acknowledgements
This work was supported by a Chan Zuckerberg Biohub grant to A.R.A., a National Institutes of Health grant (R01 HG008978) to A.R.A., a National Institutes of Health grant (U01 AI129206) to A.R.A. and E.A.B. and National Institutes of Health grants (R01 AI1125026 and R33 AI112361) to J.I.M. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions
C.S. and A.R.A. conceived the project. C.S., L.L., X.L., Y.L. and P.X. performed the experiments. C.S. sequenced the samples and analysed the data. C.S. and A.R.A. wrote the initial draft of the manuscript. L.P. assisted with patient sample processing. J.I.M. and E.A.B. revised the manuscript. All authors read, reviewed and approved the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41551-022-00864-8.

Correspondence and requests for materials should be addressed to Adam R. Abate.

Peer review information Nature Biomedical Engineering thanks Angela Ciuffi and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2022.
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 our Editorial Policies 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
- No software was used

Data analysis
- Bowtie 2.3.4.3
- SAMtools 1.5
- BCFTools 1.4.1
- extractSoftclipped: https://github.com/dpryan79/SE-MEI
- Integration-site localization: https://indra.mullins.microbiol.washington.edu/integrationsites/
- SPAdes 3.10.1
- QUAST: http://quast.bioinf.spbau.ru
- DIVEIN: https://indra.mullins.microbiol.washington.edu/DIVEIN/
- Highlighter: https://www.hiv.lanl.gov/content/sequence/HIGHLIGHT/highlighter_top.html
- Python 3.5.2
- Labview 2017 SP1

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 and 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

The sequence data have been deposited in the Sequencing Read Archive under BioProject accession number PRJCA006195. All other data supporting the findings of this study are available within the paper and its Supplementary information. Source data are provided with this paper.

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**
No explicit calculations were performed to determine sample sizes. Rather, we aimed to capture various HIV genome structures from participants on suppressive ART with or without detectable plasma HIV-1 virus load. Therefore, we investigated 2 patients suppressed by ART and 2 patients not fully suppressed by ART. We observed both intact and defective, integrated and non-integrated genome structures.

**Data exclusions**
No data were excluded.

**Replication**
Each cell-line sample was analysed in at least 3 SIP-seq experiments, and the primary sample was analysed in at least 2 SIP-seq experiments, owing to limited availability.

**Randomization**
Randomization was not relevant to this study. All samples were processed and analysed identically, regardless of patient clinical history or other external criteria.

**Blinding**
Blinding was not relevant for this study. Sample labels were required to associate SIP-seq data with corresponding bulk measurement data, and to assess the extent of agreement between the measurements.

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 and archaeology |
|     | Animals and other organisms |
|     | Human research participants |
|     | Clinical data         |
|     | Dual use research of concern |

### Methods

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

**Antibodies**

| Antibodies used |
|-----------------|
| CD3-APC-H7 (1:80; clone SK7; BD Biosciences; Cat# 641406), CD4-BV785 (1:50; clone OKT4; BioLegend; Cat# 317442), CD8-PacBlue (1:160; clone 3B8; Invitrogen; Cat# MHCDO828), CD14-BV650 (1:80; clone M5E2; BioLegend; Cat# 301836), CD16-PerCP/Cy5.5 (1:160; clone 3G8; BioLegend; Cat# 302028), CD19-BV605 (1:40; clone SJ25C1; BD Biosciences; Cat# 562653), CD20-BV570 (1:20; clone 2H7; BioLegend; Cat# 302332), CD27-Alx700 (1:40; clone O332; BioLegend; Cat# 302814), CD32-PE (1:160; clone FUN-2; BioLegend; Cat# 302026), CD45RO-ECD (1:50; clone UCHL1; Beckman Coulter; Cat# IM2712U), CD123-PE/Cy5 (1:40; clone 9F5; BD Biosciences; Cat# 551065), and TCR-APC (1:20; clone B1; BD Biosciences; Cat# 555718). |
Validation
All antibodies used are commercial reagents, and reactivity with the relevant human antigens has been well-validated by the manufacturers listed above.

**Eukaryotic cell lines**

| Policy information about cell lines |  |
|------------------------------------|--|
| **Cell line source(s)** | JLat 5A8 and JLat 15.4 were kindly provided by Dr. Mauricio Montano at the Gladstone Institute at UCSF |
| **Authentication** | We confirmed by PCR that the cells lines contained the expected 1 copy of HIV per cell. |
| **Mycoplasma contamination** | The cells tested negative for mycoplasma. |
| **Commonly misidentified lines (See ICLAC register)** | No commonly misidentified lines were used. |

**Human research participants**

| Policy information about studies involving human research participants |  |
|-----------------------------------------------------------|--|
| **Population characteristics** | Blood samples were obtained from HIV-1 infected adults on suppressive ART with or without detectable plasma HIV-1 virus load. Supplementary Table 1 provides a summary of participant characteristics. Age would be considered identifiable information and hence is not reported. |
| **Recruitment** | Participants were recruited from healthcare settings to donate blood samples for research purposes, with financial compensation. Potential biases include a selection for individuals motivated to participate in research, which could increase the likelihood of full adherence to antiretroviral therapy in these individuals. This would benefit the study by ensuring that the data obtained arise from the scenario of effective ART, which was one of the intended areas of study. |
| **Ethics oversight** | For participants #1 and #2, the protocol was approved by the Institutional Review Board of the National Institutes of Health. For participants #3 and #4, the protocol was approved by the Institutional Review Board of the University of Washington. |

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