Henry Ford Health System
Henry Ford Health System Scholarly Commons

Infectious Diseases Articles | Infectious Diseases

12-26-2018

Hematopoietic stem and progenitor cells are a distinct HIV reservoir that contributes to persistent viremia in suppressed patients.

Thomas D. Zaikos

Valeri H. Terry

Nadia T. Sebastian Kettinger

Jay Lubow

Mark M. Painter

See next page for additional authors

Follow this and additional works at: https://scholarlycommons.henryford.com/infectiousdiseases_articles

Recommended Citation
Zaikos TD, Terry VH, Sebastian Kettinger NT, Lubow J, Painter MM, Virgilio MC, Neevel A, Taschuk F, Onafuwa-Nuga A, McNamara LA, Riddell Jt, Bixby D, Markowitz N, and Collins KL. Hematopoietic stem and progenitor cells are a distinct HIV reservoir that contributes to persistent viremia in suppressed patients. Cell Rep 2018; 25(13):3759-3773.e9

This Article is brought to you for free and open access by the Infectious Diseases at Henry Ford Health System Scholarly Commons. It has been accepted for inclusion in Infectious Diseases Articles by an authorized administrator of Henry Ford Health System Scholarly Commons.

Authors

Thomas D. Zaikos, Valeri H. Terry, Nadia T. Sebastian Kettinger, Jay Lubow, Mark M. Painter, Maria C. Virgilio, Andrew Neevel, Frances Taschuk, Adewunmi Onafuwa-Nuga, Lucy A. McNamara, James Riddell, Dale Bixby, Norman Markowitz, and Kathleen L. Collins

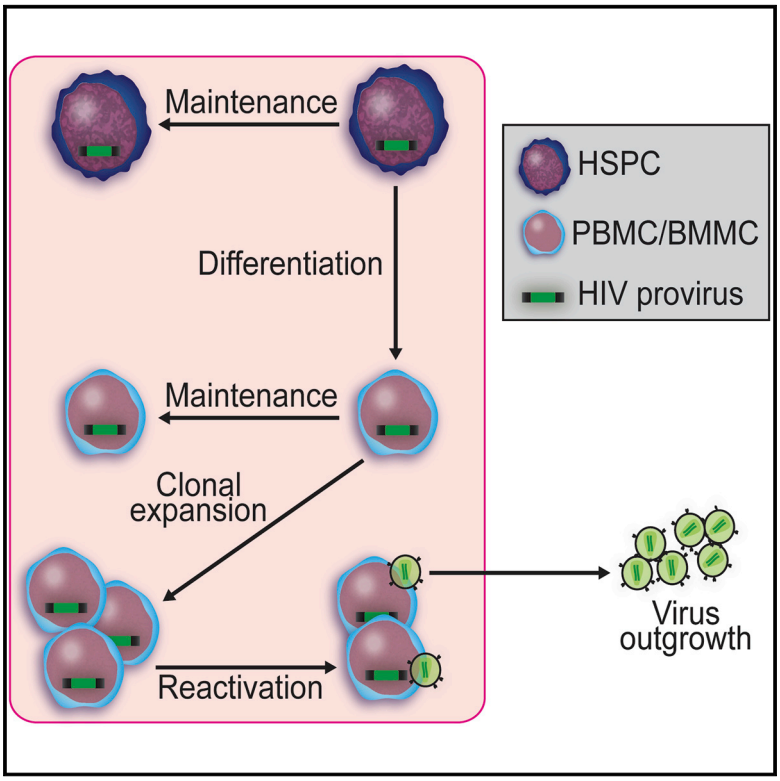
This article is available at Henry Ford Health System Scholarly Commons: https://scholarlycommons.henryford.com/infectiousdiseases_articles/28

Article

Cell Reports

Hematopoietic Stem and Progenitor Cells Are a Distinct HIV Reservoir that Contributes to Persistent Viremia in Suppressed Patients

Graphical Abstract



Authors

Thomas D. Zaikos, Valeri H. Terry, Nadia T. Sebastian Kettinger, ..., Dale Bixby, Norman Markowitz, Kathleen L. Collins

Correspondence

klcollin@umich.edu

In Brief

HIV causes an infection that persists even when optimal therapy is used. Zaikos et al. provide evidence that HIV-infected progenitor cells from the bone marrow can amplify virus through normal cellular growth pathways in some treated people.

Highlights

- Hematopoietic stem and progenitor cells can serve as long-term reservoirs of HIV
- HSPCs harbor both infectious and defective proviral genomes
- HSPCs are an important source of residual plasma virus in treated people
- Clonally amplified HIV proviruses contribute to residual plasma virus

Zaikos et al., 2018, Cell Reports *25*, 3759–3773
December 26, 2018 © 2018 The Author(s).
https://doi.org/10.1016/j.celrep.2018.11.104

CellPress

Cell Reports
Article

OPEN ACCESS
CellPress

Hematopoietic Stem and Progenitor Cells Are a Distinct HIV Reservoir that Contributes to Persistent Viremia in Suppressed Patients

Thomas D. Zaikos,[1] Valeri H. Terry,[2] Nadia T. Sebastian Kettinger,[3,4,9] Jay Lubow,[1] Mark M. Painter,[6] Maria C. Virgilio,[3] Andrew Neevel,[2] Frances Taschuk,[2,10] Adewunmi Onafuwa-Nuga,[2,11] Lucy A. McNamara,[1,12] James Riddell IV,[5] Dale Bixby,[7] Norman Markowitz,[8] and Kathleen L. Collins[1,2,3,4,5,6,13,*]

[1]Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI, USA
[2]Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA
[3]Program in Cellular and Molecular Biology, University of Michigan, Ann Arbor, MI, USA
[4]Medical Scientist Training Program, University of Michigan, Ann Arbor, MI, USA
[5]Division of Infectious Disease, Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA
[6]Graduate Program in Immunology, University of Michigan, Ann Arbor, MI, USA
[7]Division of Oncology, Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA
[8]Division of Infectious Diseases, Henry Ford Hospital, Detroit, MI, USA
[9]Present address: Internal Medicine Residency Program, University of Michigan, Ann Arbor, MI, USA
[10]Present address: Department of Microbiology, University of Pennsylvania School of Medicine, Philadelphia, PA, USA
[11]Present address: Medicine-Pediatrics Residency Program, University of Michigan, Ann Arbor, MI, USA
[12]Present address: Division of Bacterial Diseases, National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention, Atlanta, GA, USA
[13]Lead Contact
*Correspondence: klcollin@umich.edu
https://doi.org/10.1016/j.celrep.2018.11.104

SUMMARY

Long-lived reservoirs of persistent HIV are a major barrier to a cure. CD4$^+$ hematopoietic stem and progenitor cells (HSPCs) have the capacity for lifelong survival, self-renewal, and the generation of daughter cells. Recent evidence shows that they are also susceptible to HIV infection *in vitro* and *in vivo*. Whether HSPCs harbor infectious virus or contribute to plasma virus (PV) is unknown. Here, we provide strong evidence that clusters of identical proviruses from HSPCs and their likely progeny often match residual PV. A higher proportion of these sequences match residual PV than proviral genomes from bone marrow and peripheral blood mononuclear cells that are observed only once. Furthermore, an analysis of near-full-length genomes isolated from HSPCs provides evidence that HSPCs harbor functional HIV proviral genomes that often match residual PV. These results support the conclusion that HIV-infected HSPCs form a distinct and functionally significant reservoir of persistent HIV in infected people.

INTRODUCTION

Combination anti-retroviral therapy (cART) efficiently blocks HIV replication *in vivo*. However, the virus is able to persist within infected individuals and rebound viremia frequently occurs if cART is interrupted (Davey et al., 1999). Failure of cART is thought to result from persistent long-lived cells that harbor integrated HIV genomes that are unaffected by anti-retroviral therapy. Resting memory CD4$^+$ T cells are a well-characterized HIV reservoir (Chun et al., 1997; Finzi et al., 1997, 1999). In addition, emerging data support the possibility that non-CD4$^+$ T cells may also form persistent HIV reservoirs (Araínga et al., 2017; Avalos et al., 2016; Honeycutt et al., 2017; Sundstrom et al., 2007; Zhu et al., 2002). In particular, hematopoietic stem and progenitor cells (HSPCs) are a long-lived cell type that has been shown to be infected *in vivo* and is capable of propagating integrated provirus to CD4$^+$ and CD4$^-$ progeny (Carter et al., 2010; Sebastian et al., 2017). Moreover, infected HSPCs can produce virions upon latency reversal *in vitro* (Zaikos et al., 2018). Therefore, it is possible that HSPCs could contribute to persistent and rebound viremia directly or by serving as a source of infected daughter cells that can be activated to produce virus.

Residual plasma virus (PV), defined here as a PV level of less than 48 copies per mL, can be detected in treated people by ultra-sensitive techniques even when virus is undetectable by standard clinical tests. Sequence analysis of residual PV and rebounding PV in HIV-infected people indicates that virions likely come from the activation of latent provirus that had been archived since before the initiation of therapy rather than from low-level replication and spread of cART-resistant virus (Eisele and Siliciano, 2012; Kearney et al., 2014).

The cellular source of residual virus remains poorly understood. A number of studies have shown that proviral DNA from CD4$^+$ T cells and other peripheral blood mononuclear cells (PBMCs) frequently does not match residual PV (Bailey et al., 2006; Brennan et al., 2009; Buzon et al., 2014; Chun et al., 2000; Sahu et al., 2009). One small study of two donors found that virus induced *ex vivo* from a subset of activated T cells matched residual PV, leading the authors to conclude that it is

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

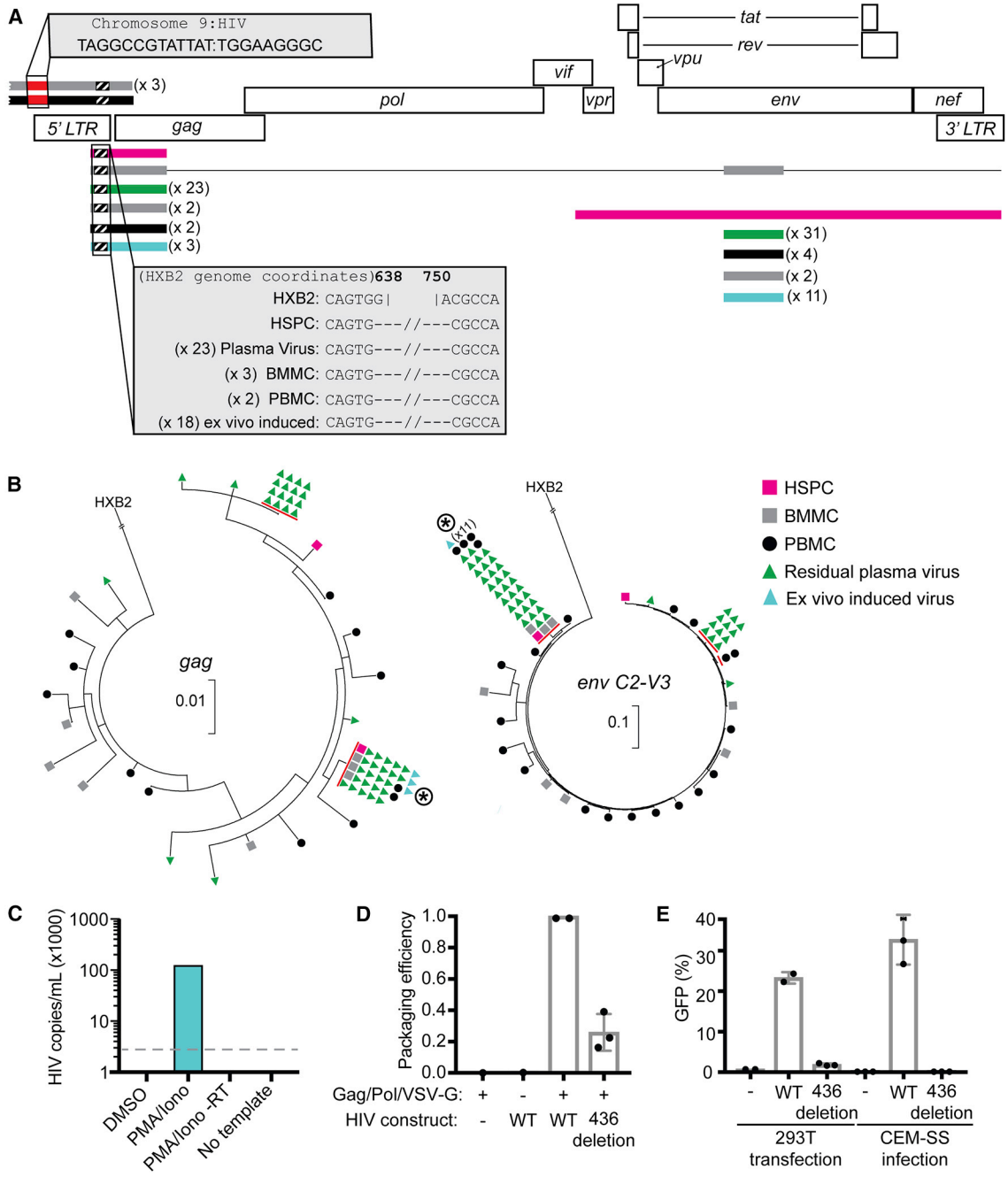


Figure 1. A Signature Deletion Marks a Clonal, HSPC-Associated-Proviral Genome that Releases Non-infectious Clonal Virus *In Vivo* and *Ex Vivo*

(A) HIV genome map and identical viral sequences from indicated tissue source from donor 436000. The solid bars indicate fully sequenced, genetically intact amplicons. The black and white hatched bars indicate the location of a signature deletion with nucleotide sequence shown in the figure insert. The solid red bar indicates sequenced human chromosome 9:HIV integration site, with nucleotide sequence shown in figure insert. The thin line connecting gag and *env* amplicons indicates sequences that originated from the same first-round reaction performed at limiting dilution. The number of times a sequence was observed is indicated by a separate rectangle or by a number (x n).

(B) Maximum-likelihood phylogenetic trees of *gag* and *env* HIV sequences from the indicated tissue sources. Identical sequence groups are designated with a red bar, and the number of symbols represents the number of times the sequence was observed. Identical sequence groups depicted in (A) are marked with the circled asterisk. Scale indicates nucleotide substitutions per site.

(C) Summary graph of HIV mRNA abundance in supernatant from a viral outgrowth assay using $CD4^+$ cells isolated from donor 436000 BMMCs. Cells were stimulated with PMA and ionomycin or matched DMSO solvent control. The dashed line indicates limit of detection for the assay. The viral outgrowth assay was performed once with all samples analyzed in duplicate in the qPCR assay.

(legend continued on next page)

possible that residual PV may originate from a minor population of circulating $CD4^+$ T cells (Anderson et al., 2011). Whether HSPCs contribute to residual PV is unknown.

Genetic characterization of provirus from $CD4^+$ T cells has demonstrated that many are defective and that defective proviruses rapidly accumulate (Bruner et al., 2016; Hiener et al., 2017; Imamichi et al., 2016). Analyses of near-full-length proviral genomes indicate that 2%–12% of $CD4^+$ T cell-derived HIV proviral genomes have full open reading frames (Bruner et al., 2016; Hiener et al., 2017; Ho et al., 2013). Whether provirus from HPSCs is defective has not previously been examined.

A determination that proviral genomes from HSPCs contain open reading frames that match residual PV would provide evidence that HSPC-derived genomes are indeed functional. Here, we examine provirus and residual PV from HSPCs and other cell types from a cohort of 24 donors, providing evidence in support of this conclusion.

RESULTS

Single-Genome Amplification of Proviral DNA from HSPCs and Other Cellular Subsets

We recruited 53 HIV-infected donors on cART with viral loads <48 copies per mL. We obtained samples of blood and bone marrow from each donor. Ten donors were excluded for insufficient sample size or suboptimal HSPC sort purity. Table S1 summarizes the characteristics of the included donors. Ten donors contributed samples more than once over periods spanning 4 months to nearly 5 years (Table S1). For each included donor, we isolated two different types of HSPCs in sequential sorts ($CD133^+$ [sort 1] or $CD34^+CD133^-$ [sort 2]). Each sort was analyzed individually for purity. We excluded any individual HSPC sorts that were <80% HSPCs or >1% $CD3^+$ cells (Figure S1A). In most cases, included samples were well within these criteria, with mean purity 90% or higher and mean $CD3^+$ T cell contamination rates less than 0.4% (Table S2).

To determine whether HSPCs serve as a source of PV, we amplified individual HIV proviral genome sequences from highly purified HSPCs, HSPC-depleted bone marrow mononuclear cells (BMMCs), and PBMCs. For this analysis, we used a highly sensitive multiplex PCR protocol that was performed at limiting dilution to amplify individual *gag* or *env* sequences in separate second-round reactions (Sebastian et al., 2017). This protocol was applied to DNA from both HSPC populations ($CD133^+$ [sort 1] and $CD34^+CD133^-$ [sort 2]), providing four assays for each donor's progenitor cells (*gag* and *env* PCRs using samples from sort 1 and sort 2). Based on the sort purity and the frequency of HIV provirus in $CD3^+$ samples, we performed a statistical analysis to determine the likelihood that provirus was amplified from contaminating $CD3^+$ cells rather than an HSPC (McNamara et al., 2013; Sebastian et al., 2017). Only samples that yielded HIV proviral DNA that was unlikely to be due to T cell contamination were included in our final analyses (Figure S1B). Figure S1C summarizes the proportion of donors who had positive results in each of the four assays.

Additionally, all viral sequences were compared phylogenetically to every previously acquired sequence and lab strain to rule out cross-contamination. All included donor sequences clustered appropriately except for outliers that contained frameshifts.

We determined that 25 of the 43 included donors had detectable provirus in HSPC DNA. Sort purity and $CD3^+$ T cell contamination levels were very similar between the donors who had detectable HIV provirus in HSPC DNA and those that did not (Table S2). However, we did note a trend toward a lower yield of HSPCs available to screen for the negative donors (Table S2).

Isolation and Characterization of Residual PV

For each donor, we also harvested plasma from which we isolated residual PV. Purified viral RNA was converted to cDNA, and HIV sequences were amplified using the multiplex PCR described above. We successfully detected residual PV in samples from 24 out of 25 donors who had provirus isolated from HSPC DNA, including one donor who initiated cART during the acute phase (donor 503501; Table S1).

To identify potential cellular sources of residual PV, we compared all PV sequences to all proviral DNA sequences. As expected, some of the PV matched provirus from PBMCs and BMMCs, a large percentage of which are T cells. However, phylogenetic analysis revealed that proviral sequences isolated from highly purified HSPCs exactly matched amplified residual PV in 8 of 24 donors. We noted significantly higher mean CD4 counts at the time of sampling and a greater number of PV amplicons recovered per donor in the 8 donors with matching HSPC and PV sequences ($p = 0.0005$ and $p = 0.04$, respectively; Table S3). The phylogenetic analysis for these sequences for each of the 8 donors is shown in Figures 1, 2, 3, and 4, and Figures S2, S3, S4, and S5. An analysis of genetic diversity based on average pairwise genetic distance (APD) indicated that donors with matching HSPC proviral DNA and PV had a similar degree of genetic diversity as those that did not (Table S4). The mean APD for provirus was >1% and the mean APD for PV was >0.5% for both groups. As expected, *env* sequences were significantly more diverse than *gag* sequences ($p < 0.05$). Based on the APD, the length of the amplicons and the number of observed versus expected identical sequences, the probability the observed

(D) Summary graph of packaging efficiency of wild-type HIV reporter construct (NL4-3 ΔGPE-GFP wild-type [WT]) or the reporter construct containing the deletion shown in (A). Reporter virus constructs were transfected into 293T cells with helper plasmids as indicated. Packaging efficiency was calculated by dividing the abundance of HIV mRNA in the supernatant by the normalized cell-associated HIV mRNA from each sample. Results are normalized to that of WT reporter co-transfected with helper plasmids. Results are from two biological replicate experiments with the deleted construct tested in duplicate in one. Samples were analyzed in triplicate in the PCR assay.

(E) Summary graph of reporter virus Env expression in transfected 293T cells with replicates as in (D) (left panel) and infected CEM-SS T cells as assessed by the frequency of GFP^+ cells detected by flow cytometry (three biological replicates; right panel). For infections, CEM-SS cells were treated with transfected cell supernatant containing equal numbers of HIV copies by qPCR. Error bars indicate SD.

See also Tables S1–S5 and Figures S1, S6, and S7.

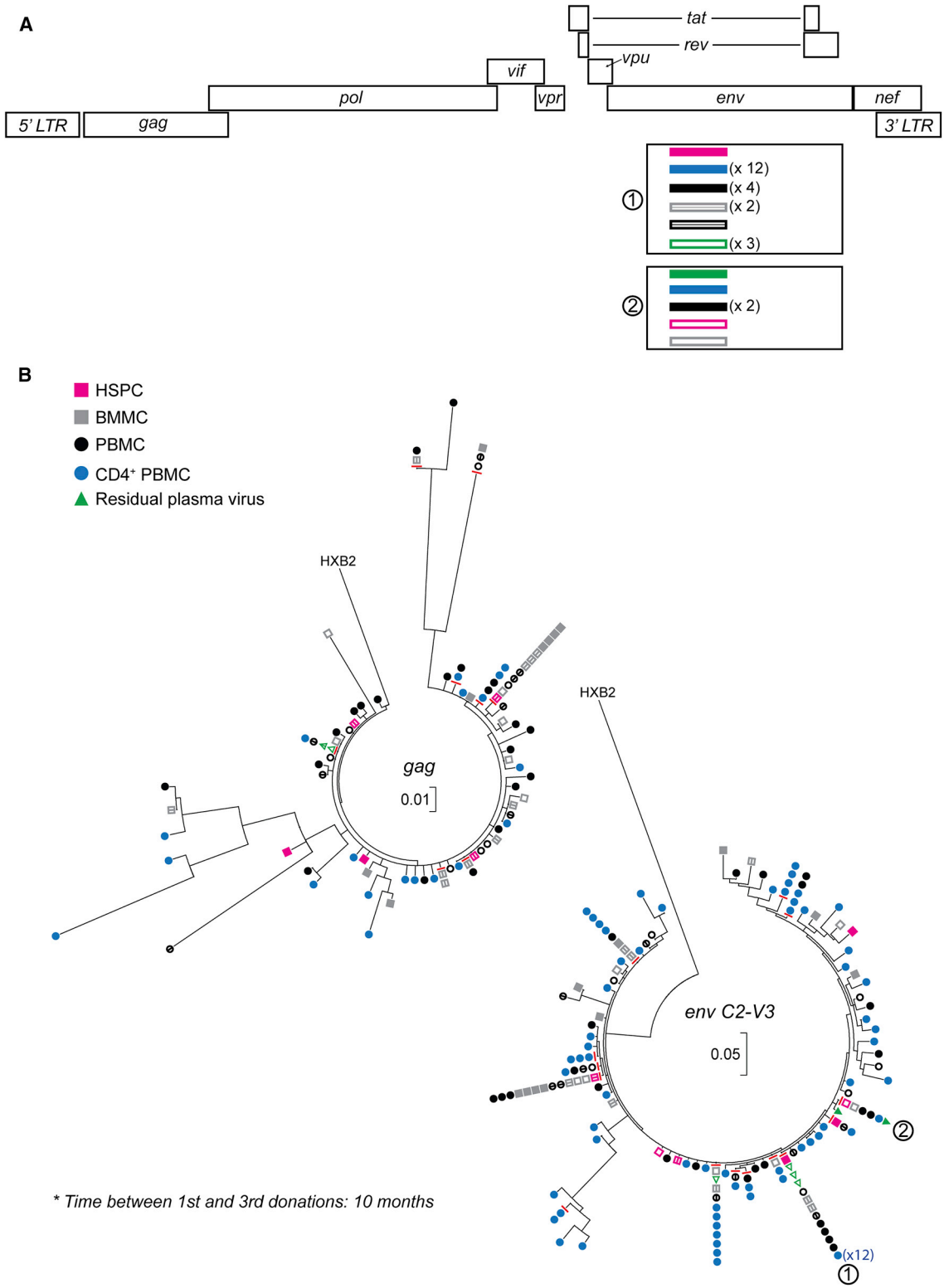


Figure 2. HSPC-Associated Sequences Match PV and Can Be Recovered from Two Donations Separated by 10 Months

(A) HIV genome map and identical viral sequences from the indicated tissue source from donor 435412406. Boxes indicate fully sequenced, genetically intact amplicons. The number of times a sequence was observed is indicated by a separate rectangle or by a number (x n).

(legend continued on next page)

identities across these amplicons occurred by chance sampling from the population was very low ($p < 10^{-6}$; Table S5) (Bui et al., 2017).

Sequence identity across entire amplicons is suggestive that PV was derived from the matching proviral genome. The likelihood that sequences identical over the C2-V3 *env* region represent identity over the entire HIV genome can be estimated at 83% using a previously published clonal prediction score (Laskey et al., 2016). In some cases, the multiplex single-genome amplification (SGA) PCR generated both an ~828-bp fragment from the 5′ long terminal repeat (LTR) into the *gag* open reading frame as well as ~429 bp of *env* C2-V3 allowing us to link two amplicons to the same genome (Figures 1, 3, 4, and S4). Linkage between the two loci was also achieved by the isolation of ~9,000-bp near-full-length genomes (Figure 3). When matching was achieved at both locations, the likelihood that the genomes were clonal was even higher. Our findings of sequence identity across multiple locations within the HIV proviral genome strongly support the conclusion that PV sequences were derived from matching proviral genomes and are consistent with our analysis showing that these identities were unlikely to have occurred by chance (Table S5).

Evidence that Residual PV Is Derived from Clusters of HSPC-Associated Identical Proviral Genomes

It was relatively common in our cohort for PV to exactly match proviral sequences from HSPCs; across the 24 donors analyzed, when unique sequences were compared, a higher proportion of proviral genomes from HSPCs matched PV (9% of *gag* and 13% of *env* amplicons) than genomes derived from PBMC or BMMC (3% *gag* and 4% *env*; $p < 0.05$) (Figure 5A). All of the HSPC-associated proviral genome sequences that matched PV also matched identical clusters of proviral genomes from PBMCs or BMMCs (Figure 5B). When these identical sequences include at least four proviral amplicons from non-HSPC cells, we refer to them as clusters of HSPC-associated identical proviruses (CHIPs). Remarkably, approximately 50% of CHIPs could be matched to PV (Figure 5B).

The impact of CHIPs is apparent when comparing Figure 5A (unique sequences only) to Figure 5C (all sequences). Because CHIPs commonly match PV, when all identical copies of provirus from BMMCs and PBMCs were included in the analysis, significant differences observed in Figure 5A were lost. These results suggest that clonally amplified sequences make important contributions to residual PV.

To further examine the relationship between clonally amplified sequences and PV, we examined a sub-group of 16 donors that had (1) clonal proviral sequences that matched an HSPC-derived provirus and (2) clonal proviral sequences for which we did not identify an HSPC match. We found that both types of "clonal" proviral sequences more frequently matched PV than sequences that were only isolated once (2% total versus 0.3% for non-clonal sequences; Figure 5D). Proviral sequences matching HSPC-derived provirus were prominent among PV-matching clonal sequences, contributing about one-half of the total clonal proviruses matching PV, despite contributing a smaller fraction of total proviruses (3% versus 9%; Figure 5D).

To avoid potential clustering effects of individual participants, we also analyzed the data by determining the proportion of residual PV matching provirus within each donor rather than pooling sequences across all donors. This analysis, shown in Figure 6A, similarly demonstrated that clonal sequences more frequently contributed to PV. Notably, the differences were most significant for sequences that were likely to be both clonal and derived from HSPCs ($p < 0.001$). These results identify a unique relationship between clonally amplified proviral genomes and residual PV. They also suggest that HSPCs are important for generating some of these clusters of amplified sequences in some donors.

Evidence that Virions Matching CHIPs Are Often Predominant in the Plasma

Five donors (434423, 436000, 449000, 454304, and 458311) had large groups of identical residual PV sequences that qualify as predominant plasma clones (PPCs). A PPC is a single sequence representing more than 50% of a large sample of independent plasma sequences from a given patient (Bailey et al., 2006). For two donors (434423 and 436000), these sequences matched CHIPs. In the remaining three donors, no cellular source of PPCs was identified (Figures 6B and 6C). Thus, CHIPs account for PPCs in at least a subset of donors.

A Signature Deletion Found in Both a CHIP and a PPC Supports HSPCs as an Original Source of Viremia

While exact matches over *gag* and *env* C2-V3 regions provide strong evidence that genomes are clonal, it remains theoretically possible that sequence differences are present elsewhere in the genome. This is an important consideration as infection by genetically similar viruses could mimic clonality that arises by cellular proliferation and expansion of genetically identical viruses. We were able to exclude coincidental infection as an explanation in one donor (Figure 1). In this donor, proviral genomes from multiple cellular sources, including HSPCs, contained a signature deletion. This deletion, which likely occurred at the time of reverse transcription, eliminated the tRNA(Lys3) primer binding site (*pbs*), the major splice donor, the dimerization initiation site, and the first two packaging stem loops. Due to the loss of the *pbs* and the major splice donor, virus generated from this proviral genome cannot initiate reverse transcription or produce Env and is thus noninfectious. However, the defects harbored by the provirus are not expected to block transcription or eliminate packaging of the RNA genome as SL3 is sufficient for RNA genome packaging (Abbink and Berkhout, 2008). Consistent with this, we identified 23 clonal

(B) Maximum-likelihood phylogenetic trees of *gag* and *env* HIV sequences from indicated tissue sources. Identical sequence groups are designated with a red bar, and the number of symbols represents the number of times the sequence was observed. Identical sequence groups depicted in (A) are marked with the circled numbers. Scale indicates nucleotide substitutions per site. Sequences from the first, second, and third donations are represented by symbols that are empty, struck-through, and filled, respectively.
See also Tables S1–S5 and Figure S1.

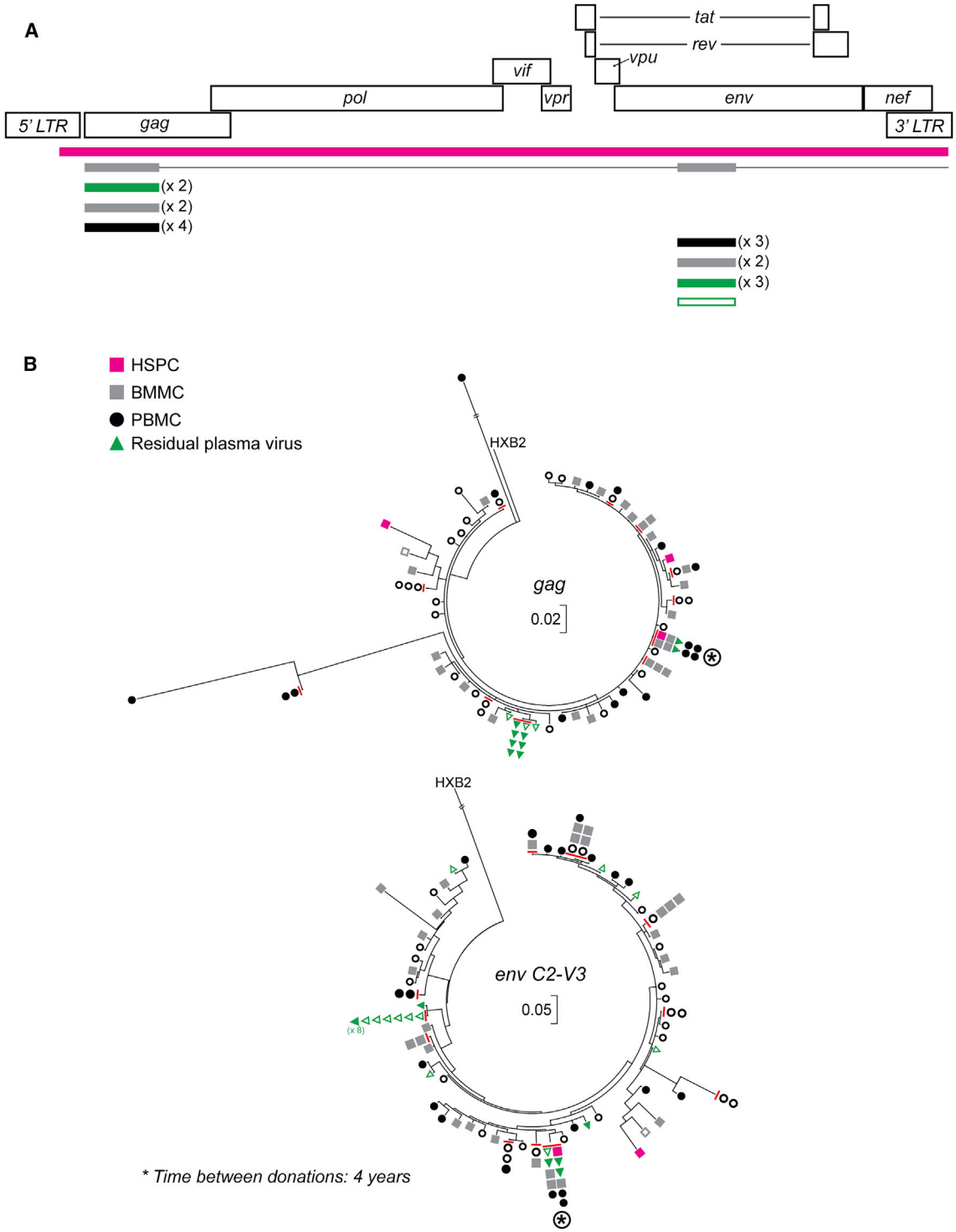


Figure 3. A Near-Full-Length Proviral Genome with Intact Open Reading Frames from HSPCs Matches PV Detected in Two Donations Separated by 3.8 Years

(A) HIV genome map and identical viral sequences from indicated tissue sources from donor 454304. The solid bars indicate fully sequenced, genetically intact amplicons. The thin line connecting *gag* and *env* amplicons indicates sequences that originated from the same first-round reaction performed at limiting dilution. The number of times a sequence was observed is indicated by a separate rectangle or by a number (x n).

(legend continued on next page)

copies of residual PV harboring the deletion (Figure 1). Additionally, a viral outgrowth assay performed using $CD4^+$ T cells isolated from BMMCs activated with phorbol 12-myristate 13-acetate (PMA) and ionomycin *ex vivo*, yielded virions containing genomic RNA with the identical signature deletion in the LTR/*gag* amplicon plus the identical C2/V3 *env* region (Figures 1A–1C). The deleted RNA was a major component of total released viral RNA sequences (~80%). The isolated genomes were derived from RNA as we obtained no sequences when reverse transcriptase was omitted (Figure 1C). This evidence supports the conclusion that the deleted virus is packaged and released from cells.

To verify this conclusion, we sub-cloned the deletion into an HIV reporter construct that requires *gag* in *trans* for virion production (Figure S6). Wild-type reporter virus was released into the supernatant 335-fold more efficiently when Gag was provided in *trans* (Figure S6B). The deletion found in the donor 436000-derived provirus, which removes the portion of the packaging signals contained in stem loops 1 and 2, resulted in a 4-fold reduction in packaging efficiency compared to wild-type, but still packaged over 80-fold more efficiently than wild-type reporter virus in the absence of Gag, demonstrating that genomic RNA containing this deletion was specifically released in Gag-containing particles (Figure 1D). A positive result in this assay depended on the addition of HIV template and reverse transcriptase, confirming that the positive signal did not originate from contamination or residual HIV DNA (Figure S6). Contaminating DNA was eliminated by the addition of DNase prior to reverse transcription, which we show efficiently eliminated signal from an HIV DNA control (Figure S6). Thus, the deleted proviral genome from donor 436000 retained the capacity to produce virions containing HIV genomic RNA.

Although the genomic RNA was packaged, it was unlikely to be infectious due to the loss of the primer binding site as well as the major splice donor. To confirm this, we measured cellular HIV mRNA and expression of the Env-GFP fusion protein expressed by this reporter construct (Figure S6). We detected HIV mRNA in cells expressing both wild-type and deleted reporter constructs (Figure S6), but only the wild-type version produced Env-GFP in transfected cells (Figure 1E). Supernatants containing virions with the deleted viral genome did not infect a $CD4^+$ T cell line, whereas control supernatants containing equal copy numbers of wild-type viral genomes were infectious in our assay (Figure 1E). These results confirm that the deleted provirus is non-infectious and that spread to multiple cells *in vivo* must have occurred by cellular proliferation of an infected progenitor cell or its progeny. To provide further proof of this conclusion, we used an established protocol to isolate HIV integration sites. We identified an integration site on chromosome 9 associated with the signature deletion. We then used chromosome 9-specific primers to confirm the location of the integration site. Using SGA PCR, we isolated multiple clonally amplified copies of the identical integration site associated with the same signature deletion (Figures 1A and S7).

HSPCs Harbor Intact, Near-Full-Length HIV Genomes

To determine whether provirus from HSPCs was infectious, we amplified near-full-length proviral genomes from HSPCs isolated from 28 donors. For this assay, we used first-round PCR primers within the viral LTRs and screened these products for near-full-length genomes using primers that amplified a ~400-bp *env* fragment in a second-round reaction. Positive first-round reactions were amplified with nested internal LTR primers and sequenced to determine whether or not each individual near-full-length genome was functional. In all, we identified 21 near-full-length genomes from 10 donors. Twenty amplicons from 9 donors were unlikely to have been due to T cell contamination of our HSPC samples based on the analysis described in Figure S1. We obtained PCR products sufficient to assess the integrity of 18 HIV genomes and fully sequenced 9. Five of 18 genomes, or 28%, had open reading frames for all 9 HIV genes (Figures 7A and 7B). Four near-full-length genomes contained intact *cis* elements in the 3′ LTR. One proviral genome from donor 409000 was missing the tRNA(Lys3) pbs (Figures 7A and S2). The four proviral genomes with intact LTR sequences were further tested to assess LTR function. As shown in Figure 7C, the reconstructed LTRs were functional based on expression of a downstream marker gene and their expression was, as expected, Tat responsive. To provide further evidence that these proviral genomes were infectious, we reconstructed the provirus isolated from HSPCs from donor 454304 and demonstrated that it produced fully infectious virus *in vitro* (Figures 7D–7F).

Strikingly, of the 5 genomes with intact open reading frames, 3 (60%) exactly matched residual PV sequences as well as proviral genomes from PBMC and BMMC. This includes the infectious proviral genome from donor 454304 in which both *env* and *gag* were identical to residual PV (Figures 3 and 7A). Donor 409000 produced 2 intact genomes, one of which had identity to a PV *env* amplicon (Figures 7A and S2). From donor 421, we found an HSPC-derived near-full-length sequence that was identical to a PV *env* amplicon (Figures 7A and S3). These intact genomes account for 10%–14% of unique residual PVs found in these donors. Moreover, sequence identity to proviral genomes from PBMCs and BMMCs provides evidence that these genomes can clonally expand *in vivo* (Figure 7).

Virions Matching HSPC Proviral Genomes Are Persistently Present in Plasma

Three of the 8 donors with identical PV and HSPC-derived viral sequences provided tissue samples on more than one occasion (435412406, 454304, and 434423). For each of these donors, we

(B) Maximum-likelihood phylogenetic trees of *gag* and *env* HIV sequences from the indicated tissue sources. Identical sequence groups are designated with a red bar, and the number of symbols represents the number of times the sequence was observed. Identical sequence groups depicted in (A) are marked with the circled asterisk. Scale indicates nucleotide substitutions per site. Sequences from the first and second donations are represented by symbols that are empty and filled, respectively.
See also Tables S1–S5 and Figure S1.

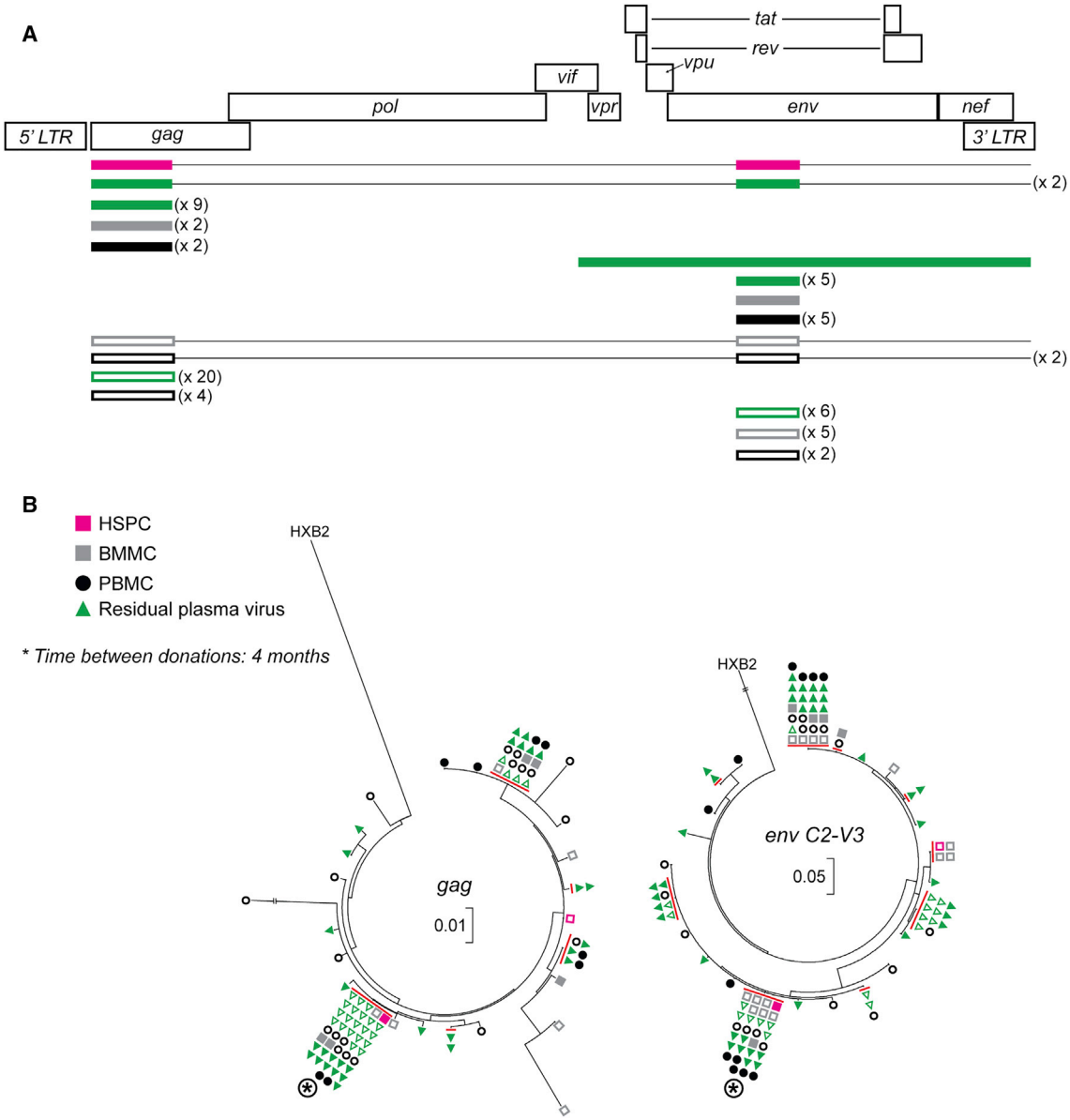


Figure 4. Multiple (*gag* and *env*) HSPC-Associated HIV Sequences Matching a Predominant Plasma Clone Are Detectable in Two Donations Separated by 4 Months

(A) HIV genome map and identical viral sequences from the indicated tissue source from donor 434423. The solid bars indicate fully sequenced, genetically intact amplicons. The thin lines connecting *gag* and *env* amplicons indicate sequences that originated from the same first-round reaction performed at limiting dilution. The number of times a sequence was observed is indicated by a separate rectangle or by a number (x n).

(B) Maximum-likelihood phylogenetic trees of *gag* and *env* HIV sequences from indicated tissue sources. Identical sequence groups are designated with a red bar, and the number of symbols represents the number of times the sequence was observed. Identical sequence groups depicted in (A) are marked with the circled asterisk. Scale indicates nucleotide substitutions per site. Sequences from the first and second donations are represented by symbols that are empty and filled, respectively.

See also Tables S1–S5 and Figure S1.

were able to demonstrate persistent matching between PV and HSPC proviral genomes over time.

Donor 435412406 contributed blood and bone marrow three times over 10 months, and we were able to detect unique proviral genomes in HSPCs from all three donations. Of note, a CHIP from the first donation matched sequences found in PV from the third donation. Conversely, a CHIP from the third donation was identical to PV from the first donation (Figure 2).

Donor 454304 had a suppressed viral load for 3.3 years at the time of the first donation and provided tissue again almost 4 years later. Described as donor 304000 in an earlier study, HSPCs from this donor's first donation had detectable HIV

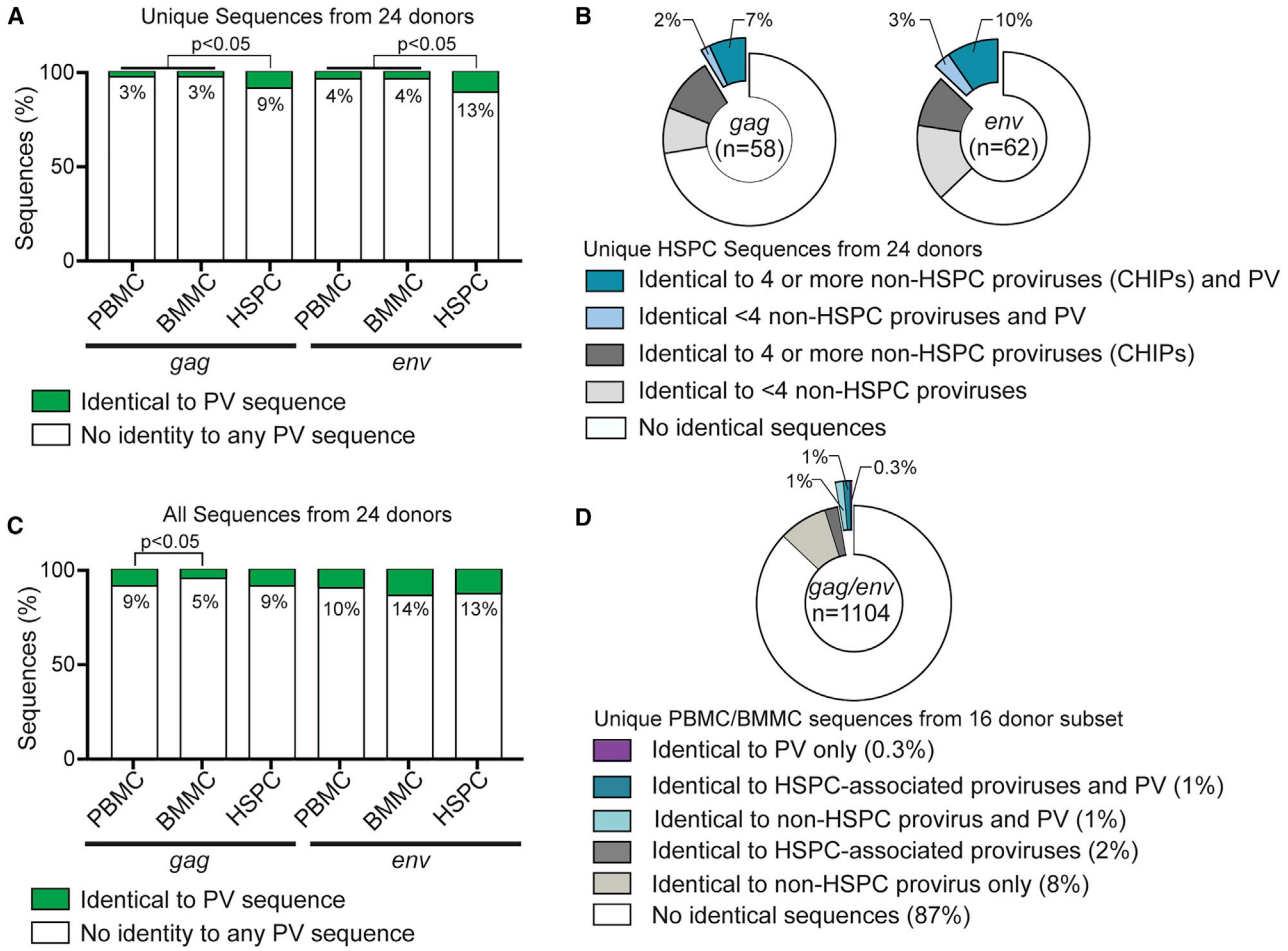


Figure 5. Residual PV Is Often Derived from Clusters of HSPC-Associated Identical Proviral Genomes

(A) Proportions of unique *gag* and *env* proviral sequences from HSPCs, PBMCs, and BMMCs with identity to plasma virus (PV). Comparison of proportions of sequence populations was performed using online calculators for two-tailed Z-test comparison of 1- or 2-sample proportions (https://www.socscistatistics.com/tests/ztest). The total number of sequences in each category is provided in Table S6.

(B) Proportions of HSPC-derived *gag* and *env* sequences with sequence identity to PV or other proviruses. CHIPs represent groups of identical proviral sequences from at least 1 HSPC-derived amplicon and 4 or more non-HSPC cell sources. The number of sequences analyzed is indicated in the center of the plot.

(C) Proportions of HSPC, PBMC, and BMMC proviral sequences that are identical to PV when all sequences are analyzed, including multiples copies of identical genomes. Analysis included sequences only from donors from whom we detected HIV nucleic acid in HSPCs and peripheral blood plasma (N = 24). (The total number of sequences in each category is provided in Table S6.)

(D) Proportions of PBMC and BMMC-derived *gag* and *env* sequences matching PV and/or (1) any other PBMC or BMMC-associated HIV sequence that was not also identical to an HSPC-associated provirus or (2) HSPC-associated proviruses. Data were derived from donors described in (A) who had both types of proviral identities (HSPC and non-HSPC; N = 16 donors). Only unique sequences are shown. The number of sequences analyzed is indicated in the center of the plot.

provirus by a *LTR/gag* qPCR assay (McNamara et al., 2013). Using our multiplex SGA, we isolated residual PV from both donations that generated *env* amplicons identical to an infectious genome (Figures 7D–7F). Two *gag* amplicons derived from residual PV from the second donation also exactly matched the functional proviral genome (Figure 3).

Donor 434423 provided blood and bone marrow twice over 4 months. A CHIP from the second donation, 434, matched large clonal clusters of residual PV for both *gag* and *env* amplicons (Figure 4). Among the reactions that were performed at limiting dilution, 2 PV PCRs produced both *gag* and *env* that were identical to HSPC-derived *gag* and *env* amplicons (Figure 4A). As discussed above, the presence of matching *gag* and *env* sequences from multiplex SGA reactions for both HSPC and PV strongly supports identity across the entire genome. These examples provide strong support for the conclusion that HSPC-associated genomes contribute to persistent residual viremia.

DISCUSSION

Here, we confirm a prior study providing evidence that HSPCs propagate HIV proviruses by cellular proliferation (Sebastian et al., 2017) and provide evidence that clonally expanded

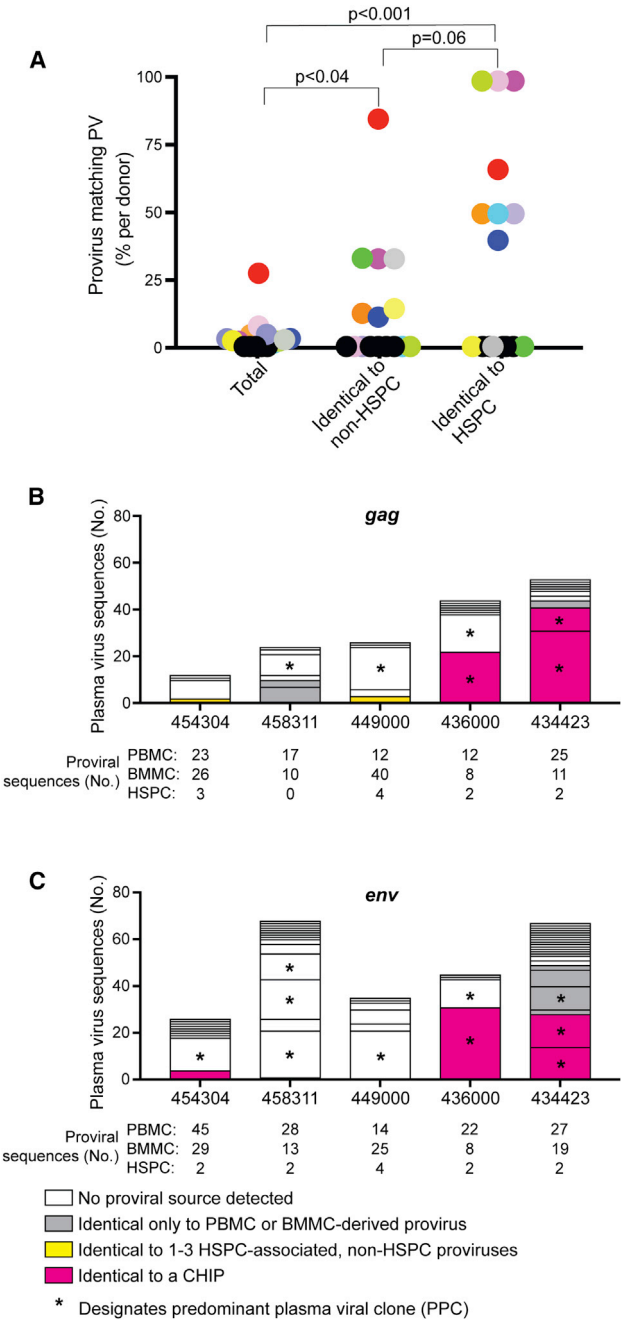


Figure 6. Proviral Genomes that Are Likely Clonal More Frequently Match Residual PV, and Predominant Plasma Clones that Can Be Attributed to a Cellular Source Often Match CHIPs

(A) Graphical analysis showing the relationship between PBMC and BMMC proviral genomes that are identical to other sequences (non-HSPC or HSPC proviral genomes) and PV. Data were derived from donors described in Figure 5D who had both types of proviral identities (HSPC and non-HSPC; N = 16). Individual donors with non-zero values are indicated by colored symbols. Only unique *gag* and *env* sequences were included. p values were determined using two-tailed Wilcoxon matched pairs signed-rank test.

(B and C) Bar graphs showing number of (B) *gag* and (C) *env* PV amplicons that matched amplicons from the indicated groups of cellular proviruses from the 5 donors from whom we identified predominant plasma clones (PPCs). Groups of identical PV sequences are indicated by horizontal bars within each stacked bar graph. Groups of PPCs are indicated with an asterisk. Number of proviral sequences analyzed is shown below x axis donor labels.

See also Tables S1–S5 and Figures S2–S5.

proviruses from HSPC progeny contribute to residual PV. Proviral genome sequences from HSPCs were often identical to sequences isolated from PBMCs and BMMCs. In two donors, predominant plasma viral clones exactly matched HSPC-associated sequences. Based on these studies, we propose that HSPCs and their daughter cells containing clonal genomes are capable of producing clonal virus that can predominate in the plasma. Determining the exact identity of the infected daughter cells will require additional studies. In a prior investigation, we found that HSPCs could pass genomes to both CD4-positive and CD4-negative progeny in some donors (Sebastian et al., 2017). These relatively rare occurrences are consistent with our model that HSPCs can spread HIV genomes by cell division and differentiation.

Although viremia is relatively low during cART, it likely requires substantial HIV production from infected cells. Thus, it is unlikely that HSPCs alone, which represent ~1% of BMMCS, are directly responsible for this virus production. Rather, we propose that a subset of PV is derived from clonal proviral sequences found in both HSPCs and PBMC/BMMC populations (clusters of HSPC-associated identical proviruses [CHIPs]). In this model, HSPCs themselves remain latently infected, preserving their longevity and ensuring viral persistence. HPSCs contribute to residual PV indirectly via the generation of progeny containing clonal proviral genomes that become activated to produce PV. Consistent with this, most HSPCs are maintained in a quiescent state *in vivo*, which we have shown is associated with viral latency *in vitro* (Painter et al., 2017).

Although HSPCs represent a small population of cells, in a subset of donors, HSPC-associated proviral genomes were a quantitatively more important source of residual PV than proviral genomes that were not associated with HSPCs. Future studies are needed to understand the mechanism behind this observation. It could reflect lower cytopathicity of the virus in HSPCs compared to T cells. Unlike T cell models of latency, latency in HSPCs is immediately established upon integration (Carter et al., 2010, 2011; McNamara et al., 2012; Painter et al., 2017; Sebastian et al., 2017; Zaikos et al., 2018). This mechanism avoids cytotoxicity due to transient viral protein expression. In contrast, most T cell models require a period of active viral expression followed by the establishment of latency with return of the T cell to a quiescent state (Bosque and Planelles, 2009; Kim et al., 2014). Another possibility is that most virus-producing T cells are non-circulating, hence untested here. However, if that were the case, we would expect to have detected a greater variety of viruses in the plasma. For example, the CHIP containing a deleted primer binding site was one of 8 *gag* PV sequences isolated and represented 51% of the *gag* amplicons we obtained (23 of 45 *gag* amplicons). For *env* amplicons, this clone was one of 4 *env* PV sequences and represented 67% of *env* amplicons (31 of 45 *env* amplicons). Bailey et al. (2006)

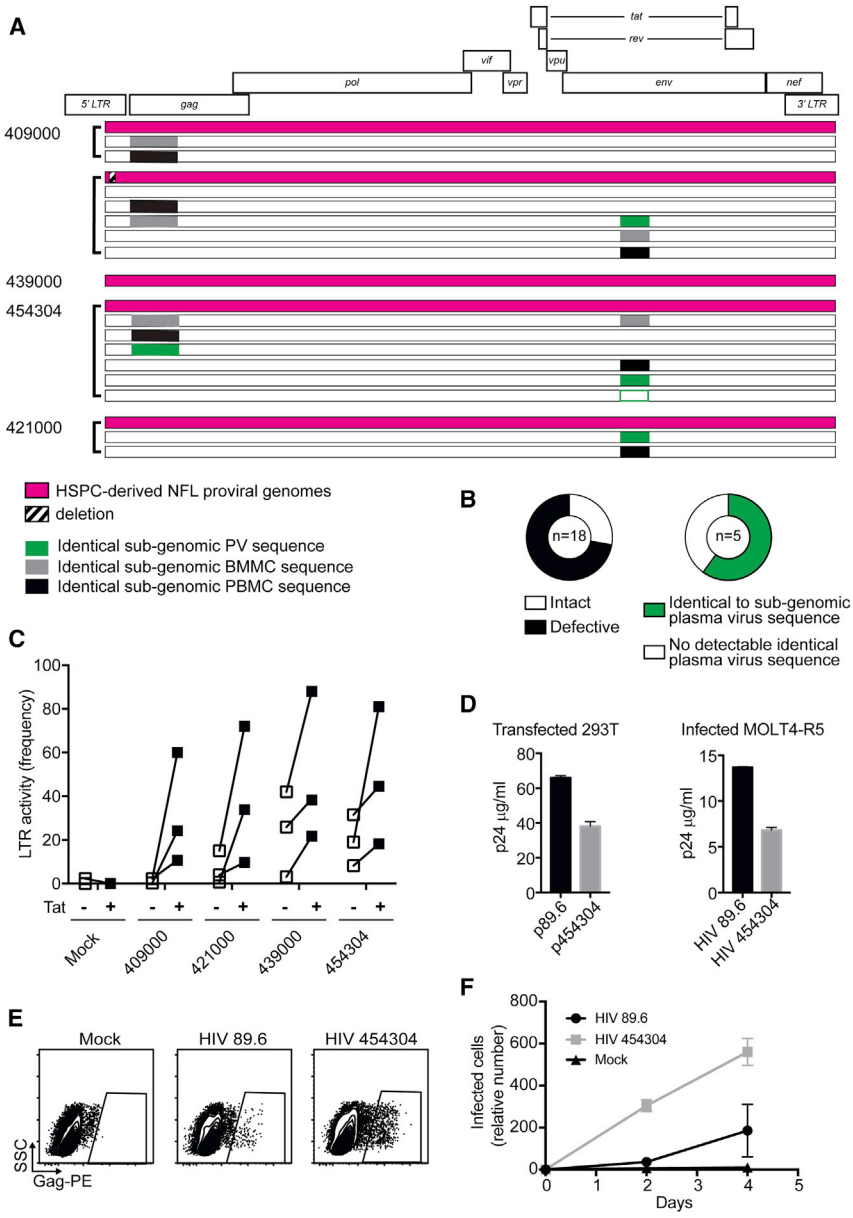


Figure 7. HSPCs Contain Proviral HIV Genomes with Complete Open Reading Frames, Active LTRs, and Infectious Virus that Are Represented in Residual PV with Relatively High Frequency

(A) HIV genome map and near-full-length proviral sequences from HSPCs and genetically identical sub-genomic amplicons from other tissue sources. All amplicons represented are from PCRs that were performed at limiting dilution. The tissue or cell source of a sequence is indicated by a separate rectangle. Donor IDs for groups of amplicons are indicated on the left. For donor 454304, amplicons from the first and second donations are represented by empty and filled bars, respectively.

(B) Pie graphs showing the proportion of intact near-full-length HIV genomes from HSPCs (left) and the proportion of these intact sequences that are genetically identical to sub-genomic amplicons from residual PV (right). The number of sequences analyzed is indicated in the center of the plot.

(C) Summary graph depicting the activity of reconstructed LTR sequences from near-full-length genomes isolated from HSPCs as assessed by the frequency of transfected 293T cells expressing a marker gene (GFP) that was under the control of the LTR. Responsiveness to Tat was assessed by co-transfection of a Tat-expressing plasmid where indicated. Three biological replicates are shown. The frequency of GFP-expressing cells are presented for each replicate.

(D) Summary graph of p24 production by the indicated molecular clone following transfection of 293T cells (left panel) or infection of MOLT4-R5 T cells (right panel). Three technical replicates of p24 ELISAs for three technical replicate infections of MOLT4-R5 are shown.

(E) Representative flow-cytometric analysis of MOLT4-R5 cells infected with the indicated virus produced in MOLT4-R5 cells from (D) and stained for intracellular Gag 4 days post-infection.

(F) Summary graph showing relative number of infected cells over time from the infection described in (E). After staining, all wells were resuspended in a fixed volume of fluorescence-activated cell-sorting (FACS) buffer and collected on a BD FACSCAN DxP for 40 s so relative numbers of cells could be compared across wells. MOLT4-R5 infection was performed in three technical replicates, and the mean number of cells expressing intracellular Gag for each replicate is displayed. Error bars are SD.

similarly observed dominant viral clones for most patients they examined on long-term cART. More recently, Wang et al. (2018) performed longitudinal sampling of residual PV *env* amplicons. In four of eight individuals, they observed dominant PV populations. The evidence showing only a few latently infected clones (or progeny thereof) contributing to the bulk of residual viremia in some donors supports the conclusion that amplified clonal populations are an important source of residual PV.

Clonally expanded T cells have been described, and it is important to consider these cells as a potential source of clonal residual PV whether or not they are progeny of an infected HSPC (Bruner et al., 2016; Hiener et al., 2017; Lee et al., 2017; Maldarelli et al., 2014; Simonetti et al., 2016; Sun et al., 2015; Wagner et al., 2014). Interestingly, a recent study suggests that these clonal populations can wax and wane over time (Wang et al., 2018). In several publications, expansion of clonal T cell populations was explained by insertion of the provirus into genes that regulate cell growth and/or by antigenic stimulation of the T cell receptor (Cohn et al., 2015; Maldarelli et al., 2014; Simonetti et al., 2016). We report a clonally amplified provirus that likely originated from an HSPC. This proviral genome was integrated into a region of chromosome 9 that is devoid of known genes (Figures 1 and S7). Amplification of this provirus by cellular proliferation is unlikely to be due to insertion into genes that regulate growth. Thus, we provide an example of clonal expansion of

an HIV-infected cell due to the innate proliferative capacity of the infected cell(s).

To the best of our knowledge, the frequency with which clonally expanded populations contribute to residual PV has not been reported. In our study, a significantly higher frequency of clonally expanded proviral genomes matched residual PV, and this difference was most significant when the clonally expanded proviral populations were likely derived from an HSPC. The evidence from our cohort supports the conclusion that HSPC-derived clonal populations are an important source of PV, at least in some donors, but are unlikely to be the only source of PV.

We also provide evidence that HSPC-associated HIV genomes are functional. A relatively high percentage of proviral genomes isolated from HSPCs contained complete open reading frames for all 9 viral genes (28%). This compares to 2%–12% for $CD4^+$ T cells (Bruner et al., 2016; Hiener et al., 2017; Ho et al., 2013). All of the intact proviral genomes tested had LTRs that were functional *in vitro*. Moreover, we confirmed that one of these genomes was infectious by showing that the fully reconstructed provirus encoded infectious virus when tested *in vitro*. Remarkably, HSPC-derived near-full-length genomes often matched PV (60%). This "*in vivo* outgrowth assay" provides strong evidence that HSPCs can harbor infectious HIV.

The relative paucity of HSPC-derived HIV sequences compared to those of PBMCs and BMMCs may have led to under-sampling of the genomes harbored within HSPCs. Given this, it is even more remarkable that HSPC-associated genomes matched PV more frequently than unique genomes from PBMCs or BMMCS. Based on the frequency with which *gag* and *env* PBMC-derived amplicons matched PV sequences (3% and 4%, respectively), we would have expected that only 2 (*gag*) and 3 (*env*) HSPC-derived proviral amplicons matched PV sequences. Instead, we found 5 (*gag*) and 8 (*env*) HSPC-derived amplicons matched PV. Based on the observed 3 (*gag*) and 5 (*env*) additional matching HSPC-derived amplicons and the expected proportion of matching sequences (3% and 4%), the probability that under-sampling overestimated that frequency of HSPC genomes matching PV is extremely low ($<10^{-4}$).

Determining whether genomes with identical sequences across an amplified segment are truly clonal can be challenging. It is theoretically possible that identical sub-genomic proviral sequences harbored by HSPCs and other cell types were derived from genetically similar viruses rather than truly clonal ones. However, this is an unlikely explanation for our results for a number of reasons: (1) identity was observed across the entire variable C2-V3 *env* amplicon, which is associated with a clonal prediction score of 83% (Laskey et al., 2016); (2) we were often able to exactly match both *gag* and *env* amplicons from the same first-round SGA PCR, increasing our confidence in clonality; (3) one CHIP that matched a PPC contained a signature deletion that removed the primer binding site (an 18-bp sequence complementary to tRNA[Lys3] that is used to prime reverse transcription), the major splice donor, and two stem loops. This deletion caused the virus to be non-infectious, allowing the conclusion that this viral genome was amplified by cellular proliferation.

Multiple defects are present in the non-infectious genome shown in Figure 1 (loss of the primer binding site, the major splice donor, and two stem loops). However, our evidence indicates deletion of the primer binding site, which is needed for reverse transcription, is the dominant mechanism for loss of infectivity. Deletion of two stem loops reduced but did not prevent viral genome packaging and virion release. Loss of the major splice donor did not prevent *in vivo* expression or detection of viral RNA in *ex vivo* outgrowth assays. The absence of the splice donor element is expected to affect expression of factors (Tat and Rev) that are needed for optimal expression of Gag and the packaged genome. Residual expression could occur from low-level Tat and Rev-independent transcription and nuclear export. It is also possible that alternative splice donor sites present in the proviral genome or in nearby genomic sequences rescue Tat and Rev expression. Importantly, the virus remained defective for infection even when Tat and Rev were provided in *trans*. Thus, the primer binding site defect is sufficient to render the virus non-infectious. The presence of this defective proviral genome in multiple cells provides strong evidence for spread of HIV genomes via cellular proliferation from an infected progenitor. The isolation of multiple copies of identical integration sites containing this signature deletion confirms our conclusion.

Loss of the tRNA binding site sequence by a genome that is producing large amounts of virus raises the possibility that this genetic element could regulate viral gene expression. Interestingly, a number of studies have shown that transcriptional repressors utilize tRNA binding sites in other retroviruses to restrict viral transcription. Specifically, this element is required for transcriptional silencing of murine leukemia viruses (MLVs) that have integrated into genomes of embryonic stem cells. The repression is largely mediated by *trans*-acting factors that recognize the primer binding site (Wolf and Goff, 2007, 2008, 2009; Wolf et al., 2008a, 2008b). Whether a related *trans*-acting factor might similarly regulate HIV gene expression via the HIV tRNA binding site in cells with stem cell-like characteristics is not yet known. Here, we report that PMA and ionomycin treatment is required for release of the deleted virus from donor $CD4^+$ T cells cultured *ex vivo*. The apparent absence of viral gene expression by unstimulated cells cultured *ex vivo* is discordant with the relatively high level of gene expression observed *in vivo*. The explanation for this intriguing observation remains elusive and will require further study. It is possible that tissue-resident progeny not sampled in this study maintain a different level of HIV gene expression than circulating $CD4^+$ T cells.

In summary, our work provides evidence that HSPCs are a functionally significant reservoir of persistent HIV infection. Proviral genomes from HSPCs are associated with expanded clonal HIV proviral genomes found in peripheral blood and bone marrow that contribute to residual PV. These studies shed light on how the latent reservoir is maintained *in vivo* and suggest a need for novel therapeutic options. Selective targeting of HIV-infected progenitor cells may reduce the number of clonally expanded HIV proviral genomes, promote a reduction of the HIV reservoir, and limit residual and rebound PV.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Human subjects
  - Cell lines
- METHOD DETAILS
  - Donor cell isolation and fractionation
  - Plasma virus and peripheral blood mononuclear cell (PBMC) isolation
  - Flow cytometry and antibodies
  - DNA extraction
  - RNA isolation
  - Reverse transcription
  - PCR contamination control procedures
  - SGA
  - Quantitative reverse transcription PCR (qRT-PCR)
  - Construction of reporter constructs and infectious proviral genomes
  - Transfections and virus production
  - HIV p24 ELISA
  - HIV transduction and infection of T cell lines
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Donor and sample exclusion criteria
  - Phylogenetic Analysis
  - Statistical analyses
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and seven tables and can be found with this article online at https://doi.org/10.1016/j.celrep.2018.11.104.

ACKNOWLEDGMENTS

This research was supported by the NIH (R01AI096962 to K.L.C., T32GM007863 to N.T.S.K., T32AI007413 to M.M.P., and T32GM007315 to N.T.S.K. and M.C.V.), National Center for Research Resources (UL1RR024986 to A.O.-N.), National Science Foundation (DGE 0718128 to L.A.M.), and The Burroughs Wellcome Fund (to K.L.C.). T.D.Z. was supported by a fellowship from HHMI. Cores we used also were supported by funds as follows: Flow Cytometry Core and DNA Sequencing Core (NIH grant P30CAO46592), as well as the Michigan Clinical Research Unit (CTSA grant 2UL1TR000433-06). We are grateful to Mary Reyes, Lisa Mac, and Justin Bell for recruitment of donors and help with regulatory documentation, and Heather Fox and Henry Ford Hospital physicians for the bone marrow aspirations, and we especially thank the donors themselves. We thank the University of Michigan Clinical Research Unit. We thank Ryan Yucha and Steve King for technical assistance and John Moran for helpful scientific discussions.

AUTHOR CONTRIBUTIONS

Conceptualization, K.L.C. and T.D.Z.; Methodology, V.H.T., N.T.S.K., F.T., A.O.-N., M.C.V., J.L., M.M.P., and L.A.M.; Investigation, T.D.Z., V.H.T., N.T.S.K., F.T., A.O.-N., J.L., M.M.P., A.N., and L.A.M.; Validation, T.D.Z., V.H.T., N.T.S.K., F.T., A.O.-N., M.C.V., J.L., M.M.P., A.N., and L.A.M.a; Visualization, K.L.C., T.D.Z., and V.H.T.; Formal Analysis, T.D.Z., V.H.T., and K.L.C.; Writing – Original Draft, T.D.Z., V.H.T., and K.L.C.; Writing – Review & Editing, K.L.C., T.D.Z., V.H.T., N.T.S.K., F.T., A.O.-N., J.R., N.M., M.C.V., J.L., M.M.P., and L.A.M.; Funding Acquisition, K.L.C., T.D.Z., L.A.M., and A.O.-N.; Resources, J.R., D.B., and N.M.; Supervision, K.L.C., N.M., and J.R.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 26, 2018
Revised: September 12, 2018
Accepted: November 29, 2018
Published: December 26, 2018

SUPPORTING CITATIONS

The following references appear in the Supplemental Information: Añez et al. (2008); Salvi et al. (1998).

REFERENCES

Abbink, T.E., and Berkhout, B. (2008). RNA structure modulates splicing efficiency at the human immunodeficiency virus type 1 major splice donor. J. Virol. *82*, 3090–3098.

Anderson, J.A., Archin, N.M., Ince, W., Parker, D., Wiegand, A., Coffin, J.M., Kuruc, J., Eron, J., Swanstrom, R., and Margolis, D.M. (2011). Clonal sequences recovered from plasma from patients with residual HIV-1 viremia and on intensified antiretroviral therapy are identical to replicating viral RNAs recovered from circulating resting CD4$^+$ T cells. J. Virol. *85*, 5220–5223.

Añez, M., Putonti, C., Fox, G.E., Fofanov, Y., and Willson, R.C. (2008). Exhaustive computational identification of pathogen sequences far-distant from background genomes: identification and experimental verification of human-blind dengue PCR primers. J. Biotechnol. *133*, 267–276.

Araínga, M., Edagwa, B., Mosley, R.L., Poluektova, L.Y., Gorantla, S., and Gendelman, H.E. (2017). A mature macrophage is a principal HIV-1 cellular reservoir in humanized mice after treatment with long acting antiretroviral therapy. Retrovirology *14*, 17.

Avalos, C.R., Price, S.L., Forsyth, E.R., Pin, J.N., Shirk, E.N., Bullock, B.T., Queen, S.E., Li, M., Gellerup, D., O'Connor, S.L., et al. (2016). Quantitation of productively infected monocytes and macrophages of simian immunodeficiency virus-infected macaques. J. Virol. *90*, 5643–5656.

Bailey, J.R., Sedaghat, A.R., Kieffer, T., Brennan, T., Lee, P.K., Wind-Rotolo, M., Haggerty, C.M., Kamireddi, A.R., Liu, Y., Lee, J., et al. (2006). Residual human immunodeficiency virus type 1 viremia in some patients on antiretroviral therapy is dominated by a small number of invariant clones rarely found in circulating CD4$^+$ T cells. J. Virol. *80*, 6441–6457.

Bosque, A., and Planelles, V. (2009). Induction of HIV-1 latency and reactivation in primary memory CD4$^+$ T cells. Blood *113*, 58–65.

Brennan, T.P., Woods, J.O., Sedaghat, A.R., Siliciano, J.D., Siliciano, R.F., and Wilke, C.O. (2009). Analysis of human immunodeficiency virus type 1 viremia and provirus in resting CD4+ T cells reveals a novel source of residual viremia in patients on antiretroviral therapy. J. Virol. *83*, 8470–8481.

Bruner, K.M., Murray, A.J., Pollack, R.A., Soliman, M.G., Laskey, S.B., Capoferri, A.A., Lai, J., Strain, M.C., Lada, S.M., Hoh, R., et al. (2016). Defective proviruses rapidly accumulate during acute HIV-1 infection. Nat. Med. *22*, 1043–1049.

Bui, J.K., Sobolewski, M.D., Keele, B.F., Spindler, J., Musick, A., Wiegand, A., Luke, B.T., Shao, W., Hughes, S.H., Coffin, J.M., et al. (2017). Proviruses with identical sequences comprise a large fraction of the replication-competent HIV reservoir. PLoS Pathog. *13*, e1006283.

Bullen, C.K., Laird, G.M., Durand, C.M., Siliciano, J.D., and Siliciano, R.F. (2014). New ex vivo approaches distinguish effective and ineffective single agents for reversing HIV-1 latency in vivo. Nat. Med. *20*, 425–429.

Buzon, M.J., Sun, H., Li, C., Shaw, A., Seiss, K., Ouyang, Z., Martin-Gayo, E., Leng, J., Henrich, T.J., Li, J.Z., et al. (2014). HIV-1 persistence in CD4$^+$ T cells with stem cell-like properties. Nat. Med. *20*, 139–142.

Carter, C.C., Onafuwa-Nuga, A., McNamara, L.A., Riddell, J., 4th, Bixby, D., Savona, M.R., and Collins, K.L. (2010). HIV-1 infects multipotent progenitor cells causing cell death and establishing latent cellular reservoirs. Nat. Med. *16*, 446–451.

Carter, C.C., McNamara, L.A., Onafuwa-Nuga, A., Shackleton, M., Riddell, J., 4th, Bixby, D., Savona, M.R., Morrison, S.J., and Collins, K.L. (2011). HIV-1 utilizes the CXCR4 chemokine receptor to infect multipotent hematopoietic stem and progenitor cells. Cell Host Microbe *9*, 223–234.

Chun, T.W., Carruth, L., Finzi, D., Shen, X., DiGiuseppe, J.A., Taylor, H., Hermankova, M., Chadwick, K., Margolick, J., Quinn, T.C., et al. (1997). Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. Nature *387*, 183–188.

Chun, T.W., Davey, R.T., Jr., Ostrowski, M., Shawn Justement, J., Engel, D., Mullins, J.I., and Fauci, A.S. (2000). Relationship between pre-existing viral reservoirs and the re-emergence of plasma viremia after discontinuation of highly active anti-retroviral therapy. Nat. Med. *6*, 757–761.

Cohn, L.B., Silva, I.T., Oliveira, T.Y., Rosales, R.A., Parrish, E.H., Learn, G.H., Hahn, B.H., Czartoski, J.L., McElrath, M.J., Lehmann, C., et al. (2015). HIV-1 integration landscape during latent and active infection. Cell *160*, 420–432.

Davey, R.T., Jr., Bhat, N., Yoder, C., Chun, T.W., Metcalf, J.A., Dewar, R., Natarajan, V., Lempicki, R.A., Adelsberger, J.W., Miller, K.D., et al. (1999). HIV-1 and T cell dynamics after interruption of highly active antiretroviral therapy (HAART) in patients with a history of sustained viral suppression. Proc. Natl. Acad. Sci. USA *96*, 15109–15114.

Eisele, E., and Siliciano, R.F. (2012). Redefining the viral reservoirs that prevent HIV-1 eradication. Immunity *37*, 377–388.

Finzi, D., Hermankova, M., Pierson, T., Carruth, L.M., Buck, C., Chaisson, R.E., Quinn, T.C., Chadwick, K., Margolick, J., Brookmeyer, R., et al. (1997). Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. Science *278*, 1295–1300.

Finzi, D., Blankson, J., Siliciano, J.D., Margolick, J.B., Chadwick, K., Pierson, T., Smith, K., Lisziewicz, J., Lori, F., Flexner, C., et al. (1999). 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.

Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. *22*, 160–174.

Hiener, B., Horsburgh, B.A., Eden, J.S., Barton, K., Schlub, T.E., Lee, E., von Stockenstrom, S., Odevall, L., Milush, J.M., Liegler, T., et al. (2017). Identification of genetically intact HIV-1 proviruses in specific CD4$^+$ T cells from effectively treated participants. Cell Rep. *21*, 813–822.

Ho, Y.C., Shan, L., Hosmane, N.N., Wang, J., Laskey, S.B., Rosenbloom, D.I., Lai, J., Blankson, J.N., Siliciano, J.D., and Siliciano, R.F. (2013). Replication-competent noninduced proviruses in the latent reservoir increase barrier to HIV-1 cure. Cell *155*, 540–551.

Honeycutt, J.B., Thayer, W.O., Baker, C.E., Ribeiro, R.M., Lada, S.M., Cao, Y., Cleary, R.A., Hudgens, M.G., Richman, D.D., and Garcia, J.V. (2017). HIV persistence in tissue macrophages of humanized myeloid-only mice during antiretroviral therapy. Nat. Med. *23*, 638–643.

Imamichi, H., Dewar, R.L., Adelsberger, J.W., Rehm, C.A., O'Doherty, U., Paxinos, E.E., Fauci, A.S., and Lane, H.C. (2016). Defective HIV-1 proviruses produce novel protein-coding RNA species in HIV-infected patients on combination antiretroviral therapy. Proc. Natl. Acad. Sci. USA *113*, 8783–8788.

Kearney, M.F., Spindler, J., Shao, W., Yu, S., Anderson, E.M., O'Shea, A., Rehm, C., Poethke, C., Kovacs, N., Mellors, J.W., et al. (2014). Lack of detectable HIV-1 molecular evolution during suppressive antiretroviral therapy. PLoS Pathog. *10*, e1004010.

Kim, M., Hosmane, N.N., Bullen, C.K., Capoferri, A., Yang, H.C., Siliciano, J.D., and Siliciano, R.F. (2014). A primary CD4$^+$ T cell model of HIV-1 latency established after activation through the T cell receptor and subsequent return to quiescence. Nat. Protoc. *9*, 2755–2770.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol. Biol. Evol. *33*, 1870–1874.

Laird, G.M., Bullen, C.K., Rosenbloom, D.I., Martin, A.R., Hill, A.L., Durand, C.M., Siliciano, J.D., and Siliciano, R.F. (2015). Ex vivo analysis identifies effective HIV-1 latency-reversing drug combinations. J. Clin. Invest. *125*, 1901–1912.

Laskey, S.B., Pohlmeyer, C.W., Bruner, K.M., and Siliciano, R.F. (2016). Evaluating clonal expansion of HIV-infected cells: optimization of PCR strategies to predict clonality. PLoS Pathog. *12*, e1005689.

Lee, G.Q., Orlova-Fink, N., Einkauf, K., Chowdhury, F.Z., Sun, X., Harrington, S., Kuo, H.H., Hua, S., Chen, H.R., Ouyang, Z., et al. (2017). Clonal expansion of genome-intact HIV-1 in functionally polarized Th1 CD4$^+$ T cells. J. Clin. Invest. *127*, 2689–2696.

Liu, H., Dow, E.C., Arora, R., Kimata, J.T., Bull, L.M., Arduino, R.C., and Rice, A.P. (2006). Integration of human immunodeficiency virus type 1 in untreated infection occurs preferentially within genes. J. Virol. *80*, 7765–7768.

Maldarelli, F., Wu, X., Su, L., Simonetti, F.R., Shao, W., Hill, S., Spindler, J., Ferris, A.L., Mellors, J.W., Kearney, M.F., et al. (2014). HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. Science *345*, 179–183.

McNamara, L.A., Ganesh, J.A., and Collins, K.L. (2012). Latent HIV-1 infection occurs in multiple subsets of hematopoietic progenitor cells and is reversed by NF-κB activation. J. Virol. *86*, 9337–9350.

McNamara, L.A., Onafuwa-Nuga, A., Sebastian, N.T., Riddell, J., 4th, Bixby, D., and Collins, K.L. (2013). CD133$^+$ hematopoietic progenitor cells harbor HIV genomes in a subset of optimally treated people with long-term viral suppression. J. Infect. Dis. *207*, 1807–1816.

Painter, M.M., Zaikos, T.D., and Collins, K.L. (2017). Quiescence promotes latent HIV infection and resistance to reactivation from latency with histone deacetylase inhibitors. J. Virol. *91*, e01080-17.

Sahu, G.K., Paar, D., Frost, S.D., Smith, M.M., Weaver, S., and Cloyd, M.W. (2009). Low-level plasma HIVs in patients on prolonged suppressive highly active antiretroviral therapy are produced mostly by cells other than CD4 T-cells. J. Med. Virol. *81*, 9–15.

Salvi, R., Garbuglia, A.R., Di Caro, A., Pulciani, S., Montella, F., and Benedetto, A. (1998). Grossly defective nef gene sequences in a human immunodeficiency virus type 1-seropositive long-term nonprogressor. J. Virol. *72*, 3646–3657.

Sebastian, N.T., Zaikos, T.D., Terry, V., Taschuk, F., McNamara, L.A., Onafuwa-Nuga, A., Yucha, R., Signer, R.A.J., Riddell, J., IV, Bixby, D., et al. (2017). CD4 is expressed on a heterogeneous subset of hematopoietic progenitors, which persistently harbor CXCR4 and CCR5-tropic HIV proviral genomes in vivo. PLoS Pathog. *13*, e1006509.

Simonetti, F.R., Sobolewski, M.D., Fyne, E., Shao, W., Spindler, J., Hattori, J., Anderson, E.M., Watters, S.A., Hill, S., Wu, X., et al. (2016). Clonally expanded CD4$^+$ T cells can produce infectious HIV-1 in vivo. Proc. Natl. Acad. Sci. USA *113*, 1883–1888.

Sun, H., Kim, D., Li, X., Kiselinova, M., Ouyang, Z., Vandekerckhove, L., Shang, H., Rosenberg, E.S., Yu, X.G., and Lichterfeld, M. (2015). Th1/17 polarization of CD4 T cells supports HIV-1 persistence during antiretroviral therapy. J. Virol. *89*, 11284–11293.

Sundstrom, J.B., Ellis, J.E., Hair, G.A., Kirshenbaum, A.S., Metcalfe, D.D., Yi, H., Cardona, A.C., Lindsay, M.K., and Ansari, A.A. (2007). Human tissue mast cells are an inducible reservoir of persistent HIV infection. Blood *109*, 5293–5300.

Wagner, T.A., McLaughlin, S., Garg, K., Cheung, C.Y., Larsen, B.B., Styrchak, S., Huang, H.C., Edlefsen, P.T., Mullins, J.I., and Frenkel, L.M. (2014). HIV

latency. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. Science *345*, 570–573.

Wang, Z., Gurule, E.E., Brennan, T.P., Gerold, J.M., Kwon, K.J., Hosmane, N.N., Kumar, M.R., Beg, S.A., Capoferri, A.A., Ray, S.C., et al. (2018). Expanded cellular clones carrying replication-competent HIV-1 persist, wax, and wane. Proc. Natl. Acad. Sci. USA *115*, E2575–E2584.

Wolf, D., and Goff, S.P. (2007). TRIM28 mediates primer binding site-targeted silencing of murine leukemia virus in embryonic cells. Cell *131*, 46–57.

Wolf, D., and Goff, S.P. (2008). Host restriction factors blocking retroviral replication. Annu. Rev. Genet. *42*, 143–163.

Wolf, D., and Goff, S.P. (2009). Embryonic stem cells use ZFP809 to silence retroviral DNAs. Nature *458*, 1201–1204.

Wolf, D., Cammas, F., Losson, R., and Goff, S.P. (2008a). Primer binding site-dependent restriction of murine leukemia virus requires HP1 binding by TRIM28. J. Virol. *82*, 4675–4679.

Wolf, D., Hug, K., and Goff, S.P. (2008b). TRIM28 mediates primer binding site-targeted silencing of Lys1,2 tRNA-utilizing retroviruses in embryonic cells. Proc. Natl. Acad. Sci. USA *105*, 12521–12526.

Zaikos, T.D., Painter, M.M., Sebastian, N.T., Terry, V.H., and Collins, K.L. (2018). Class 1-selective histone deacetylase inhibitors enhance HIV latency reversal while preserving the activity of HDAC isoforms necessary for maximal HIV gene expression. J. Virol *92*, e02110–17.

Zhu, T., Muthui, D., Holte, S., Nickle, D., Feng, F., Brodie, S., Hwangbo, Y., Mullins, J.I., and Corey, L. (2002). Evidence for human immunodeficiency virus type 1 replication in vivo in $CD14^+$ monocytes and its potential role as a source of virus in patients on highly active antiretroviral therapy. J. Virol. *76*, 707–716.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
CD133 monoclonal antibody, phycoerythrin conjugated	Miltenyi Biotec	Cat# 130-090-853, clone 293C3; RRID:AB_244346
CD133 monoclonal antibody, biotin conjugated	eBioscience	Cat# 13-1338-80, clone AC133; RRID:AB_1210585
CD34 monoclonal antibody, fluorescein isothiocyanate conjugated	BD Biosciencess	Cat# 555821, clone 581; RRID:AB_396150
CD3 monoclonal antibody, allophycocyanine conjugated	eBioscience	Cat# 17-0037, clone OKT3; RRID:AB_1907373
CD3 monoclonal antibody, phycoerythrin conjugated	eBioscience	Cat# 12-0037-42, clone OKT3; RRID:AB_1272078
CD3 monoclonal antibody, Pacific Blue conjugated	BioLegend	Cat# 317314, clone OKT3; RRID:AB_571909
HIV gag, PE (RD-1) conjugate	Beckman Coulter	Cat# 6604667, clone KC57
HIV gag ELISA capture antibody	NIH AIDS Reagent Program	Cat# 3537, clone 183-H12-5C
HIV gag ELISA detection antibody, purified from hybridoma ascites	ATCC	Cat# HB-9725, clone 31-90-25
Bacterial and Virus Strains		
E. coli Stbl2	Fisher	Cat# 10268019
Biological Samples		
Fetal bovine serum, qualified, non-US origin, USDA-approved regions	GIBCO	Cat#10437028
Human AB serum	Fisher	Cat# BP2525-100
Total Control RNA (from Raji cell line)	Invitrogen	Cat# 4307281
Chemicals, Peptides, and Recombinant Proteins		
Ficoll-Paque PLUS media	GE Healthcare	Cat# 17-1440-02
StemSpan SFEM	StemCell Technologies	Cat# 09650
Hybri-Max dimethyl sulfoxide	Millipore Sigma	Cat# D2650-100ML
Tumor necrosis factor alpha (TNFα)	R&D Systems	Cat# 210-TA-005
Streptavidin, Pacific Blue conjugated	Invitrogen	Cat# S-11222
Trizol Reagent	Invitrogen	Cat# 15596026
DNase I, Amp grade	Invitrogen	Cat# 18068015
Phorbol 12-myristate 13-acetate (PMA)	Sigma-Aldrich	Cat# P1585-1MG
Ionomycin calcium	Sigma-Aldrich	Cat# I0634-1MG
Trizol LS Reagent	Invitrogen	Cat# 10296028
qScript Flex cDNA Synthesis kit	Quanta Biosciences	Cat# 101414-112
qScript cDNA Supermix	Quanta Biosciences	Cat# 101414-108
Phusion Hot Start II High Fidelity DNA polymerase	Fisher	Cat# F549L
GelRed nucleic acid gel stain, 10,000X in water	Biotium	Cat# 41003
TaqMan Gene Expression MasterMix	Applied Biosystems	Cat# 4369016
TaqMan Fast Advanced MasterMix	Applied Biosystems	Cat# 4444557
Polyethylenimine	Polysciences Incorporated	Cat# 23966-2
Plasmocin	InvivoGen	Cat# ant-mpt
Nunc Immuno Maxi-Sorp 96 well plate	Fisher	Cat# 12-565-135
Streptavidin Poly-HRP40	Fitzgerald Industries International	Cat# 65R-S104PHRP
HIV-1 p24 core antigen, ELISA standard	ViroGen	Cat# 00177-V

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical Commercial Assays		
CD133 MicroBead Kit - Hematopoietic Tissue, Human	Miltenyi Biotec	Cat#130-100-830
StemSep CD34 Positive Selection Cocktail	StemCell Technologies	Cat#14756
CD4 MicroBead sort kit, untouched	Miltenyi Biotec	Cat# 130-096-533
CD4 MicroBeads	Miltenyi Biotec	Cat# 130-045-101
ACTB TaqMan Gene Expression assay, FAM-MGB	Applied Biosystems	assay ID Hs99999903_m1
POLR2A TaqMan Gene Expression assay, FAM-MGB	Applied Biosystems	Assay ID Hs00172187_m1
pcDNA3.1/V5-His-TOPO kit	Invitrogen	Cat# K480001SC
Pierce EZ-Link Micro Sulfo-NHS-Biotinylation Kit	Fisher	Cat# PI-21925
Deposited Data		
Sequence files	GenBank	MH895379 -MH897920
Experimental Models: Cell Lines		
ACH-2	NIH AIDS Reagent Program	Cat# 349; RRID:CVCL_0138
293T	ATCC	Cat# CRL-3216; RRID:CVCL_0063
CEM-SS	NIH AIDS Reagent Program	Cat# 776
MOLT4-R5	NIH AIDS Reagent Program	Cat# 4984
Oligonucleotides		
See Table S7 for PCR primers		N/A
Recombinant DNA		
pVQA	NIH AIDS Reagent Program	Cat# 12666
pNL4-3-ΔGPE-GFP	McNamara et al., 2012	N/A
p436-5LTRdel	This paper	N/A
pCMV-HIV-1	S.-J.-K. Yee (City of Hope National Medical Center)	N/A
pVSV-G	Nancy Hopkins (Massachusetts Institute of Technology)	N/A
p409HX-gagGFP	This paper	N/A
p421HX-gagGFP	This paper	N/A
p439BX-gagGFP	This paper	N/A
p454BX-gagGFP	This paper	N/A
pHIV454304	This paper	N/A
p89.6	R.G. Collman (University of Pennsylvania)	NIH AIDS Reagent Program Cat# 3552
Software and Algorithms		
FlowJo		https://www.flowjo.com; RRID:SCR_008520
Lasergene suite (SeqMan, SeqBuilder)	DNAStar	V10.1.1; RRID:SCR_000291
MEGA7	(Kumar et al., 2016)	https://www.megasoftware.net
Poisson distribution calculator		https://keisan.casio.com/exec/system/1180573179
HIVAlign	Los Alamos National Laboratory	https://www.hiv.lanl.gov/content/sequence/VIRALIGN/viralign.html
HyperMut	Los Alamos National Laboratory	https://www.hiv.lanl.gov/content/sequence/HYPERMUT/hypermut.html; RRID:SCR_014933
Other		
100 μm nylon cell strainer, sterile	Falcon	Cat# 352360
MACS MS columns	Miltenyi Biotec	Cat# 130-042-201
MACS LS columns	Miltenyi Biotec	Cat# 130-042-401
MiniMACS Separator	Miltenyi Biotec	Cat# 130-042-102

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
MidiMACS Separator	Miltenyi Biotec	Cat# 130-042-302
MACS MultiStand	Miltenyi Biotec	Cat# 130-042-303
Pre-Separation filter, 30 μm	Miltenyi Biotec	Cat# 130-041-407
MagNA Pure Compact	Roche	Cat# 03731146001
MagNA Pure LC Total Nucleic Acid Isolation kit	Roche	Cat# 03246779001
Low retention, RNase-, DNase-free 1.5 ml tubes	Denville Scientific	Cat# C2171
NucleoSpin Gel and PCR Clean Up Columns	Macherey-Nagel	Cat# 740609.250S

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Kathleen Collins (klcollin@med.umich.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human subjects

Fifty-three HIV-infected individuals were recruited through the University of Michigan HIV-AIDS Treatment Program and the Henry Ford Health System. Written informed consent was obtained according to a protocol approved by the University of Michigan Institutional Review Board and Henry Ford Institutional Review Board (U-M IRB number HUM00004959 and HFH IRB number 7403). Donors were between 21 and 69 years old (Table S1), with normal white blood cell counts and plasma viral loads < 48 copies/mL for at least 6 months on antiretroviral therapy (Table S1). We obtained 100 mL of peripheral blood and 20 mL of bone marrow from most donors. The 43 donors who met the sample inclusion criteria were comprised of 38 males and 5 females (Table S1). We did not note sex differences in our ability to detect provirus in donor HSPCs or in the detection of residual PV that matched HSPC provirus. However, this study was not designed to detect such differences. The smaller number of females than males recruited to the study greatly limits this analysis. 41 subjects initiated therapy in chronic phase of HIV-1 infection, two males began therapy in the acute/early stage. All collected samples were coded and anonymized. A sample size of approximately 50 subjects was chosen based on a power calculation that assumed half the HSPC samples would be positive. We estimated that with a sample size of 50 subjects we could be 95% certain that our estimate of the prevalence of HSPC harboring infectious provirus would be accurate to within 7%.

Cell lines

293T cells were authenticated by genotyping performed by the Duke DNA Analysis Facility and grown in DMEM. CEM-SS cells were genotyped but not authenticated as a genotyped comparison cell line was not available to Duke DNA Analysis Facility. CEM-SS, MOLT4-R5 and ACH2 cells were grown in RPMI-1640 plus 10mM HEPES. All cells were grown in medium supplemented with 100 Units/mL penicillin, 100 μg/mL streptomycin, 2 mM glutamine, 10mM HEPES and 10% fetal bovine serum. All cell lines were maintained at 37°C in 5% CO_2 humidified atmosphere.

METHOD DETAILS

Donor cell isolation and fractionation

Isolation of bone marrow mononuclear cells (BMMC)

Bone marrow was aspirated from the posterior iliac crest into 10 cc syringes containing 0.5 mL heparin. Cell and plasma isolations were performed in a dedicated clean room distinct from PCR set up and product gel analysis areas to minimize potential contamination. Marrow was expelled into a 50 mL conical tube and diluted with one volume bone marrow wash buffer (BM Wash), which was composed of phosphate buffered saline [PBS (Invitrogen)], 10% FBS and 2 mM EDTA (Lonza). The diluted marrow was passed through a 100 μm nylon cell strainer by gravity flow. The marrow was further diluted and overlayed on Ficoll-Paque for gradient separation at 720 x *g* for 30 min at 25°C, acceleration setting 2, deceleration setting 2. After the separation, bone marrow plasma in the upper layer was carefully removed and filtered through a 0.2 μm pore to remove cells and stored at −80°C. Bone marrow mononuclear cells (BMMC) at the plasma/Ficoll interface were pooled, washed with BM Wash, and incubated in 20 mL StemSpan SFEM in a prone T75 culture flask for 1.5 – 2 hours at 37°C in 5% CO_2. Non-adherent cells were removed by gently rolling the flask and pipetting. Flasks were rinsed with 5 mL cold MACS Buffer (2% FBS, 2 mM EDTA in PBS) and detached cells were pooled with non-adherent cells.

Purification of CD133$^+$ cells (sort 1)

CD133$^+$ BMMCs were purified using the CD133 MicroBead Kit – (Hematopoietic Tissue, Human) according to the manufacturer's protocol for maximal purity with the following modifications: for donations 453000 and onward, we used 1.5 times the recommended bead to cell ratio to increase CD133$^+$ cell recovery. MACS MS columns in a MiniMACS magnetic separator or MACS LS columns in a MidiMACS Separator on a MACS MultiStand (Miltenyi Biotec) were prepared, depending on the number of cells to be sorted. A 30 μm pore Pre-Separation filter was placed over the first of the two columns used in maximal purity protocol, and columns and filter were equilibrated with MACS Buffer before applying cell samples. Each column was washed twice. Sort purity was ascertained at each step by analyzing an aliquot for CD133, CD34, and CD3 expression by flow cytometry (see below).

Isolation of CD34$^+$ cells (sort 2)

The flow through cells from the first column were used for the second sort in which CD133-depleted CD34$^+$ HSPCs were isolated. For CD34$^+$ cell isolation, flow through cells were concentrated to 200 million cells/mL in MACS Buffer then the StemSep CD34 Positive Selection kit was applied and incubated according to the manufacturer's instructions for maximal purity. MACS Buffer and columns were substituted for washes and magnetic separation as described for the CD133 sort (sort 1). Sort purity was ascertained at each step by analyzing an aliquot for CD133, CD34, and CD3 expression by flow cytometry (see below).

Following purification, HSPCs and matching numbers of flow through cells were washed with 4 mL of StemSpan medium and divided into equal aliquots for DNA extraction on a MagNAPure Compact Nucleic Acid Isolation unit and cryogenic preservation in 10% Hybri-Max DMSO in FBS. The bulk flow through of the second sort was viably frozen and served as a source of non-HSPC BMMCs.

Isolation of BMMC CD4$^+$ cells for viral outgrowth assay

Bulk flow through cells described above were quick-thawed at 37°C and washed twice with R10 medium supplemented with 40 U/mL DNase I (Sigma). CD4$^+$ cells were isolated by negative selection with the CD4 MicroBead Sort Kit, Untouched, according to the manufacturer's protocol.

Plasma virus and peripheral blood mononuclear cell (PBMC) isolation

100 mL of peripheral blood was collected into 10 mL purple top ethylenediaminetetraacetic acid dipotassium salt vacutainer tubes (BD) on the same day as the bone marrow aspirate. The blood was diluted with PBS, then overlayed onto Ficoll-Paque. Gradient separation was performed as described for bone marrow samples. Following the centrifugation, plasma was filtered through a 0.2 μm pore to remove cells. Virus was pelleted at 112,400xg for 1.5 – 2 h at 4°C. Viral pellets were resuspended with 1 mL TRIzol Reagent and stored at −80°C until RNA extraction. Mononuclear cells at the plasma:Ficoll interface were washed twice with PBS before extracting DNA from two aliquots of 10^6 cells and viably freezing the remainder.

CD4$^+$ cells were positively selected from thawed PBMC by incubation with CD4 MicroBeads at a 50% higher bead:cell ratio than in the product insert and sequential isolation over two MACS MS columns. Bead bound CD4$^+$ cells were depleted of adherent cells overnight at 37°C in R10 medium in a tissue culture treated 6 well plate. DNA was extracted from one million cells with a MagNAPure Compact Nucleic Acid Isolation unit.

Flow cytometry and antibodies

All cell samples were fixed with 2% paraformaldehyde (Sigma) in FACS Buffer (2% FBS, 1% human AB serum, 2 mM HEPES, 0.025% sodium azide (Sigma) in PBS without calcium or magnesium) for 30 minutes at room temperature before acquisition on a BD Biosciences FACSCanto cytometer or FACScan cytometer with Cytek 6-color upgrade.

Flow cytometric analysis of cell purity

BMMC aliquots removed for purity assessment were stored at 4-8°C overnight before staining. The purity of the recovered HSPC populations was measured for expression of the following human proteins: CD133 [phycoerythrin (PE) conjugated or biotin-conjugated with streptavidin-Pacific blue], CD34 [conjugated with fluorescein isothiocyanate (FITC)], CD3 [conjugated with allophycocyanine (APC), PE, or Pacific Blue]. Nonviable cells were identified and excluded from analyses by staining with 1 μg/mL 7-aminoactinomycin D (7-AAD, Sigma) or 40 ng/mL 4,6-Diamidino-2-phenylindole (DAPI, Pierce) in FACS Buffer. An un-stained pre-adherence depletion cell sample, isotype controls and single-color compensation controls using OneComp eBeads (eBioscience) were processed in parallel to aid acquisition and compensation settings.

Flow cytometric analysis of GFP expression by transfected and transduced cells

Transfected 293T cells were washed once in PBS, detached from the plate with 0.05% trypsin EDTA (GIBCO), and fixed with 2% paraformaldehyde for 30 minutes at room temperature. Transduced suspension cells were fixed as described for 293T cells. Expression of GFP was assessed on viable cells gated by light scatter.

Flow cytometric analysis of intracellular Gag stain to measure infection

Cells treated with virus were fixed in 2% paraformaldehyde for 30 minutes at room temperature, permeabilized in 0.1% Triton X-100 (Fisher) at room temperature for 5 minutes, and stained with PE-conjugated anti-Gag antibody at room temperature for 20 min. Fluorescence was measured on a BD FACScan cytometer with Cytek 6-color upgrade. Cells were gated by forward scatter versus side scatter, and infected cells were gated by Gag-PE versus side scatter using mock infected cells to set the Gag$^+$ gate.

DNA extraction

Cellular DNA was prepared using a MagNA Pure Compact System in room one of the two dedicated PCR clean rooms described below. Cell suspensions in PBS were extracted with program "DNA Blood 100_400" and with the "Cultured cells" setting.

RNA isolation

RNA was prepared in room one of the two dedicated PCR clean rooms described below. In addition, work space, pipets and equipment surfaces were wiped down with RNase Away (Invitrogen) to reduce RNase transfer before handling samples and reagents.

Plasma virus RNA isolation

Frozen samples in TRIzol were thawed at room temperature. Two micrograms of control human Raji cell line RNA were spiked into donor TRIzol samples prior to organic extraction to serve as carrier for improved RNA recovery as well as a means to monitor RNA extraction and reverse transcription efficiency. A separate tube of TRIzol and Raji RNA was prepared as a reagent only control. A fresh aliquot of chloroform was prepared daily and the organic extraction was performed according to the manufacturer's directions. The aqueous phase of the TRIzol suspensions was further purified with the RNeasy Micro kit (QIAGEN). Samples were divided into two aliquots and processed according to the manufacturer's protocol except that the DNase step was performed after RNA elution as described below as we found this achieved a higher yield. RNA was eluted with 20 μL of RNase-free water.

DNase treatment was performed immediately before reverse transcription. Enzyme reactions were prepared on ice or in a cooler rack. RNA samples were divided into two 20 μL reactions containing 2U of Amp Grade DNaseI and 10 μL RNA following the manufacturer's protocol.

Viral outgrowth supernatants

For RNA isolation from viral supernatants, we harvested supernatants from 5 million $CD4^+$ cells cultured in 1mL of R10 medium supplemented with: 0.3% DMSO solvent control or 50 ng/mL PMA and 1 μM ionomycin incubated at 37°C in 5% CO_2 for 48 hours. Culture supernatants were centrifuged at 700xg for 5 minutes to remove cells and debris, and were processed as previously published (Laird et al., 2015) using TRIzol LS reagent according to the manufacturer's protocol with the modification that 1 μg of Raji RNA was added to the TRIzol LS to allow for quantification of the efficiency of RNA recovery. RNA was dissolved in 10 μL RNase-free water.

Cellular RNA extraction

Cellular RNA was extracted using the QIAGEN RNeasy kit according to the manufacturer's protocol. Samples were frozen at −80°C after homogenization with QIAGEN Qiashredders and thawed at room temperature to complete sample binding, on-column DNase digestion with QIAGEN RNase-free DNase set, and elution in 22 μL water.

Reverse transcription

Reactions were performed for each sample as described below. RNA samples were treated with DNase prior to reverse transcription and/or a reverse transcriptase leave-out was performed to avoid and/or detect contaminating HIV DNA contributing to the downstream qPCR signal. In all samples tested, reverse transcriptase leave-out controls were negative. Control samples indicated that DNase treatment was fully effective.

Plasma virus RNA

Plasma virus cDNA synthesis with oligo dT priming was performed using the qScript Flex cDNA Kit scaled up from the manufacturer's protocol to 200 μL final reaction volumes incubated in 50 μL aliquots. cDNA was pooled and transferred to a low retention 1.5 mL snap cap tube and stored at −20°C.

$CD4^+$ T cell supernatants (VOA)

RNA was reverse transcribed with qScript cDNA SuperMix per manufacturer's instructions scaled up to 200 μL reactions. cDNA was pooled and concentrated to 30 μL in Buffer EB with the QIAQuick PCR Purification kit (QIAGEN) according to the kit protocol.

Cell-associated RNA

cDNA was generated from cell-associated RNA as described for the viral outgrowth assay, but in 50 μL reactions per sample which were tested without volume reduction.

PCR contamination control procedures

To minimize carryover and cross contamination with lab strains, three separate work rooms were used for cell and nucleic acid isolation, PCR reaction set up, and amplified product analysis. Work flow proceeded unidirectionally each day. The first room was for tissue preparation and nucleic acid extraction, as described above. The PCR set-up room (room 2) utilized a dedicated still air hood for reagents and unamplified templates and a separate hood for adding first round PCR products to reactions. Hoods were wiped down with 10% bleach and 70% ethanol before commencing any work. The entire interior of each hood was decontaminated with 10% bleach and UV irradiation monthly. Reactions were assembled and stock tubes removed before tubes containing DNA were opened. Gloves were changed after handling template and before touching refrigerator or freezer handles. First round PCR products were added to second round reactions while wearing disposable sleeve protectors. Second round products were never opened in the reaction set-up room. Tubes containing stock solutions were not opened after handling amplified DNA. Gel loading and DNA extraction from gel bands occurred in the third room.

SGA

Cellular genomic DNA or cDNA prepared from plasma were used in a two-step, limiting dilution SGA reaction. The cut off for limiting dilution was 30% positive reactions in the donor samples. ACH2 cells, which stably harbor one copy of the HIV genome were used for this as a positive control. PBMC DNA from each donor was used to select optimal primers for detection of HIV from that donor. All PCRs were 50 μL reactions containing 1X Phusion Buffer HF, 200 nM dNTP (from 100 mM dNTP set, Roche), 500 nM each primer, and 1 U Hot Start II High Fidelity DNA Polymerase. First round reactions received 5 μL DNA or cDNA template, second round reactions amplified 1-2 μL of the first-round reaction. Primer sequences and sources are listed in Table S7.

Gag/env multiplex SGA PCR

First round reactions were multiplexed for LTR/gag with 5′ LTR forward primer U5-557.9662-f and either tagD4.6b-p24R1d or long1316-D4.6b Dengue-tagged *gag* reverse primer, and *env* C2-V3 with forward primer C2F2 and reverse primer C4R1. Reactions were denatured at 98°C for 30 s, followed by 35 cycles of 98°C for 10 s, 68°C for 10 s, and 72°C for 25 s, with a final extension at 72°C for 5 minutes. For some donors, alternative primers were used because of sequence variation. If first round primers extending from *vif* (5036d) or *tat* (5956d) to the 3′ LTR (LTR-pA-R) were used instead, the 72°C extension time per cycle was 2 minutes and the final extension was for 10 minutes. For full length envelope amplification, first round primers were 5036d and LTR-pA-R.

Second round LTR-*gag* PCR were performed with primers 626 s and Dengue tag D4.6b and incubated at 98°C for 30 s, followed by 35 cycles of 98°C for 10 s, 70°C for 10 s, and 72°C for 40 s and a final extension at 72°C for 10 minutes. Second round *env* C2-V3 reactions consisting of forward primer env1in5 and reverse primer env1in3 were incubated at 98°C for 30 s and were cycled 40 times at 98°C for 10 s, 56°C for 10 s, and 72°C for 15 s, concluding with a 5 minute extension at 72°C. Cycling conditions for second round reactions spanning full length *env* (primers 5956d and LTR-pA-R) began with denaturation at 98°C for 30 s, followed by 35 cycles of 98°C for 10 s, 71°C for 10 s, and 72°C for 2 minutes, with a final incubation at 72°C for 10 minutes. Some donor viruses required an alternative hemi-nested LTR-*gag* primer set U5-577.9662-f and 1294r, and were incubated at 98°C for 30 s with 35 cycles of 98°C for 10 s, 64°C for 10 s, and 72°C for 40 s, and a final extension at 72°C for 5 minutes.

Optimal amplification of *env* C2-V3 from donors 409000, 415000, 421000 and 432000 required donor specific primers which were designed for compatibility with the cycling conditions described for primer sets C2F2/C4R1 and env1in5/env1n3.

Near full length genome PCR

For detection of intact near full length proviral genomes, first round reactions performed at limiting dilution (fewer than 30% positive reactions) for the amplification of single genomes. Reactions contained primers U5-577.9662-f and a Dengue-tagged version of LTR-pA-R (Sebastian et al., 2017). The incubation conditions were 98°C for 30 s, 35 cycles of 98°C for 10 s, 67°C for 10 s, and 72°C for 4 minutes, followed by 72°C for 10 minutes. First round reactions were screened for the presence of *env* C2-V3. Only reactions with open reading frames were further amplified to generate a hemi-nested second round product of approximately 9 kilobase pairs (kbp) and/or three overlapping sub-genomic regions. The hemi-nested reaction contained forward primer 626 s and either LTR-pA-R or tag primer D4.6b and was incubated at 98°C for 30 s and cycled 35 times at 98°C for 10 s, 70°C for 10 s, and 72°C for 4 minutes followed by incubation at 72°C for 10 minutes.

In some cases, we also amplified overlapping subgenomic fragments consisting of a ~700 bp *gag* product (from U5-577.9662-f and 1294r primers), a 3.7 kbp fragment (from 5956d in *vif* and LTR-pA-R through the poly adenylation site), and a 5.3 kbp fragment (from the 3′ end of gag to 5′ end of *env* amplified with primers 1204s and E30HX rc). The three overlapping fragments have exactly matching overlaps ranging from 67-461 base pairs.

The cycling conditions for the 3.7 kbp and 700 bp fragments are described above. The cycling conditions for the 5.3 kbp fragment were 98°C for 30 s, followed by 35 cycles of 98°C for 10 s, 66°C for 10 s, and 72°C for 2 minutes 40 s, and ended with 10 minutes at 72°C. All reactions were amplified in a BioRad C1000 thermocycler.

PCR amplification of integration sites

Amplification of integration sites used a modification of an assay that was previously published (Liu et al., 2006). The assay was modified by the design of new primers that allowed the inclusion of the 5′ deleted region described in Figure 1 along with the integration site. To improve restriction digestion and ligation efficiency, MagNA Pure DNA elution buffer was removed by ethanol precipitation and the DNA was resuspended to the original volume with QIAGEN Buffer EB (10 mM Tris). Recovered DNA was digested in 50 μL with PstI at 37°C x 5 hours followed by an incubation at 80°C x 30 minutes, then applied to a 200 μL ligation reaction containing 1X Ligation Buffer with 1 mM ATP and 3200 units T4 DNA Ligase (New England Biolabs). Ligation reactions were divided into 50 μL aliquots and incubated at 16°C for 15 hours followed by 80°C for 30 minutes. 5 μL of ligation reaction was added to 50 μL PCR for a total of 80 reactions. First round cycling was performed using primers outer LTR3 and outer gag at 98°C for 30 s for an initial cycle followed by 35 cycles of 98°C for 10 s, 63°C for 10 s, and 72°C for 2 minutes. A final incubation was performed at 72°C for 5 minutes. Second round PCR was performed with primers inner LTR3 and inner gag at 98°C for 30 s for the initial cycle followed by 30 cycles of 98°C for 10 s, 66°C for 10 s, and 72°C for 2 minutes. A final incubation was performed at 72°C for 5 minutes. Potential integration sites were purified from gel bands and sequenced.

Once the integration site was identified and sequenced, we designed primers outer Chr9 and inner Chr9 to amplify additional copies. These primers were used with HIV inner LTR3 reverse primer to generate amplicons. Five μL of genomic DNA were added neat to 80 - 50 μL first round PCR reactions and incubated at 98°C for 30 s on the initial cycle followed by 35 cycles of 98°C for 10 s, 70°C for 10 s, and 72°C for 40 s. We also included a final incubation of 72°C for 10 minutes. One μL of first round product was added to second round reactions, which were cycled under the same conditions as the first round.

Gel extraction of DNA fragments

Amplicons were purified from TAE (40 mM Tris, 20 mM acetate, 1 mM disodium EDTA; Fisher)/agarose (Fisher) gels containing 1X GelRed with the QIAQuick Gel Extraction kit (QIAGEN) according to the manufacturer's instructions with the following modifications: Nucleospin Gel and PCR Clean-up Columns were substituted for kit columns and maximum sample and wash volumes were reduced to 650 μL. Buffer PE wash and Buffer EB elutions were incubated for 4 minutes at room temperature before centrifugation through the column membrane. Eluates were sequenced directly in both directions by the Sanger method at the University of Michigan DNA Sequencing Core.

Quantitative reverse transcription PCR (qRT-PCR)

All qRT-PCR reactions were cycled on an Applied Biosystems 7300 thermocycler, with the following conditions: 50°C for 2 minutes, then 95°C for 10 minutes, followed by 45 cycles of 95°C for 15 s and 60°C for 60 s.

Viral supernatants

Plasma and supernatant cDNA yields were determined by quantitative real-time RT-PCR for β-actin mRNA (*ACTB*). cDNAs were diluted 1:10 and assayed with TaqMan Gene Expression MasterMix or Taqman Fast Advanced Gene Expression MasterMix as previously described (Laird et al., 2015). A cDNA synthesis reaction performed on 500 ng of direct input Raji RNA was serially diluted to generate a standard curve.

HIV mRNA was quantified with primers forward 9501 mRNA-F, reverse 9629-polyA-R and 0.25 μM probe L9531FM as previously described (Bullen et al., 2014) with either Taqman FAST Gene Expression MasterMix or Taqman Gene Expression MasterMix. HIV mRNA was assayed on the same plate as the no reverse transcriptase and β-actin reactions. Standard curves were generated by serial dilution of the pVQA HIV plasmid down to 10 HIV copies.

Cellular RNA

Levels of HIV mRNA in VOA and 436 5′LTR deletion cell-associated RNA were quantified as for supernatants. RNA input was normalized for *POLR2A* levels. A standard curve was prepared by serially diluting one sample from neat to 10,000-fold diluted, while all experimental samples were diluted 10-fold for comparison to the standard curve.

Construction of reporter constructs and infectious proviral genomes

Generation of reporter construct containing the deletion identified in donor 436000 (p436-5LTRdel)

A 2238 bp DNA fragment containing the donor 436000 5′LTR and *gag* sequence within NL4-3-ΔGPE-GFP BstZ171 to SwaI sites (see Figure S6) was synthesized by GenScript with flanking 5′ SalI and 3′ XbaI restriction sites and cloned by GenScript into the multiple cloning cassette of pUC18, the same vector as the parental viral plasmid for NL4-3-ΔGPE-GFP. The synthesized DNA was verified by sequencing. The 2.7 kbp AatII to SwaI fragment was gel purified using the QIAquick Gel Extraction kit according to the manufacturer's protocol, but substituting NucleoSpin Gel and PCR Clean-up Columns. This fragment was ligated into NL4-3-ΔGPE-GFP plasmid DNA digested with the same enzymes. Diagnostic digests confirmed the isolation of the desired plasmid (p436-5LTRdel).

Synthesis of LTR reporter constructs from donor LTR sequences

To test the activity of LTR sequences from intact near full length HSPC provirus recovered from donors 409000, 421000, 439000 and 454304, LTR-driven Gag-EGFP (GenBank: LC311024.1) fusion inserts were synthesized by Thermo to provide reporter constructs and cloning intermediates for 454304 reconstruction. Because of problems with PCR amplifying duplicated sequences, HIV near full-length PCR does not completely regenerate full LTRs at each end. For synthetic reconstruction, full LTR sequences were inferred from the available donor derived sequences for all but 89 (40900, 421000, 454304) or 108 bp (439000). This small gap was filled in with NL4-3 sequence. In some cases, additional *gag* sequence was also included to allow for the possibility of full genomic reconstruction in the future. This resulted in some variation in the resultant Gag-GFP fusion protein. The synthetic constructs were provided in Thermo's pMK (p421HX-gagGFP, p439BX-gagGFP, p454BX-gagGFP) or pMS (p409HX-gagGFP) plasmids along with quality control sequence verification. LTR-gag regions were independently confirmed by sequencing.

Reconstruction of donor proviral genomes to produce infectious virus

The donor 454304 near full length genome amplicon generated by PCR with primers U5-577.9662-f and LTR-pA-R was gel purified and TOPO cloned with the pcDNA3.1/V5-His-TOPO kit. To accomplish this, the PCR product was desalted with the QIAQuick PCR Purification kit according to QIAGEN's protocol. An A-overhang was added with Taq DNA polymerase (New England Biolabs) and 1 mM ATP (Roche) at 70°C for 20 minutes before combining with salt solution and vector provided in the TOPO kit in proportions recommended for chemical transformation. The TOPO product was transformed into competent Stbl2 cells according to the manufacturer's instructions and DNA minipreps were screened with SbfI and KpnI (New England Biolabs). Potentially intact clones were regrown for larger scale plasmid isolation to completely sequence the HIV insert.

The whole genome for the near full-length genome isolated from donor 454304 DNA was reconstructed by inserting the BsshII-XhoI fragment from the TOPO-cloned PCR product into the corresponding LTR vector using a three-way ligation strategy using fragments generated with XhoI, FspI and BssHII digestions. Transformed Stbl2 cultures were grown at 30°C to minimize deletions. Colonies were screened for the correctly ligated product by digesting individually with HindIII-HF, XhoI, FspI, and BssHII (all from New England Biolabs). The properly reconstructed plasmid, pHIV454304, lacked the GFP reporter and contained the complete, fully infectious HIV genome.

Transfections and virus production

HEK293T cells were seeded to provide 50%–75% confluence at the time of transfection. 24 hours post-seeding, cells were transfected using a 1:4 mass ratio of DNA:polyethylenimine in 150 mM NaCl. Where pCMV-HIV-1 and pVSV-G plasmids were included, they were added in equal mass ratios with ΔNL4-3-ΔGPE-GFP or LTR construct plasmids. DNA and polyethylenimine were gently vortexed for 10 s and incubated at room temperature for 15 minutes. The transfection solution was then added to the plate drop-wise while agitating, and the cells were maintained at 37°C, 5% CO_2. 48 hours post-transfection, viral supernatants were collected and clarified by pelleting cell debris at 500-1,000xg for 5-15 minutes, aliquoted and stored at −80°. To achieve higher viral titers, equal copy numbers of HIV RNA from the wild-type NL4-3-ΔGPE-GFP or LTR-deleted viral supernatants were concentrated by ultracentrifugation at an average RCF of 50,000xg for 2 hours. Viral pellets were resuspended in D10 medium with 56 μg/mL plasmocin and used for transductions.

HIV p24 ELISA

Gag p24 capture antibody was diluted to 1 μg/ml in carbonate coating buffer (100 mM $NaHCO_3$, 32.35 mM Na_2CO_3 in water). 100 μL of the antibody mixture was applied per well of Nunc Immuno Maxi-Sorp 96 well plates and incubated at 4°C for at least 16 hours. Antibody coated plates were used within 2 weeks of coating. Before use, each well was washed 3 times with wash buffer [200 mM NaCl, 0.05% Tween-20 (both from Fisher) in PBS] and blocked with blocking buffer [1% casein (Sigma) in PBS] for 45 min. Supernatant from HIV transfected or infected cultures was clarified at 6000xg for 5 minutes and serially diluted in ELISA lysis buffer (0.05% Tween-20, 0.5% Triton X-100, 0.5% casein in PBS) to produce readings within the dynamic range of the standard curve. HIV p24 core antigen was also diluted in lysis buffer to generate a standard curve from 1.5625 to 100 ng/mL. Plain lysis buffer was used to blank absorbance readings. 100 μL were applied per well and incubated for 2 hours at room temperature. The plates were then washed four times with a 1 minute incubation per wash. Anti-p24 detection antibody was biotinylated using EZ-Link Micro Sulfo-NHS-Biotinylation kit per manufacturers protocol. 0.5 μg/mL biotinylated antibody in diluent buffer (0.5% casein, 0.5% Tween-20 in PBS) was applied to each well, incubated at room temperature for 1 hour then washed before addition of 1 μg/mL strepavidin-horse radish peroxidase (Streptavidin Poly-HRP40). The plates were incubated for 30 minutes, washed, and developed with TMB substrate (3, 3′, 5, 5′ tetramethylbenzidine, Sigma) for 20 minutes. The reaction was stopped by adding 1/4 volume of 0.5M sulfuric acid (Mallinckrodt) and mixing gently. Absorbance was read at 450 and 650 nm. Peroxidase activity was calculated as A450 minus A650 and p24 concentration interpolated from the standard curve.

HIV transduction and infection of T cell lines

Transduction of CEM-SS cells and infection of MOLT4-R5 cells were by spin inoculation at 1050-1400xg for 2 hours at 25°C in medium containing 4 μg/mL polybrene (Sigma). Transduction was monitored by GFP expression while infection was quantified by intracellular Gag expression as described in the flow cytometry section.

QUANTIFICATION AND STATISTICAL ANALYSIS

Donor and sample exclusion criteria

Ten donors were excluded due to lack of sufficient volume of bone marrow, HSPC sort purity below 80% $CD133^+$ (sort 1) and $CD34^+$ (sort 2), > 1% $CD3^+$ cells, or inability to assess $CD3^+$ T cell content in HSPC population at any donation time point. For the included donors, there were instances where one sort had adequate purity, but the other did not. These exclusions are summarized in Figure S1A. Proviral sequences from HSPC samples were subjected to a statistical test for likelihood that they were contributed by contaminating T cells (see legend for Figure S1). Hypermutated sequences were not excluded from phylogenetic trees but were excluded from the average pairwise distance calculation presented in Table S4 as is standard.

Phylogenetic Analysis

Condensed consensus Maximum Likelihood phylogenetic trees of all *gag* and *env* C2-V3 amplicons were estimated for appropriate clustering. Individual donor phylogenetic trees were inferred by using the Maximum Likelihood method based on the Hasegawa-Kishino-Yano model (Hasegawa et al., 1985). Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites [5 categories (+G, parameter values as noted)]. Codon positions included were 1st+2nd+3rd+Noncoding. All positions with less than 95% site coverage were eliminated. HIV sequence database tools on the Los Alamos National Laboratory website were used to translate HIV genes (HIVAlign) and determine if sequences were hypermutated (HyperMut).

Statistical analyses

Statistical details of experiments can be found in Figure legends. All results are expressed as the mean ± SD unless otherwise specified. Comparison of proportions of sequence populations was performed using online calculators for two-tailed Z-test comparison of 1 or 2 sample proportions (https://www.socscistatistics.com/tests/ztest). P values were determined using two-tailed

Wilcoxon matched pairs signed rank test except for Table S3, which utilized two-tailed student's t-test. All experiments were repeated in at least two independent experiments unless otherwise indicated in the figure legend.

DATA AND SOFTWARE AVAILABILITY

The accession numbers for the HIV sequences from donor cells and tissues reported in this paper are Genbank: MH895359- Genbank: MH897920.