Title

- **Full Title**: Early Emergence and Long-Term Persistence of HIV-Infected T Cell Clones in Children
- **Short Title**: Early and Persistent HIV-1 Cell Clones in Children

Authors

Michael J. Bale[1ǂ], Mary Grace Katusiime[1ǂ*], Daria Wells[2], Xiaolin Wu[2], Jonathan Spindler[1], Elias K. Halvas[3], Joshua C. Cyktor[3], Ann Wiegand[1], Wei Shao[3], Mark F. Cotton[4], Stephen H. Hughes[1], John W. Mellors[3], John M. Coffin[5], Gert U. Van Zyl[6], Mary F. Kearney[1]

Affiliations

[1]HIV Dynamics and Replication Program, CCR, National Cancer Institute, Frederick, MD, US.

[2]Leidos Biomedical Research, Inc., Frederick National Laboratory for Cancer Research, Frederick, MD, US.

[3]Department of Medicine, University of Pittsburgh, Pittsburgh, PA, US.

[4]Department of Pediatrics and Child Health, Tygerberg Children's Hospital and Family Center for Research with Ubuntu, Stellenbosch University, Cape Town, ZA.

[5]Department of Molecular Biology and Microbiology, Tufts University, Boston, MA, US.

[6]Division of Medical Virology, Stellenbosch University and National Health Laboratory Service Tygerberg, Cape Town, ZA.

ǂThese authors contributed equally to this study

*Corresponding Author:
Mary Grace Katusiime, PhD
HIV Dynamics and Replication Program
National Cancer Institute at Frederick
1050 Boyles Street, Building 535, Room 134
Frederick, Maryland, US 21702
1-301-846-6522
marygrace.katusiime@nih.gov

Abstract

Little is known about the emergence and persistence of HIV-infected T cell clones in perinatally-infected children. We analyzed peripheral blood mononuclear cells for clonal expansion in 11 children who initiated antiretroviral therapy (ART) between 1.8-17.4 months of age and with viremia suppressed for 6-9 years. We obtained 8,662 HIV-1 integration sites from pre-ART and 1,861 sites on ART. Expanded clones of infected cells were detected pre-ART in 10/11 children. In 8 children, infected cell clones detected pre-ART persisted for 6-9 years on ART. A comparison of integration sites in the samples obtained on ART with healthy donor PBMC infected *ex-vivo* showed selection for cells with proviruses integrated in *BACH2* and *STAT5B*. Our analyses indicate that, despite marked differences in T cell composition and dynamics between children and adults, HIV-infected cell clones are established early in children, persist for up to 9 years on ART, and can be driven by proviral integration in proto-oncogenes.

Introduction

Human Immunodeficiency Virus (HIV) remains a worldwide health crisis. Approximately 37.9 million people are living with HIV globally and about 1 million die each year (1). Although current antiretroviral therapy (ART) is able to fully suppress HIV-1 replication in the blood (2-5), lymph nodes (6-9), and other tissues (10, 11), it does not cure the infection. If treatment is initiated before the immune system is heavily compromised and if there is lifelong adherence, ART can lead to a partial restoration of CD4+ T cell numbers (12, 13) and can prevent immunodeficiency in most individuals.

The main obstacle to a cure for HIV-1 is the persistence of replication-competent proviruses in long-lived and/or proliferating populations of infected T cells (14, 15). Most of the infected cells that persist on ART contain defective proviruses that are incapable of producing infectious virus (16, 17), although they may be complemented to generate infectious virus upon ART interruption (18, 19). These defective proviruses do not directly contribute to the HIV-1 reservoir that persists on ART but complicate its measurement and may contribute to persistent immune activation. The fraction of infected cells that contains replication-competent (intact/infectious) proviruses has been estimated to be between 1 and 5% in individuals on long-term ART (16, 17, 20). Although the fraction of intact proviruses is small relative to the total number of infected cells, there are sufficient replication-competent proviruses, or defective proviruses that can be readily complemented, to fuel rapid viral rebound if ART is interrupted (21, 22). In both adults and children, when ART is initiated soon after infection, the number of infected cells is reduced, sometimes to levels below the detection limit of current assays (20, 23, 24) and rebound viremia can be significantly delayed (25-28).

Studies of HIV-1 integration sites were initially performed in cell lines and showed that sites were widely distributed but favored highly expressed genes (29-31). Two studies in 2014 were the first to demonstrate expansion of HIV-infected T cells *in vivo* (32, 33). These clones of infected T cells can be detected as early as Fiebig IV in acute infection (34), can persist in adults for at least 3 years on ART (35), and are distributed among different tissues (6). Studies of clones persisting in adults on ART revealed selection against proviruses in expressed genes with a stronger selection against those that are integrated in the same orientation as the host gene (32, 36) and selection for proviruses integrated into some proto-oncogenes—e.g. *BACH2, MKL2, and STAT5B* (32, 33). Although much is now known about the HIV-1 integration site landscape in adults prior to and on ART, there is little information on clonal expansion of infected cells in children who acquired HIV perinatally (PHIV). The largely anti-inflammatory and immunoregulatory environment of the immune systems in children (37, 38) could affect the behavior of infected T cells in ways that would alter the integration site landscape and selection of proviruses at specific sites in children, leading to differences compared to adults. Furthermore, infants have a high fraction of naïve T-cells and fewer clonally expanded T-cells than adults (39, 40), a difference that could affect integration site selection and clonal expansion of infected cells.

To our knowledge, only two reports have investigated clonal expansion of infected cells in children, to date (33, 41). However, only a few children were studied and the integration site sampling in these studies was shallow because it is difficult to collect large numbers of PBMC from infants and children. One study reported clones of infected cells in 3 children initiating ART during chronic infection (33) and the other in 3 neonates on ART who were followed for 2 years (41). Here, we expand on these studies to perform a deep look at the integration site landscape in 11 children initiating ART early and followed for 6-9 years of continual suppression of viremia. We performed a detailed analysis of the integration site landscape by comparing the findings to those in *ex vivo* infected adult PBMC and to those in infected adults on ART (6, 32). To study the emergence of infected CD4+ T cell clones before ART initiation, the dynamics of their long-term persistence, and their potential survival within select genes, we obtained 10,523 integration sites from CD4+ T cells in the perinatally-infected children using samples obtained prior to and during long-term ART. We compared longitudinal integration site datasets to look for evidence of long-term persistence of clones of infected T cells and to investigate the frequency and size of the infected cell clones in the children. Finally, to determine if there exists selective maintenance of infected cells within single genes, we analyzed the integration sites in children compared to sites obtained from *ex vivo*-infected, CD8-depleted PBMC (deposited at rid.ncifcrf.gov) (42, 43).

We report here that clones of infected cells are found in children as early as 1.8 months after birth and that some of the clones that arose early persisted for up to 9 years on ART. Strikingly, although there are noted differences between the immune environments in children compared to adults, our findings on the population of infected T cell clones are similar to what has been reported for adults, suggesting that clonal expansion is the main mechanism for persistence of HIV-1 in children whose viremia is suppressed by ART. We also found that the selection for proviruses integrated in certain genes is similar in adults and children and, importantly, that this selection occurs pre-ART. Integration events and selection for proviruses in these genes in children born with HIV-1 could have long-term effects in adulthood that have not been investigated and are not observed in adults who were not born with HIV infection.

Results

Participants and Sampling

PBMC were obtained from children enrolled in the Children with HIV and Early Antiretroviral therapy (CHER) randomized trial and post-CHER cohort (44) who were identified as plasma HIV-1 RNA positive by 7 weeks of age, initiated ART within 18 months of age (median: 5.1 months; range: [1.8 to 17.4] months), and had long-term, sustained suppression on ART (3) (Table S1). Children were included based on the availability of pre-ART PBMC and PBMC obtained after at least 6 years of continuous suppression of viremia (median: 8.1 years; range 6.8 to 9.1 years). The sex, pre-treatment plasma HIV-1 RNA, ART regimen, time to viral load suppression, and CD4 percentage after long-term ART are shown in Table S1. The pre-ART and on-ART enchriched-CD4+ T cells were analyzed for the presence and persistence of clones of infected cells. We obtained between 197 and 1386 (median: 655) integration sites from each of the samples taken before ART was initiated and between 77 and 432 (median: 137) integration sites from those after at least 6 years on ART (Table 1). In total, we obtained 10,523 HIV-1 integration sites from the 11 children.

Clones of HIV-1 infected cells are detected in children pre-ART and persist on long-term ART

Clonal expansion of cells infected with replication-competent proviruses or defective proviruses that can be complemented during active replication (45, 46), is an important mechanism for HIV-1 persistence on ART (6, 14, 35, 47, 48). The detection of identical integration sites within a sample is the hallmark of clonal expansion of an infected cell, independent of the replication-competence of the integrated provirus. We defined an integration site as being from a clone using three separate criteria: 1) detection of the same integration site at least 3 times in pre-ART samples (to account for recently-infected cells that had duplicated their DNA but would die before establishing a clone), 2) detection of the same integration site at least twice in an on-ART sample (if a cell is dividing after long term ART, it is almost certainly part of a clone), and 3) detection of the same integration site in two different samples from the same donor. Additionally, the method we use to identify integration sites recovers the host-virus DNA junctions from both the 5' and 3' LTRs (32). Therefore, integration sites observed at both junctions were considered as a single integration site under the conservative assumption that they could have originated from the same provirus. In all but one of the 11 donors [Participant Identifier (PID) ZA009], we found at least one clone of infected cells in the pre-ART samples (range: [1 to 27]) (Table 1, column 4). We found at least 3 clones of infected cells in all on-ART samples (range: [3 to 32]) (Table 1, column 9). Although we did not detect any clones of infected cells in the pre-ART samples from donor ZA009 by the stringent criteria described above, 16 of the integration sites were detected twice, suggesting that clones of infected T cells could have been present in this donor pre-ART (Table 1, column 4 parenthetical). We identified clones of infected cells in the pre-ART samples that persisted for up to 6-9 years on ART in 8 of the 11 children (range: [1, 7] clones) (Table 1, column

11). These data show that clonal expansion contributes to the persistence of total HIV-1 DNA in children, as was shown previously for adults (6, 15, 32, 35, 47).

Size of infected cell clones is similar in children and adults

We analyzed the size and frequency of infected cell clones using a modified Gini coefficient called the "oligoclonality index" (OCI) (49). Briefly, the OCI, which has a value between 0 and 1, is a measure of the non-uniformity of a given dataset; 0 indicates complete heterogeneity and 1 indicates complete homogeneity. In our analysis, 0 would mean each detected integration site was detected only once while a value of 1 would mean that all the integration sites would be from a single large clone. In the pre-ART samples, most integration sites were detected only once (Table 1, column 5). The pre-ART samples contained large numbers of recently infected cells that had not undergone clonal expansion. Thus, all pre-ART OCI values were less than 0.1 (range: 0.006 to 0.085; median: 0.027). The pre-ART OCI positively correlated with the age at which ART was initiated – presumably because clones increase in size with time, which makes it easier for us to detect them (Adj. R^2=0.53; p=0.011) (Figure 1A). Stated differently, although clones can arise soon after infection (35, 50), they may require time to expand to a size that can be detected using the integration sites assay (35). As expected, the OCIs were significantly higher during long-term suppression on ART (range: 0.055 to 0.403; median: 0.161; p=0.002) (Table 1, column 10, Figure 1B), suggesting that the short survival of most recently-infected T cells makes it easier to detect clones of infected cells after long-term ART (32, 35). It should be noted; however, that the on-ART OCI does not correlate with time on ART (Adj. R^2=-0.08; p=0.63), suggesting that clonal expansion during ART is not just a function of time, but rather a complex dependence on homeostatic, antigen-driven, and integration-driven proliferation (Figure 1B). We further compared the on-ART OCI in children to published datasets from 9 infected adults (6, 32) on long-term ART and found no statistical difference (p>0.99; Figure 2; numerical data found in Jupyter Notebook – see methods) (51).

Selection for cells with proviruses integrated in certain genes

Recent reports show that HIV-1 proviruses integrated in one of a small number of genes (15, 32, 33, 35, 52) contribute to the growth, survival, and persistence of the infected cell clones *in vivo*. To look for evidence of similar selection in children born with HIV-1 and treated early with ART, we compared the distribution of integration sites from the children (pre-ART and on-ART) to integration sites obtained from *ex vivo* HIV-1 infected, CD8-depleted PBMC from healthy donors [deposited at rid.ncifcrf.gov; (42)]. We asked if there was evidence for enrichment of proviruses in specific genes *in vivo* (relative to *ex vivo*). We also analyzed the orientation of the proviruses relative to the host gene. Enrichment in the fraction of proviruses within, and oriented in the same direction as, the gene are evidence of post-integration selection. Enrichment of the

integration sites was determined by comparing the *ex vivo*-infected PBMC dataset against the *in vivo* datasets. For this analysis, clonally amplified sites were removed from the in vivo datasets by collapsing identical integration sites. Integration sites in intergenic regions (mapped to hg19) were not included in the analysis. The resulting datasets consisted of 335,614 integration sites from the *ex vivo* infected PBMC (87.2% of the initial data), 7039 sites from the pre-ART dataset from the children (83.9%), and 1202 (76.8%) sites from the on-ART dataset from the children. To detect enrichment in both the pre-ART and on-ART datasets relative to the *ex vivo* PBMC dataset, Fisher's Exact Tests were performed on genes in each library with post-hoc multiple tests correction. Adjusted p-values are reported with $p_{adj} \leq 0.05$ being considered significant.

Consistent with what has been observed in virally suppressed adults (32), we found a strong enrichment for proviruses during ART integrated into both *BACH2* ($p_{adj}=2.7*10^{-15}$) and *STAT5B* ($p_{adj}=4.0*10^{-29}$) (Table 2), but not *MKL2* ($p_{adj}>0.05$) during ART. The question of enrichment in samples prior to ART initiation in either adults or children has not previously been addressed. Strikingly, we observed a signal for enrichment of integrations into *BACH2* in children even prior to ART initiation ($p_{adj}=8.9*10^{-17}$) showing that selection can occur early in PHIV infection. Although not statistically significant, we also observed a trend toward selection for integration events in *STAT5B* ($p_{adj}=0.14$) (Table 2) prior to ART initiation.

Previous studies in adults have shown that, if there is post-integration selection for an HIV provirus in a gene, like STAT5B and *BACH2*, the proviruses are highly enriched for the same orientation as the gene (32, 33). We analyzed the genes for which there were at least 15 unique integrations in the *ex vivo* dataset and at least 1 integration in the *in vivo* dataset so that there would be a signal sufficient to detect selection. Although 18 genes were retained for analysis in the pre-ART dataset, only 2 met these criteria in the on-ART dataset (Table S2, S3). Despite the global preference for proviruses detected on ART to be integrated against the gene (*ex vivo* PBMC: 50.0% vs. children on-ART: 54.7%; p=0.0011), there was no evidence for such global selection prior to initiation of ART (*ex vivo* PBMC: 50.0%; children pre-ART: 50.7%; p=0.26 for the difference) (Figure 3A). However, of the 18 genes in which there were sufficient numbers of integrations in pre-ART samples, we found selection for with-the-gene integration in both *BACH2* ($p_{adj}=2.0*10^{-3}$) and *STAT5B* ($p_{adj}=7.8*10^{-3}$) (Table S2, Figure 3B) and an against-the-gene bias in an ankyrin repeat protein, *ANKRD11* ($p_{adj}=0.028$) (Table S3). Although these data provide evidence for strong selection for both *BACH2* and *STAT5B* pre-ART, we do not consider the against-gene bias for *ANKRD11* to be evidence of selection specific to that gene because of the global bias for against-gene integrations and the lack of an enrichment signal in this and previous datasets.

Likewise, integration sites recovered from children on ART in *BACH2* and *STAT5B* were significantly selected for with-the-gene orientation ($p_{BACH2}=0.034$; $p_{STAT5B}=6.2*10^{-5}$) (Table S3, Figure 3B). Taken together with the enrichment analyses, we conclude that cells containing proviruses integrated in *BACH2* and *STAT5B* in the same orientation as the genes were selected in children both prior to and on-ART.

We also compared the within-gene distribution of proviruses in *BACH2* and *STAT5B* in the children vs. the *ex vivo*-infected PBMC using an in-house mapping application (36) (Figure 4). In both genes, clearly visible clusters of integration sites in the same orientation as the gene (shown in blue) in a single intron upstream of the start of translation were observed in the children both before and during ART (Figure 4B, C, E, F). The *ex vivo*-infected PBMC have a broader, randomly oriented (equal red to blue) distribution (Figure 4A, D) in comparison. The different distributions highlight the selection for directional and clustered integration events into *BACH2* and *STAT5B* in children both prior to and on-ART.

Sub-genomic sequencing datasets do not accurately characterize clonality within individuals

Proviruses in a subset of the children in this study were previously characterized using single-genome sequencing (SGS) of the *gag-pol* genes (encoding P6, protease, and the first 900 nucleotides of reverse transcriptase) (3). We assessed the clonality of the infected cells using integration site analysis compared to the identical sequences found in the SGS analysis. We found that proviruses with identical sub-genomic sequences were more common and constituted larger fractions of the data than the clones detected by sequencing integration site analyses (Figure 5, Figure S1). We also calculated the OCI for each set of data and found that the OCIs were significantly higher (average fold difference: 3.4x) for the sub-genomic single-genome sequences than for the integration site datasets ($p = 0.0078$) (Figure 5). These data suggest that either proviruses with identical sub-genomic sequences have different sites of integration, as has been shown for adults (47), or that many of the integration sites that were detected contained proviruses for which the *gag-pol* regions could not be amplified and sequenced due to deletions, PCR primer mismatches, or both.

Discussion

Despite effective therapies, which have reduced the rate of mother-to-child HIV-1 transmission (53-55), approximately 180,000 infants were infected worldwide in 2018 (53). These children must be included in the larger quest for effective HIV-1 curative interventions and such interventions may need to be tailored to their developing immune systems. Although the contribution of clonal expansion to HIV-1 persistence is well-studied in adults (6, 32, 35), this mechanism has not been well-described in children. Additionally, no analysis has been done in children on the clonal expansion of infected cells prior to the initiation of ART. To compare the mechanisms that underlie the persistence of HIV-1 infected cells during ART in adults and vertically-infected children, we performed HIV-1 integration site analysis on samples obtained from perinatally infected infants (prior to ART initiation) and from the same children during long-term suppression of viremia on ART (6-9 years of full suppression on ART). Despite inherent differences in T cell composition between children and adults (40) the clones of HIV-1 infected cells obtained from the blood of PHIV children in our study were not statistically different from adults (6, 32).

A study by Coffin et al. showed that infected cell clones can arise in adults in the first few weeks post-infection (35). In this study, we found that infected cell clones were detectable, using the integration sites assay, in 4 of the 5 samples collected from infants <3 months of age, consistent with early detection of clones in adults (35). In 2 of the 5 infants first sampled at <3 months old, we detected multiple proviruses with identical integration sites in both the pre-ART sample and the 6-9 years on-ART sample, demonstrating that clones of cells arose prior to ART initiation and persisted for years on ART. The other 7 donors, who were >3 months of age when initiating ART, also had detectable infected cell clones that persisted for at least 6 years of treatment. The frequency of clonal detection in the pre-ART populations tracked linearly with the estimated duration of infection prior to ART – using age as a surrogate – suggesting that the number of infected cell clones that expanded to detectable levels increased with the time of untreated infection, at least during the relatively short periods our donors were infected pre-ART. Our finding that infected cell clones had expanded and become large enough to be detected before two months of age supports the idea that the HIV-1 reservoir is generated rapidly, in actively dividing cells, in both adults and children (35, 56).

These results, in conjunction with previous studies showing that ongoing HIV-1 replication does not occur in children when viremia is fully suppressed on ART (3, 57, 58) and the fact that intact proviruses persist for years both in adults treated early (16) and in children treated early (20), supports the conclusion that the HIV-1 reservoir is maintained in vertically-infected children through the proliferation of cells infected prior to ART initiation, as it is in adults (6, 32, 35, 59). However, the available data are limited by the rarity of infected cells and the very small subset of HIV-infected cells that harbor intact, replication-competent proviruses in children (20). Although further studies are required to increase our understanding of the clonal expansion of intact proviruses as a mechanism by which the reservoir persists in both children and adults, it is possible

that defective proviruses can undergo complementation upon ART interruption and contribute to viral rebound (60).

Although the number of infected cells in children on ART is small, we were able to detect an enrichment in the number and the orientation of proviruses in both *BACH2* and *STAT5B* in the pre-ART and on-ART samples, suggesting that proviruses in a specific intron and oriented with these genes can promote the survival of these clones *in vivo*, as in adults (32). While the selection for the survival of cells harboring *BACH2* and *STAT5B* proviruses has been previously described in adults on-ART (32, 33), no data had been presented to show that such selection exists prior to ART initiation. In both pre- and on-ART, we saw clear evidence for selection of cells containing proviruses in the exon immediately upstream of the start site of translation in *BACH2* and *STAT5B*. Although the selection of BACH2 integrants pre-ART was largely driven by a single child (ZA002) who did not initiate ART until 17 months of age, this single example nonetheless shows that clonal selection due to integration in specific genes is not strictly an on-ART phenomenon. The duration of untreated infection in this child may have allowed enough time for the selection of the cells with the *BACH2* proviruses to become detectable. Similar conclusions can be drawn for selection for proviruses integrated in the first intron of STAT5B**,** where there was clear evidence of selection for cells containing proviruses in the first intron, despite it's being a very strong target for integration *ex vivo*. The trend towards enrichment of *STAT5B* integrants in pre-ART samples was due to the high level of sampling required to overcome the background of integration events in this gene compared to the *ex vivo*-PBMC infected dataset; however, the statistically significant orientation bias prior to ART demonstrates that pre-ART selection exists for *STAT5B*.

Samples from a subset of the children studied here were previously characterized in experiments that showed that ART is effective in suppressing on-going cycles of viral replication in children (3). Thus, proviral SGS data were available at the same on-ART timepoint. The OCIs obtained using the P6-PR-RT SGS results were significantly higher than the OCIs obtained from the on-ART integration site data. The observation that a higher OCI was obtained from the SGS data than the ISA data adds to the growing number of studies (15, 47, 50) suggesting that viruses with identical sub-genomic sequences may not all come from a clonal population of infected cells. These data strongly suggest that sub-genomic sequencing does not always accurately identify clones of infected cells or sufficiently characterize the genetic diversity of the intra-patient HIV-1 populations that persist on ART (47). Although the results here are consistent with previous studies showing that sub-genomic sequences are not sufficient to define clonality, it should be noted that calculating an OCI for small-N datasets can result in artificially high OCI values. Studies that are based on integration site analysis, rather than SGS, are more appropriate to study the clonal expansion of infected cells.

It is important to note that because these children were diagnosed within a few weeks of birth it is not known whether the transmission of HIV-1 occurred at birth or in utero. Because of this ambiguity, the age of the participant may not accurately reflect the duration of infection,

334 although we found evidence of clonal expansion as early as 1.8 months after birth. Furthermore,
335 the integration site libraries only represent a small fraction of the total number of infected cells in
336 the blood. It is therefore likely that many of the integration sites that were recovered only once
337 belong to clones of infected cells.

338 Despite these caveats, we have presented here the largest dataset yet of integration sites
339 from pediatric HIV-1 infections both prior to ART and after durable suppression on ART. Because
340 children primarily have naïve T cells, which do not have the HIV coreceptor CCR5 as a surface
341 marker (40), as well as an immune environment that promotes quiescence (37, 38), and a more
342 diverse T cell receptor repertoire (39), it is important to determine if there are differences between
343 the observed frequency of clones and patterns of integration and post-insertional selection in
344 children and adults. However, despite the differences in the immune systems of adults and
345 children, our data suggest that these differences do not influence the infection and clonal expansion
346 of T cells to a degree that is detectable by our integration site analysis. It is possible that by 6 to 9
347 years of age the immune system may be similar enough to that of an adult to account for the striking
348 similarities in the on-ART libraries of these children and the published data from adults. Although
349 these data suggest that the role of clonal expansion as the mechanism for HIV-1 persistence during
350 ART is similar in children and adults, further studies are warranted to better understand how the
351 developing immune system affects clonal expansion and what effects proposed curative
352 interventions might have in both children and adults.

353

354

355

Materials and Methods

Study Approval and Ethics Statement

The CHER trial is registered with ClinicalTrials.gov (NCT00102960). Guardians of all donors provided written informed consent and the study was approved by the Stellenbosch University Internal Review Board.

Total HIV-1 DNA quantification

HIV-1 DNA levels were determined using the integrase cell-associated DNA (iCAD) assay as previously described (61) with the following primers for use with HIV-1 Subtype C:

Forward primer	HIV_Int_FP	CCCTACAATCCCCAAAGTCA	4653 → 4672
Reverse primer	HIV_Int_RP	CACAATCATCACCTGCCATC	5051 → 5070

Integration Sites Assay

ISA was performed and analyzed as previously described (32, 62) using patient-specific primers to the 5' and 3' LTRs. Importantly, our protocol includes a shearing step (63) that effectively tags each DNA molecule, allowing determination of the relative numbers of cells in the initial pool with identical sites of integration (i.e., clonality). The full set of integration sites obtained has been submitted to the Retroviral Integration Sites Database (https://rid.ncifcrf.gov/) (43) and the primer sequences are available in Supplemental Table 5.

A comparison integration site dataset was prepared from CD8-depleted PBMC isolated from two HIV negative human donors infected *in vitro* with replication-competent HIV-1, subtype B (BAL) (64). After 2 days the cells were harvested and DNA was prepared and integration sites analyzed as previously described (36). The global distribution of the integration sites from the two donors, was indistinguishable; therefore all comparisons were performed with combined data from the two donors.

Oligoclonality Index

The oligoclonality Index (OCI) was calculated using a python script available at https://github.com/michaelbale/python_stuff/. Full details of the calculation are described in the supplemental text of Gillet, *et al* (49). Briefly, the LTR-corrected counts of all unique integration sites are sorted into descending order and the cumulative abundance of the clones are summed as a fraction of the total number of unique integration sites and normalized to have a maximal value of 1. Mathematically, the OCI is calculated as below:

389 s_i – LTR-corrected count of integration site i

390 S – Number of unique integration sites in library

391 $N = \Sigma_{i=1}^{S} s_i$ – Total number of integration sites in library

392 $p_i = \frac{s_i}{N}$ – Relative abundance of integration site i

393 $X_i = \Sigma_{k=1}^{i} p_k$ – cumulative abundance of all integration sites of size $\{s_i\}$ or greater

394 $OCI = 2 * (\Sigma_{k=1}^{S} \left(\frac{X_k}{S}\right) - 0.5)$ – Oligonoclonality index of library

395

396 *Statistical Analysis*

397 Clonality was assessed by grouping sequenced integration sites with identical pseudo-
398 3'LTR genomic coordinates and different shear points into count data in R. These count data were
399 used to generate the OCI. Independent integration sites into genes were pooled and assessed for
400 selection by Fisher's Exact test by either the pre-ART or on-ART library vs. the ex vivo infected
401 PBMC library as null set. P-values for this gene-enrichment analysis were corrected post-hoc by
402 the Benjamini-Hochberg method. Orientation biases were assessed in a similar manner with post-
403 hoc corrections only in the pre-ART comparison. All adjusted p-values are presented as p_{adj} where
404 appropriate. All other statistical analyses are noted where appropriate and performed in R v3.5.2.
405 A Jupyter notebook (51) with the R commands and visualizations for the unedited figures available
406 at github.com/michaelbale/cher_bale/.

407

408 *Phylogenetic Analyses*

409 HIV-1 P6-PR-RT sequences were aligned to HIV Consensus C using MUSCLE and
410 neighbor joining phylogenetic p-distance trees were built using MEGA 7
411 (https://www.megasoftware.net/) (65) and outgroup rooted to Consensus C. Distance matrix
412 generation for calculation of the sequence-based OCI was performed using Hamming distance.

413

414

415

416

417

418

References

1. Tavhi F, Carter A, Jahagirdar D, Biehl M, Douwes-Schultz D, Larson S, Arora M, Dwyer-Lindgren L, Steuben K, Abbastabar H, Abu-Raddad L, Abyu D, Adabi M, Adebayo O, Adekanmbi V, Adetokunboh O, Ahmadi A, Ahmadian E, Ahmadpour E, Ahmed M, Akal C, Alahdab F. 2019. Global, regional, and national incidence, prevalence, and mortality of HIV, 1980-2017, and forecasts to 2030, for 195 countries and territories: a systematic analysis for the Global Burden of Disease, Injuries, and Risk Factors Study 2017. THE LANCET 6:e831 - e859.
2. Kearney MF, Spindler J, Shao W, Yu S, Anderson EM, O'Shea A, Rehm C, Poethke C, Kovacs N, Mellors JW, Coffin JM, Maldarelli F. 2014. Lack of Detectable HIV-1 Molecular Evolution during Suppressive Antiretroviral Therapy. PLoS Pathog 10:e1004010.
3. Van Zyl GU, Katusiime MG, Wiegand A, McManus WR, Bale MJ, Halvas EK, Luke B, Boltz VF, Spindler J, Laughton B, Engelbrecht S, Coffin JM, Cotton MF, Shao W, Mellors JW, Kearney MF. 2017. No evidence of HIV replication in children on antiretroviral therapy. J Clin Invest doi:10.1172/JCI94582.
4. Josefsson L, von Stockenstrom S, Faria NR, Sinclair E, Bacchetti P, Killian M, Epling L, Tan A, Ho T, Lemey P, Shao W, Hunt PW, Somsouk M, Wylie W, Douek DC, Loeb L, Custer J, Hoh R, Poole L, Deeks SG, Hecht F, Palmer S. 2013. The HIV-1 reservoir in eight patients on long-term suppressive antiretroviral therapy is stable with few genetic changes over time. Proc Natl Acad Sci U S A 110:E4987-96.
5. Vancoillie L, Hebberecht L, Dauwe K, Demecheleer E, Dinakis S, Vaneechoutte D, Mortier V, Verhofstede C. 2017. Longitudinal sequencing of HIV-1 infected patients with low-level viremia for years while on ART shows no indications for genetic evolution of the virus. Virology 510:185-193.
6. McManus WR, Bale MJ, Spindler J, Wiegand A, Musick A, Patro SC, Sobolewski MD, Musick VK, Anderson EM, Cyktor JC, Halvas EK, Shao W, Wells D, Wu X, Keele BF, Milush JM, Hoh R, Mellors JW, Hughes SH, Deeks SG, Coffin JM, Kearney MF. 2019. HIV-1 in lymph nodes is maintained by cellular proliferation during antiretroviral therapy. J Clin Invest 130.
7. Bozzi G, Simonetti FR, Watters SA, Anderson EM, Gouzoulis M, Kearney MF, Rote P, Lange C, Shao W, Gorelick R, Fullmer B, Kumar S, Wank S, Hewitt S, Kleiner DE, Hattori J, Bale MJ, Hill S, Bell J, Rehm C, Grossman Z, Yarchoan R, Uldrick T, Maldarelli F. 2019. No evidence of ongoing HIV replication or compartmentalization in tissues during combination antiretroviral therapy: Implications for HIV eradication. Sci Adv 5:eaav2045.
8. Josefsson L, Palmer S, Faria NR, Lemey P, Casazza J, Ambrozak D, Kearney M, Shao W, Kottilil S, Sneller M, Mellors J, Coffin JM, Maldarelli F. 2013. Single cell analysis of lymph node tissue from HIV-1 infected patients reveals that the majority of CD4+ T-cells contain one HIV-1 DNA molecule. PLoS Pathog 9:e1003432.

9. Lee E, von Stockenstrom S, Morcilla V, Odevall L, Hiener B, Shao W, Hartogensis W, Bacchetti P, Milush J, Liegler T, Sinclair E, Hatano H, Hoh R, Somsouk M, Hunt P, Boritz E, Douek D, Fromentin R, Chomont N, Deeks SG, Hecht FM, Palmer S. 2019. The impact of antiretroviral therapy duration on the HIV-1 infection of T-cells within anatomic sites. J Virol doi:10.1128/JVI.01270-19.
10. Imamichi H, Degray G, Dewar RL, Mannon P, Yao M, Chairez C, Sereti I, Kovacs JA. 2011. Lack of compartmentalization of HIV-1 quasispecies between the gut and peripheral blood compartments. J Infect Dis 204:309-14.
11. Evering TH, Mehandru S, Racz P, Tenner-Racz K, Poles MA, Figueroa A, Mohri H, Markowitz M. 2012. Absence of HIV-1 evolution in the gut-associated lymphoid tissue from patients on combination antiviral therapy initiated during primary infection. PLoS Pathog 8:e1002506.
12. Maartens G, Celum C, Lewin SR. 2014. HIV infection: epidemiology, pathogenesis, treatment, and prevention. Lancet 384:258-71.
13. Li TS, Tubiana R, Katlama C, Calvez V, Ait Mohand H, Autran B. 1998. Long-lasting recovery in CD4 T-cell function and viral-load reduction after highly active antiretroviral therapy in advanced HIV-1 disease. Lancet 351:1682-6.
14. Simonetti FR, Sobolewski MD, Fyne E, Shao W, Spindler J, Hattori J, Anderson EM, Watters SA, Hill S, Wu X, Wells D, Su L, Luke BT, Halvas EK, Besson G, Penrose KJ, Yang Z, Kwan RW, Van Waes C, Uldrick T, Citrin DE, Kovacs J, Polis MA, Rehm CA, Gorelick R, Piatak M, Keele BF, Kearney MF, Coffin JM, Hughes SH, Mellors JW, Maldarelli F. 2016. Clonally expanded CD4+ T cells can produce infectious HIV-1 in vivo. Proc Natl Acad Sci U S A 113:1883-8.
15. Einkauf KB, Lee GQ, Gao C, Sharaf R, Sun X, Hua S, Chen SM, Jiang C, Lian X, Chowdhury FZ, Rosenberg ES, Chun TW, Li JZ, Yu XG, Lichterfeld M. 2019. Intact HIV-1 proviruses accumulate at distinct chromosomal positions during prolonged antiretroviral therapy. J Clin Invest 129:988-998.
16. Bruner KM, Murray AJ, Pollack RA, Soliman MG, Laskey SB, Capoferri AA, Lai J, Strain MC, Lada SM, Hoh R, Ho YC, Richman DD, Deeks SG, Siliciano JD, Siliciano RF. 2016. Defective proviruses rapidly accumulate during acute HIV-1 infection. Nat Med 22:1043-9.
17. Ho YC, Shan L, Hosmane NN, Wang J, Laskey SB, Rosenbloom DI, Lai J, Blankson JN, Siliciano JD, Siliciano RF. 2013. Replication-competent noninduced proviruses in the latent reservoir increase barrier to HIV-1 cure. Cell 155:540-51.
18. Lori F, Hall L, Lusso P, Popovic M, Markham P, Franchini G, Reitz MS. 1992. Effect of reciprocal complementation of two defective human immunodeficiency virus type 1 (HIV-1) molecular clones on HIV-1 cell tropism and virulence. Journal of Virology 66:5553-5560.

19. Salzwedel K, Berger EA. 2000. Cooperative subunit interactions within the oligomeric envelope glycoprotein of HIV-1: Functional complementation of specific defects in gp120 and gp41. Proceedings of the National Academy of Sciences 97:12794-12799.
20. Katusiime MG, Halvas EK, Wright I, Joseph K, Bale MJ, Kirby-McCullough B, Engelbrecht S, Shao W, Hu WS, Cotton MF, Mellors JW, Kearney MF, van Zyl GU. 2019. Intact HIV proviruses persist in children 7-9 years after initiation of ART in the first year of life. J Virol doi:10.1128/JVI.01519-19.
21. Chun TW, Justement JS, Murray D, Hallahan CW, Maenza J, Collier AC, Sheth PM, Kaul R, Ostrowski M, Moir S, Kovacs C, Fauci AS. 2010. Rebound of plasma viremia following cessation of antiretroviral therapy despite profoundly low levels of HIV reservoir: implications for eradication. AIDS 24:2803-8.
22. Hamlyn E, Ewings FM, Porter K, Cooper DA, Tambussi G, Schechter M, Pedersen C, Okulicz JF, McClure M, Babiker A, Weber J, Fidler S, Insight S, Investigators S. 2012. Plasma HIV viral rebound following protocol-indicated cessation of ART commenced in primary and chronic HIV infection. PLoS One 7:e43754.
23. de Souza MS, Pinyakorn S, Akapirat S, Pattanachaiwit S, Fletcher JL, Chomchey N, Kroon ED, Ubolyam S, Michael NL, Robb ML, Phanuphak P, Kim JH, Phanuphak N, Ananworanich J, Group RSS. 2016. Initiation of Antiretroviral Therapy During Acute HIV-1 Infection Leads to a High Rate of Nonreactive HIV Serology. Clin Infect Dis 63:555-61.
24. Ananworanich J, Chomont N, Eller LA, Kroon E, Tovanabutra S, Bose M, Nau M, Fletcher JLK, Tipsuk S, Vandergeeten C, O'Connell RJ, Pinyakorn S, Michael N, Phanuphak N, Robb ML, Rv, groups RSs. 2016. HIV DNA Set Point is Rapidly Established in Acute HIV Infection and Dramatically Reduced by Early ART. EBioMedicine 11:68-72.
25. Henrich TJ, Hanhauser E, Marty FM, Sirignano MN, Keating S, Lee TH, Robles YP, Davis BT, Li JZ, Heisey A, Hill AL, Busch MP, Armand P, Soiffer RJ, Altfeld M, Kuritzkes DR. 2014. Antiretroviral-free HIV-1 remission and viral rebound after allogeneic stem cell transplantation: report of 2 cases. Ann Intern Med 161:319-27.
26. Luzuriaga K, Gay H, Ziemniak C, Sanborn KB, Somasundaran M, Rainwater-Lovett K, Mellors JW, Rosenbloom D, Persaud D. 2015. Viremic relapse after HIV-1 remission in a perinatally infected child. N Engl J Med 372:786-8.
27. Persaud D, Gay H, Ziemniak C, Chen YH, Piatak M, Jr., Chun TW, Strain M, Richman D, Luzuriaga K. 2013. Absence of detectable HIV-1 viremia after treatment cessation in an infant. N Engl J Med 369:1828-35.
28. Henrich TJ, Hatano H, Bacon O, Hogan LE, Rutishauser R, Hill A, Kearney MF, Anderson EM, Buchbinder SP, Cohen SE, Abdel-Mohsen M, Pohlmeyer CW, Fromentin R, Hoh R, Liu AY, McCune JM, Spindler J, Metcalf-Pate K, Hobbs KS, Thanh C, Gibson EA, Kuritzkes DR, Siliciano RF, Price RW, Richman DD, Chomont N, Siliciano JD, Mellors JW, Yukl SA, Blankson JN, Liegler T, Deeks SG. 2017. HIV-1 persistence following

extremely early initiation of antiretroviral therapy (ART) during acute HIV-1 infection: An observational study. PLoS Med 14:e1002417.

29. Craigie R, Bushman FD. 2012. HIV DNA integration. Cold Spring Harb Perspect Med 2:a006890.

30. Schröder AR, Shinn P, Chen H, Berry C, Ecker JR, Bushman F. 2002. HIV-1 integration in the human genome favors active genes and local hotspots. Cell 110:521-9.

31. Singh PK, Plumb MR, Ferris AL, Iben JR, Wu X, Fadel HJ, Luke BT, Esnault C, Poeschla EM, Hughes SH, Kvaratskhelia M, Levin HL. 2015. LEDGF/p75 interacts with mRNA splicing factors and targets HIV-1 integration to highly spliced genes. Genes Dev 29:2287-97.

32. Maldarelli F, Wu X, Su L, Simonetti FR, Shao W, Hill S, Spindler J, Ferris AL, Mellors JW, Kearney MF, Coffin JM, Hughes SH. 2014. HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. Science 345:179-83.

33. Wagner TA, McLaughlin S, Garg K, Cheung CY, Larsen BB, Styrchak S, Huang HC, Edlefsen PT, Mullins JI, Frenkel LM. 2014. HIV latency. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. Science 345:570-3.

34. Fiebig EW, Wright DJ, Rawal BD, Garrett PE, Schumacher RT, Peddada L, Heldebrant C, Smith R, Conrad A, Kleinman SH, Busch MP. 2003. Dynamics of HIV viremia and antibody seroconversion in plasma donors: implications for diagnosis and staging of primary HIV infection. AIDS 17:1871-1879.

35. Coffin JM, Wells DW, Zerbato JM, Kuruc JD, Guo S, Luke BT, Eron JJ, Bale M, Spindler J, Simonetti FR, Hill S, Kearney MF, Maldarelli F, Wu X, Mellors JW, Hughes SH. 2019. Clones of infected cells arise early in HIV-infected individuals. JCI Insight 4.

36. Coffin JM, Bale MJ, Wells DW, Guo S, Luke B, Zerbato JM, Sobolewski M, Sia T, Shao W, Xiaolin W, Maldarelli F, Kearney MF, Mellors JW, Hughes SH. 2020. What Shapes the in Vivo Distribution of HIV Integration Sites? Submitted.

37. Surendran N, Simmons A, Pichichero ME. 2018. TLR agonist combinations that stimulate Th type I polarizing responses from human neonates. Innate Immun 24:240-251.

38. Kollmann TR, Crabtree J, Rein-Weston A, Blimkie D, Thommai F, Wang XY, Lavoie PM, Furlong J, Fortuno ES, 3rd, Hajjar AM, Hawkins NR, Self SG, Wilson CB. 2009. Neonatal innate TLR-mediated responses are distinct from those of adults. J Immunol 183:7150-60.

39. Wedderburn LR, Patel A, Varsani H, Woo P. 2001. The developing human immune system: T-cell receptor repertoire of children and young adults shows a wide discrepancy in the frequency of persistent oligoclonal T-cell expansions. Immunology 102:301-9.

40. Shearer WT, Rosenblatt HM, Gelman RS, Oyomopito R, Plaeger S, Stiehm ER, Wara DW, Douglas SD, Luzuriaga K, McFarland EJ, Yogev R, Rathore MH, Levy W, Graham BL, Spector SA, Pediatric ACTG. 2003. Lymphocyte subsets in healthy children from birth through 18 years of age: the Pediatric AIDS Clinical Trials Group P1009 study. J Allergy Clin Immunol 112:973-80.

41. Garcia-Broncano P, Maddali S, Einkauf K, Jiang C, Gao C, Chevalier J, Chowdhury F, Maswabi K, Ajibola G, Moyo S, Mohammed T, Ncube T, Makhema J, Jean-Philippe P, Yu X, Powis K, Lockman S, Kuritzkes D, Shapiro R, Lichterfeld M. 2019. Early antiretroviral therapy in neonates with HIV-1 infection restricts viral reservoir size and induces a distinct innate immune profile. Science Translational Medicine 11.
42. Ferris AL, Wells DW, Guo S, Del Prete GQ, Swanstrom AE, Coffin JM, Wu X, Lifson JD, Hughes SH. 2019. Clonal expansion of SIV-infected cells in macaques on antiretroviral therapy is similar to that of HIV-infected cells in humans. PLoS Pathog 15:e1007869.
43. Shao W, Shan J, Kearney MF, Wu X, Maldarelli F, Mellors JW, Luke B, Coffin JM, Hughes SH. 2016. Retrovirus Integration Database (RID): a public database for retroviral insertion sites into host genomes. Retrovirology 13:47.
44. Cotton MF, Violari A, Otwombe K, Panchia R, Dobbels E, Rabie H, Josipovic D, Liberty A, Lazarus E, Innes S, van Rensburg AJ, Pelser W, Truter H, Madhi SA, Handelsman E, Jean-Philippe P, McIntyre JA, Gibb DM, Babiker AG, Team CS. 2013. Early time-limited antiretroviral therapy versus deferred therapy in South African infants infected with HIV: results from the children with HIV early antiretroviral (CHER) randomised trial. Lancet 382:1555-63.
45. Burke DS. 1997. Recombination in HIV: an important viral evolutionary strategy. Emerg Infect Dis 3:253-9.
46. Robertson DL, Sharp PM, McCutchan FE, Hahn BH. 1995. Recombination in HIV-1. Nature 374:124-6.
47. Patro SC, Brandt LD, Bale MJ, Halvas EK, Joseph KW, Shao W, Wu X, Guo S, Murrell B, Wiegand A, Spindler J, Raley C, Hautman C, Sobolewski M, Fennessey CM, Hu WS, Luke B, Hasson JM, Niyongabo A, Capoferri AA, Keele BF, Milush J, Hoh R, Deeks SG, Maldarelli F, Hughes SH, Coffin JM, Rausch JW, Mellors JW, Kearney MF. 2019. Combined HIV-1 sequence and integration site analysis informs viral dynamics and allows reconstruction of replicating viral ancestors. Proc Natl Acad Sci U S A doi:10.1073/pnas.1910334116.
48. Bui JK, Sobolewski MD, Keele BF, Spindler J, Musick A, Wiegand A, Luke BT, Shao W, Hughes SH, Coffin JM, Kearney MF, Mellors JW. 2017. Proviruses with identical sequences comprise a large fraction of the replication-competent HIV reservoir. PLoS Pathog 13:e1006283.
49. Gillet NA, Malani N, Melamed A, Gormley N, Carter R, Bentley D, Berry C, Bushman FD, Taylor GP, Bangham CR. 2011. The host genomic environment of the provirus determines the abundance of HTLV-1-infected T-cell clones. Blood 117:3113-22.
50. Garcia-Broncano P, Maddali S, Einkauf KB, Jiang C, Gao C, Chevalier J, Chowdhury FZ, Maswabi K, Ajibola G, Moyo S, Mohammed T, Ncube T, Makhema J, Jean-Philippe P, Yu XG, Powis KM, Lockman S, Kuritzkes DR, Shapiro R, Lichterfeld M. 2019. Early antiretroviral therapy in neonates with HIV-1 infection restricts viral reservoir size and induces a distinct innate immune profile. Sci Transl Med 11.

51. Kluyver T, Ragan-Kelley B, Perez F, Granger B, Bussonnier M, Frederic J, Kelley K, Hamrick J, Grout J, Corlay S, Ivanov P, Avila D, Abdalla S, Willing C. 2016. Jupyter Notebooks – a publishing format for reproducible computational workflows.
52. Ikeda T, Shibata J, Yoshimura K, Koito A, Matsushita S. 2007. Recurrent HIV-1 integration at the BACH2 locus in resting CD4+ T cell populations during effective highly active antiretroviral therapy. J Infect Dis 195:716-25.
53. Kempton J, Hill A, Levi JA, Heath K, Pozniak A. 2019. Most new HIV infections, vertical transmissions and AIDS-related deaths occur in lower-prevalence countries. J Virus Erad 5:92-101.
54. Ghoma Linguissi LS, Sagna T, Soubeiga ST, Gwom LC, Nkenfou CN, Obiri-Yeboah D, Ouattara AK, Pietra V, Simpore J. 2019. Prevention of mother-to-child transmission (PMTCT) of HIV: a review of the achievements and challenges in Burkina-Faso. HIV AIDS (Auckl) 11:165-177.
55. Nduati EW, Hassan AS, Knight MG, Muema DM, Jahangir MN, Mwaringa SL, Etyang TJ, Rowland-Jones S, Urban BC, Berkley JA. 2015. Outcomes of prevention of mother to child transmission of the human immunodeficiency virus-1 in rural Kenya--a cohort study. BMC Public Health 15:1008.
56. Jones BR, Kinloch NN, Horacsek J, Ganase B, Harris M, Harrigan PR, Jones RB, Brockman MA, Joy JB, Poon AFY, Brumme ZL. 2018. Phylogenetic approach to recover integration dates of latent HIV sequences within-host. Proc Natl Acad Sci U S A 115:E8958-E8967.
57. Persaud D, Ray SC, Kajdas J, Ahonkhai A, Siberry GK, Ferguson K, Ziemniak C, Quinn TC, Casazza JP, Zeichner S, Gange SJ, Watson DC. 2007. Slow human immunodeficiency virus type 1 evolution in viral reservoirs in infants treated with effective antiretroviral therapy. AIDS Res Hum Retroviruses 23:381-90.
58. Ruff CT, Ray SC, Kwon P, Zinn R, Pendleton A, Hutton N, Ashworth R, Gange S, Quinn TC, Siliciano RF, Persaud D. 2002. Persistence of wild-type virus and lack of temporal structure in the latent reservoir for human immunodeficiency virus type 1 in pediatric patients with extensive antiretroviral exposure. J Virol 76:9481-92.
59. Reeves DB, Duke ER, Wagner TA, Palmer SE, Spivak AM, Schiffer JT. 2018. A majority of HIV persistence during antiretroviral therapy is due to infected cell proliferation. Nat Commun 9:4811.
60. Lu C-L, Pai JA, Nogueira L, Mendoza P, Gruell H, Oliveira TY, Barton J, Lorenzi JCC, Cohen YZ, Cohn LB, Klein F, Caskey M, Nussenzweig MC, Jankovic M. 2018. Relationship between intact HIV-1 proviruses in circulating CD4⁺ T cells and rebound viruses emerging during treatment interruption. Proceedings of the National Academy of Sciences 115:E11341-E11348.
61. Hong F, Aga E, Cillo AR, Yates AL, Besson G, Fyne E, Koontz DL, Jennings C, Zheng L, Mellors JW. 2016. Novel Assays for Measurement of Total Cell-Associated HIV-1 DNA and RNA. J Clin Microbiol 54:902-11.

62. Wells DW, Guo S, Shao W, Bale MJ, Coffin JM, Hughes SH, Wu X. 2020. An analytical pipeline for identifying and mapping the integration sites of HIV and other retroviruses. BMC Genomics 21:216.
63. Berry CC, Gillet NA, Melamed A, Gormley N, Bangham CR, Bushman FD. 2012. Estimating abundances of retroviral insertion sites from DNA fragment length data. Bioinformatics 28:755-62.
64. Gartner S, Markovits P, Markovitz DM, Kaplan MH, Gallo RC, Popovic M. 1986. The role of mononuclear phagocytes in HTLV-III/LAV infection. Science 233:215-9.
65. Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol 33:1870-4.

Acknowledgements

General: We thank the participants and their caregivers for their willingness to take part in this study as well as the clinic staff for their assistance.

Funding: This study was supported by funding from the National Institutes of Allergy and Infectious Diseases (Comprehensive International Program for Research on AIDS (CIPRA)–South Africa (U19 AI153217), U.S.-South Africa Program for Collaborative Biomedical Research, National Cancer Institute: U01CA200441, the Office of AIDS Research (to M.F.K.), Departments of Health of the Western Cape and Gauteng, South Africa, ViiV Healthcare, and by National Institute of Mental Health under Award R01MH105134. JMC was a Research Professor of the American Cancer Society and was supported by the NCI through a Leidos subcontract, l3XS110, and research grant R35 CA200421. JWM was supported by the NCI through a Leidos subcontract, 12XS547.

Author Contributions

MJB – Analyzed data, performed statistical analyses, wrote the paper
MGK – Processed samples, performed SGS, analyzed data, wrote the paper
DW – Performed ISA, analyzed data
XW – Performed ISA, analyzed data
JS – Processed samples, performed SGS
EKH – Processed samples, performed DNA quantitation
JCC – Performed DNA quantitation
AW – Performed SGS
WS – Analyzed Data
MFC – Conceived of idea, reviewed manuscript
SHH – Conceived of idea, analyzed data, wrote the paper
JWM – Conceived of idea, wrote the paper

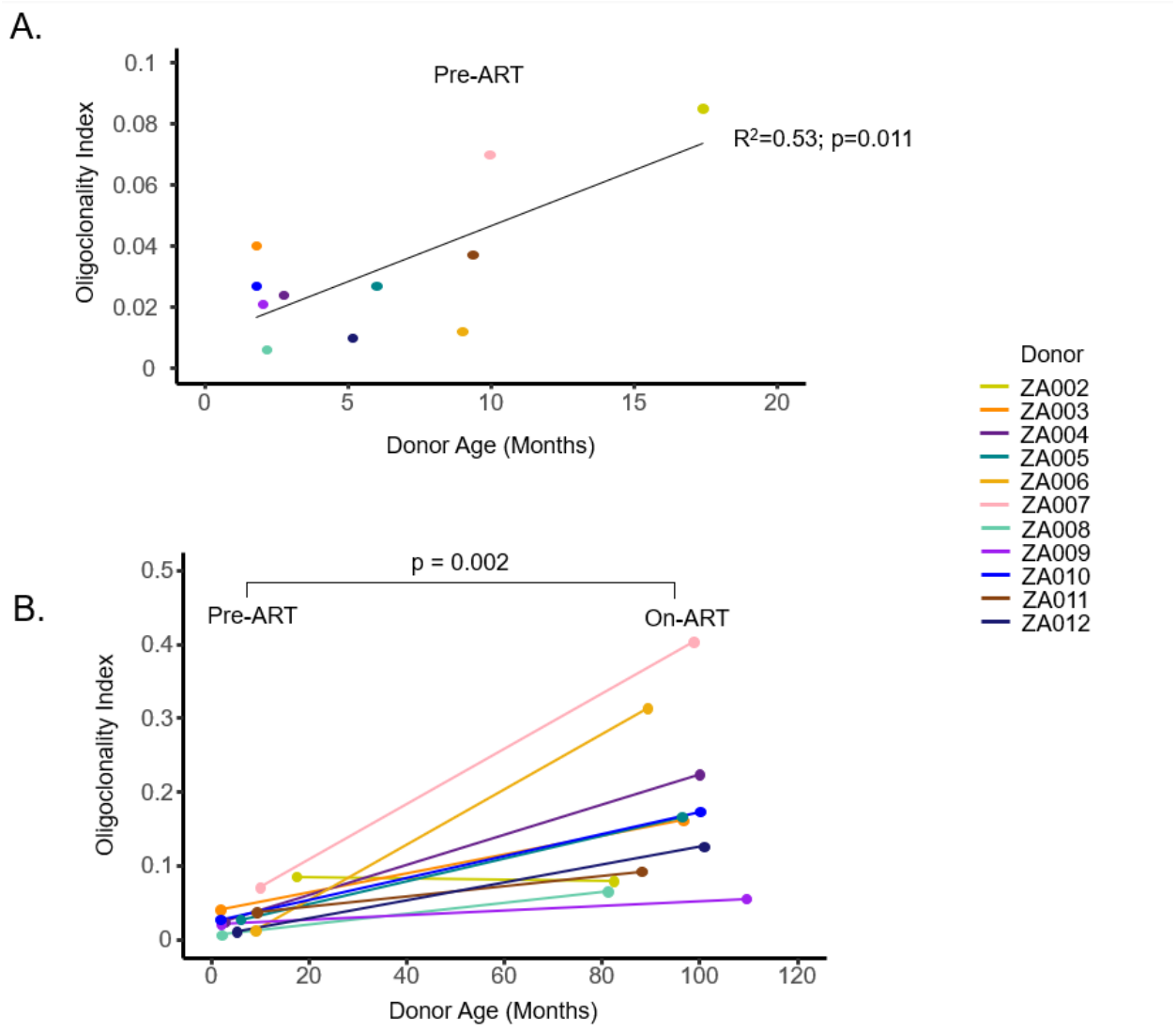
JMC–Analyzed data, wrote the paper
GUVZ – Conceived of idea, wrote the paper
MFK – Conceived of idea, analyzed data, wrote the paper

Competing interests: JWM is consultant to Gilead Sciences, Accelevir Diagnostics, Merck, and Infectious Disease Connect, has received grants from Gilead Sciences to the University of Pittsburgh, and owns share options in Co-crystal Pharmaceuticals, Inc. and Abound Bio, Inc. unrelated to the current work. The remaining authors have declared that no conflict of interest exist.

Data and materials availability: Previously published sequences from Van Zyl et al (3) were accessed from GenBank with accession numbers (KY820119-KY820376) retaining only the sequences associated with the on-ART timepoint of PIDs ZA003 – ZA010. The integration site libraries for the donors in this study are available in the Retrovirus Integration Database (RID) (43) at rid.ncifcrf.gov.The ex vivo infected PBMC integration sites library is also available in the RID under the Pubmed ID 31291371 (deposited at rid.ncifcrf.gov) (42).

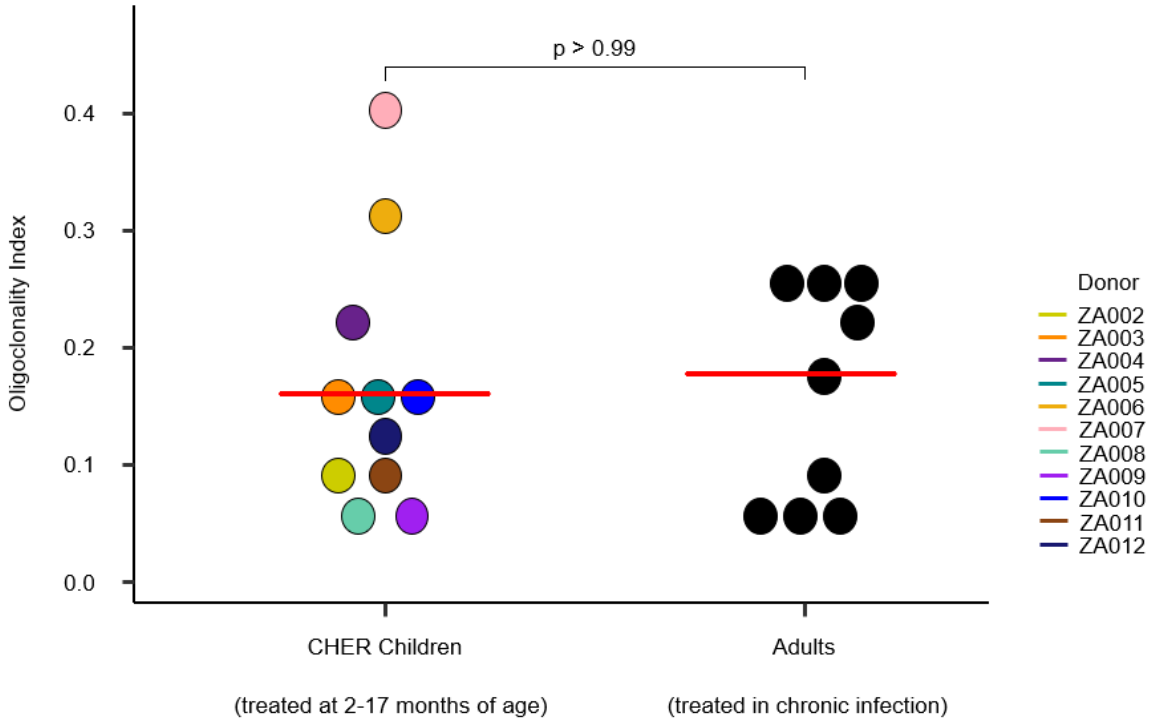
Figures and Tables

Figure 1. Oligoclonality indexes correlate both with time on ART and duration of infection prior to ART.



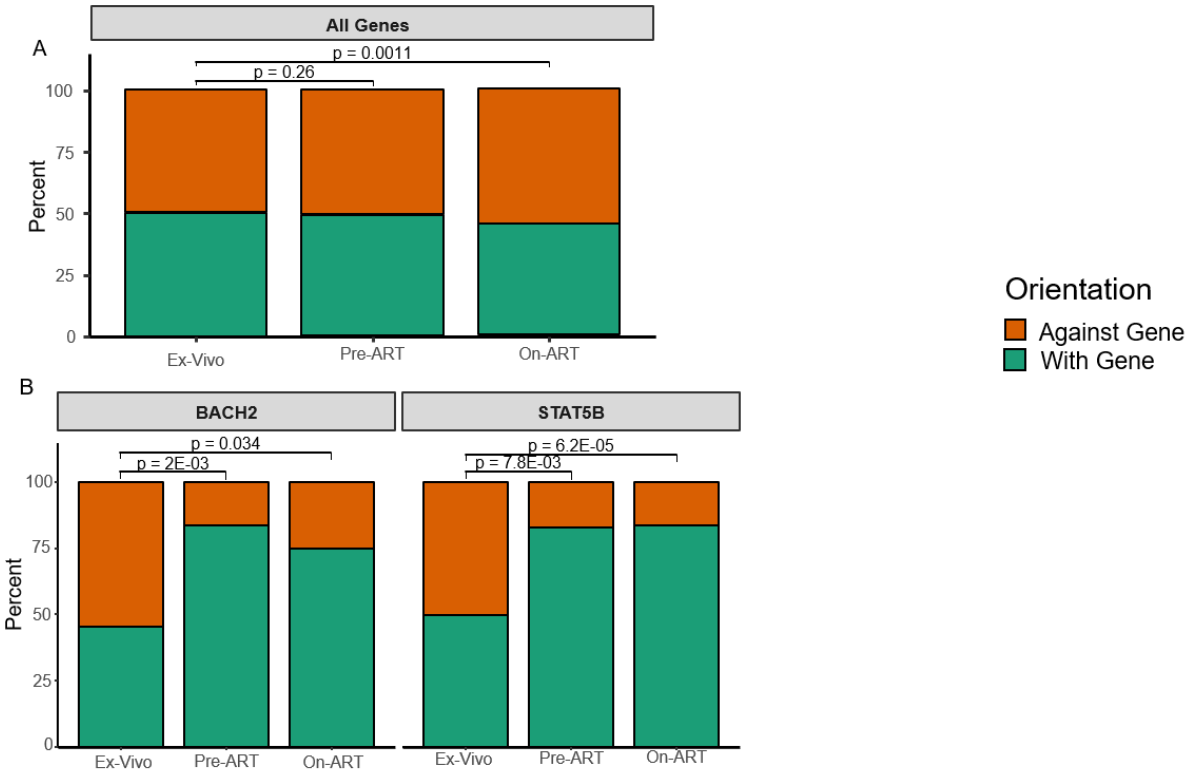
Oligoclonality indexes (OCI) were calculated from the pre-ART and on-ART libraries plotted against Donor Age in months. Pre-ART OCIs were evaluated via linear regression and F-test against donor age **(A)** while change in OCI as a function of ART status was evaluated by Wilcoxon Signed-Rank test **(B)**.

Figure 2. Oligoclonality indexes are comparable between ART-suppressed adults and children.



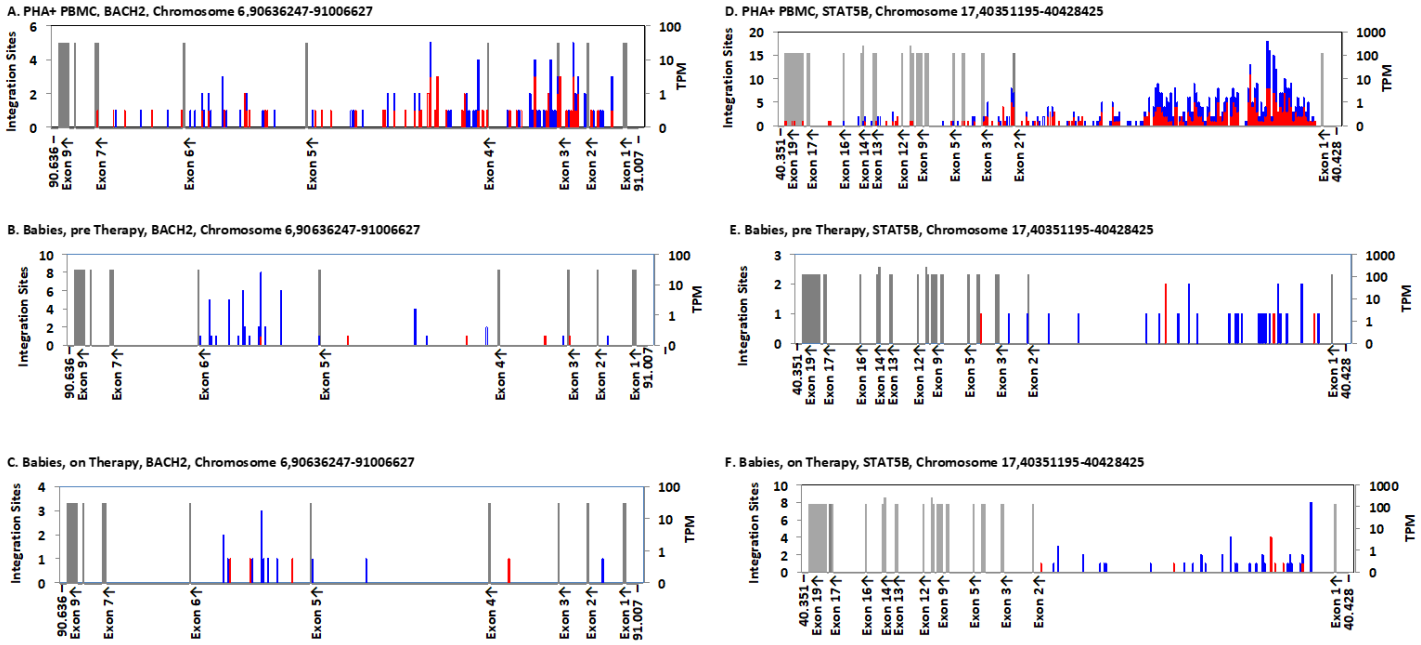
Integration site data from donors whose viremia was suppressed on ART were downloaded from the Retrovirus Integration Database (rid.ncifcrf.gov) (43) from two studies totaling 9 individuals (6, 32) and the OCIs were calculated. OCIs were compared using Mann-Whitney test. Median values for each patient group are marked by red lines.

Figure 3. Global selection against proviruses oriented with-the-gene and selection in two genes for with-the-gene proviruses.



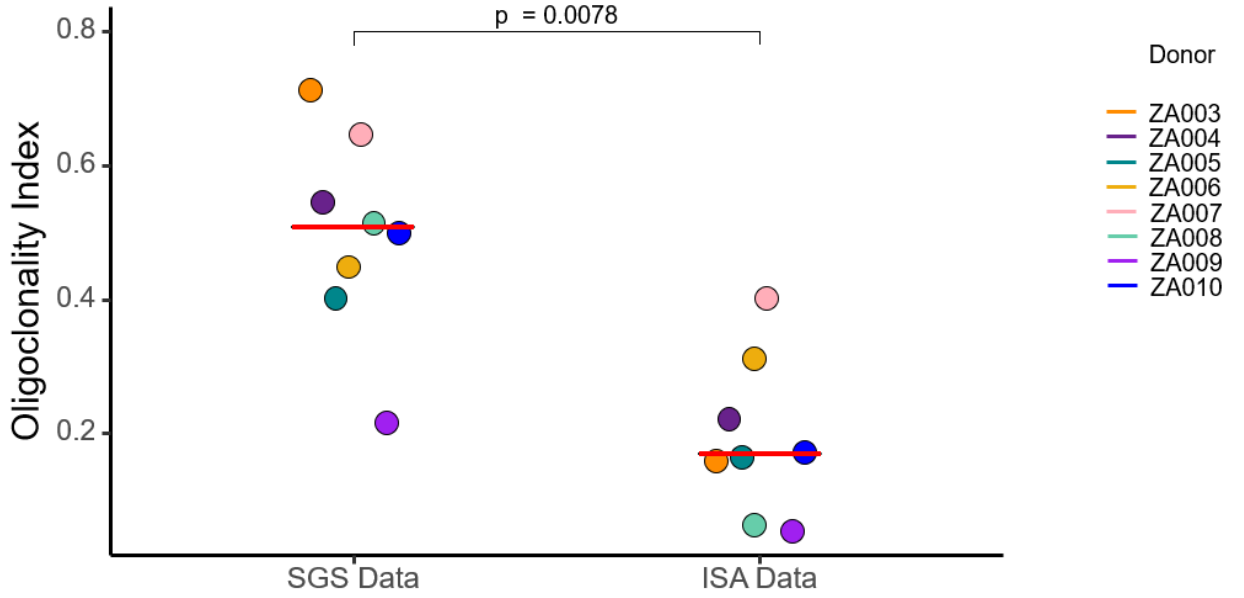
For each of the integration site libraries, unique integration sites for all genes **(A)** and for proviruses integrated in *BACH2* and *STAT5B* **(B)** were plotted as the percentage of integrations against-the-gene (orange) and with-the-gene (green). Significance was assessed via Fisher's Exact test between the *ex vivo* infected PBMC library and the pre-ART and on-ART integration site libraries from children. p-Values for pre-ART comparisons were post-hoc adjusted. The on-ART comparisons were not adjusted because of the differences in the number of independent statistical tests against each library.

Figure 4. Distribution of Integration sites in *BACH2* and *STAT5B*.



The maps show the integration sites of the proviruses in *ex vivo* infected PBMC **(A, D)** (36), infants sampled pre-ART **(B, E),** and children sampled on ART **(C, F).** Each graph shows the entire gene, divided into 250 bins. For *BACH2* **(A-C),** each bin corresponds to ca 1500 NT; and for *STAT5B* **(D-F),** ca 300 NT. Exons (labeled on the X axis, with orientation of transcription shown) are shown as grey bars, whose height indicates the level of expression, in transcripts per million (TPM), as shown on the scale on the right. Note that the resolution of the text sometimes leads to loss of labels of closely spaced exons. The numbers of integration sites in each bin are indicated by the stacked bars, according to the scale on the left, with red indicating the same transcriptional orientation as the chromosome numbering and blue indicating the opposite orientation. In these two genes, blue indicates the number of proviruses in each bin integrated in the same orientation as the gene.

788 **Figure 5. OCIs for single-genome sequencing datasets are significantly higher than OCIs**
789 **derived from integration sites analyses.**



790

791 OCIs were calculated from single-genome sequencing and integration sites data obtained from
792 PBMC of children suppressed for 6-9 years on ART. Significance was assessed by Wilcoxon
793 Signed-Rank Test. Median values are noted by red dash for each group.

794

795 **Table 1. Number of Integration Sites and Infected Cell Clones Detected in Children Prior to and On ART.**

Donor ID	Age at ART initiation (months)	No. of integration sites obtained pre-ART[a]	No. of integration sites detected with >2 breakpoints pre-ART (>1 breakpoint)	Oligoclonality index[b] (pre-ART)	Years suppressed on-ART	HIV DNA cps/10^6 PBMC on-ART[c]	Number of integration sites obtained on- ART[a]	No. of integration sites detected with >1 breakpoint on-ART	Oligoclonality index[b] (on ART)	No. of Integration site matches between pre- and on- ART[d]
ZA002	17.4	1064	27 (50)	0.085	6.87	33	113	9	0.079	7
ZA003	1.8	1386	7 (30)	0.04	8.06	2	148	16	0.161	0
ZA004	2.7	655	5 (13)	0.024	7.92/8.76	24/--	255	25	0.223	3
ZA005	6.0	583	2 (14)	0.027	8.04	9	77	4	0.166	1
ZA006	9.0	486	1 (6)	0.012	7.45	47	137	20	0.313	1
ZA007	9.9	197	4 (5)	0.07	8.24	21	85	8	0.403	2
ZA008	2.2	1293	1 (8)	0.006	6.77	42	225	11	0.065	1
ZA009	2.0	809	0 (16)	0.021	9.13	186	125	5	0.055	0
ZA010	1.8	514	1 (12)	0.027	8.35	5	115	3	0.173	0
ZA011	9.3	432	5 (9)	0.037	7.35	182	149	8	0.092	3
ZA012	5.1	1243	3 (12)	0.01	8.41	12	432	32	0.126	2

Median Values	5.1	655	3 (12)	0.027	8.04	24	137	9	0.161	1

[a]Value obtained by counting an integration from both the 5' and 3' LTR as a single integration site

[b]As described in Gillet, *et al*. and Bangham (41). This value ranges in the interval [0, 1] dependent on the relative size and contribution of integration site clones to the dataset where 0 signifies a completely uniform distribution while 1 signifies a single integration site.

[c]Integrase cell-associated DNA (iCAD) protocol (51)

[d]Matches between pre-ART and on-ART are counted as clones in columns 4 and 9.

802 **Table 2. Analysis of Enrichment of Integration into Specific Genes in vivo[a]**

Chromosome	Gene name[b]	Independent integrations *ex vivo*[c]	Independent integrations in CHER cohort pre-ART	Adjusted p-value[d]	Independent integrations in CHER cohort on-ART	Adjusted p-value[d]
17	*STAT5B*	562	29	0.14	37	4.0E-29
6	*BACH2*	132	31	8.9E-17	16	2.7E-15
All	All other genes	334,920	6,979	>0.05	1,149	>0.05

803 [a]Data shown only for integrations into genes and for which at least 1 integration was detected in
804 both libraries

805 [b]Genic coordinates mapped to hg19

806 [c]*Ex-vivo* dataset contains integration sites from CD8-depeleted PBMCs from two healthy donor
807 patients infected and PHA-stimulated *ex vivo*

808 [d]Adjusted p-value determined by Fisher's Exact Test with post-hoc Benjamini-Hochberg
809 Correction

810

811

812

813

814

815

816

817

818

819

820

821

822

Supplementary Materials

Table S1. Donor Characteristics.

PID	Sex	Time to viral load suppression in years	Pre-ART plasma HIV RNA[a]	ART regimen[b]	Years suppressed on ART	CD4% at on-ART time point
ZA002	Male	1.37	654000	AZT/3TC/LPV/r	6.87	33
ZA003	Male	0.46	>750000	ABC/3TC/LPV/r	8.06	21
ZA004	Male	1.38	>750000	AZT/3TC/LPV/r	7.92/8.76	46
ZA005	Male	0.47	>750000	AZT/3TC/LPV/r	8.04	41
ZA006	Female	0.44	635000	AZT/3TC/LPV/r	7.45	50
ZA007	Male	0.92	>750000	AZT/3TC/EFV	8.24	35
ZA008	Female	0.44	>750000	AZT/3TC/LPV/r	6.77	36
ZA009	Female	3.76	>750000	AZT/3TC/EFV	9.13	29
ZA010	Female	0.46	510000	AZT/3TC/LPV/r	8.35	39
ZA011	Female	2.29	>750000	AZT/3TC/LPV/r	7.35	54
ZA012	Male	0.93	277,000	AZT/3TC/LPV/r	8.41	31

[a]Determined by Roche Amplicor HIV Monitor assay v1.0

[b]ART abbreviations: Zidovudine (AZT), Lamivudine (3TC), Ritonavir-boosted Lopinavir (LPV/r), Efavirenz (EFV)

838 **Table S2. Orientation Bias for Genic Integrations Pre-ART[a].**

Chromosome	Gene name[b]	Unique integrations with the gene (CHER)[c]	Unique integrations against the gene (CHER)[c]	Unique integrations with the gene (*ex vivo*) [c]	Unique integrations against the gene (*ex vivo*) [c]	Adjusted p-value[d]
chr6	*BACH2*	26	5	60	72	0.0019
chr17	*STAT5B*	24	5	284	278	0.0078
chr16	*ANKRD11*	3	16	380	403	0.0281
chr22	*HORMAD2*	11	4	192	241	0.15
chrX	*MECP2*	4	12	228	267	0.45
chr17	*VMP1*	6	12	376	374	0.52
chr17	*GRB2*	11	5	339	336	0.52
chr17	*NPLOC4*	5	10	418	408	0.52
chr17	*POLR2A*	6	10	161	177	0.82
chr11	*MALAT1*	6	9	113	96	0.82
chr11	*PACS1*	18	21	852	821	0.84
chr11	*KDM2A*	9	10	806	724	0.84
chr19	*CARD8*	10	7	331	315	0.84
chr19	*VAV1*	7	8	255	222	0.84
chr17	*RPTOR*	14	12	820	807	0.89
chr17	*CYTH1*	10	9	352	363	0.89
chr22	*TNRC6B*	9	10	444	435	0.89
chr1	*ASH1L*	8	9	378	385	>0.99

839 **[a]**Data shown only for integrations into genes for which at least 15 unique integrations were
840 detected

841 **[b]**Genic coordinates mapped to hg19

842 **[c]**"With" gene and "Against" gene defined as orientation of integrated provirus compared with the
843 sense of the host gene it's integrated into

844 **[d]**Adjusted p-value determined by Fisher Test with post-hoc Benjamini-Hochberg Correction

845

846

Table S3. Orientation Bias for Genic Integrations On-ART[a].

Chromosome	Gene name[b]	Unique integrations with the gene (CHER) [c]	Unique integrations against the gene (CHER) [c]	Unique integrations with the gene (*ex vivo*) [c]	Unique integrations against the gene *(ex vivo)* [c]	p-value[d]
chr17	*STAT5B*	31	6	284	278	6.2E-05
chr6	*BACH2*	12	4	60	72	0.034

[a]Data shown only for integrations into genes for which at least 15 unique integrations were detected in vivo and at least 1 unique integration ex vivo

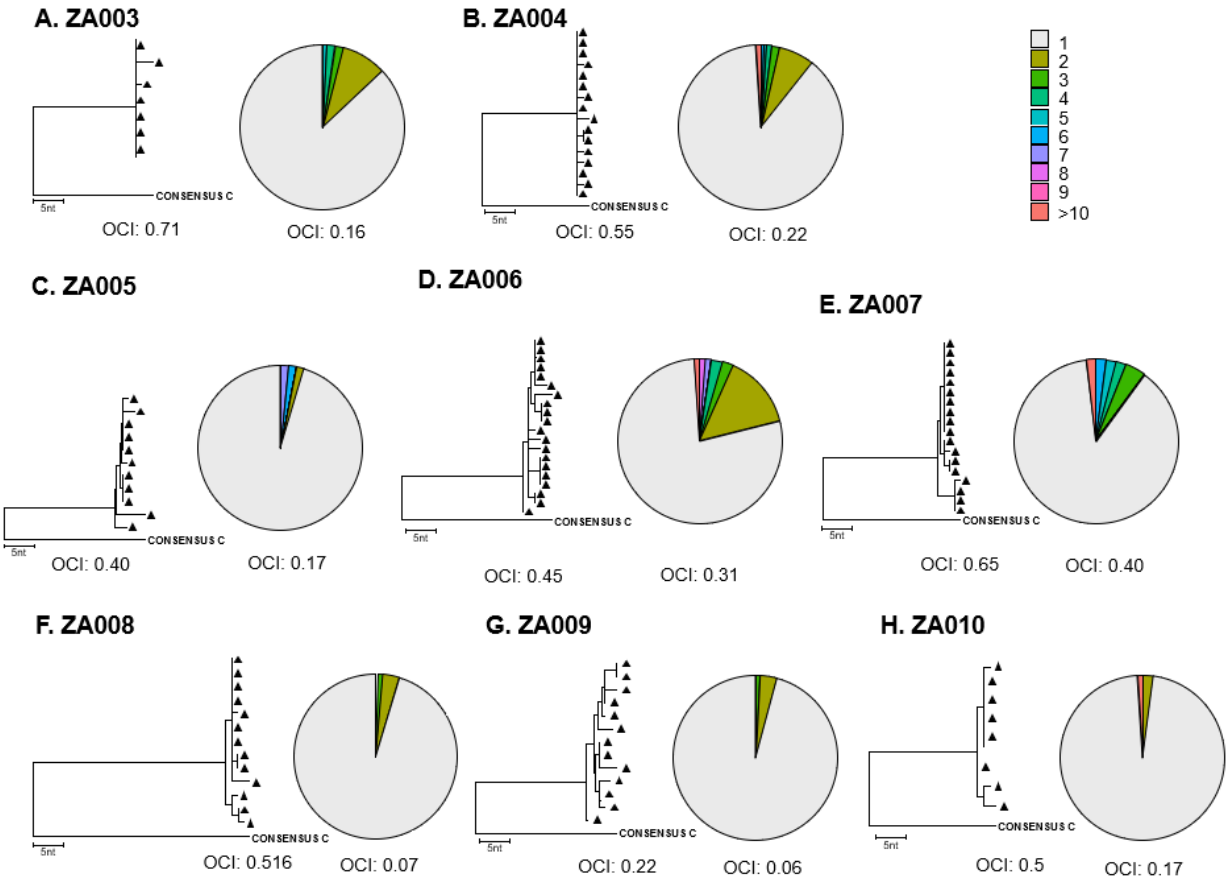
[b]Genic coordinates mapped to hg19

[c]"With" gene and "Against" gene defined as orientation of integrated provirus compared with the sense of the host gene it's integrated into

[d]p-Value determined by Fisher Test – no post-hoc adjustments performed

870 **Figure S1. Number of Detections of Integration Sites.**

871 FF



872

873 For each study participant, a neighbor-joining phylogenetic tree representing gag-pol single
874 genome sequences with its respective OCI value is shown on the left; on the right, a pie chart
875 representing the number of detections of integrations sights by ISA and the respective OCI value.