PLOS PATHOGENS

RESEARCH ARTICLE

# Integration in oncogenes plays only a minor role in determining the in vivo distribution of HIV integration sites before or during suppressive antiretroviral therapy

**John M. Coffin**[1], **Michael J. Bale**[2], **Daria Wells**[3], **Shuang Guo**[3], **Brian Luke**[3], **Jennifer M. Zerbato**[4¤a], **Michele D. Sobolewski**[4], **Twan Sia**[2¤b], **Wei Shao**[3], **Xiaolin Wu**[3], **Frank Maldarelli**[2], **Mary F. Kearney**[2], **John W. Mellors**[4], **Stephen H. Hughes**[2]*

**1** Tufts University, Boston, Massachusetts, United States of America, **2** National Cancer Institute, Frederick, Maryland, United States of America, **3** Cancer Research Technology Program, Leidos Biomedical Research Inc., Frederick, Maryland, United States of America, **4** University of Pittsburgh, Pittsburgh Pennsylvania, United States of America

¤a Current address: The Peter Doherty Institute for Infection and Immunity, Melbourne, Australia
¤b Current address: Swarthmore College, Swarthmore, Pennsylvania, United States of America
* hughesst@mail.nih.gov

**Citation:** Coffin JM, Bale MJ, Wells D, Guo S, Luke B, Zerbato JM, et al. (2021) Integration in oncogenes plays only a minor role in determining the in vivo distribution of HIV integration sites before or during suppressive antiretroviral therapy. PLoS Pathog 17(4): e1009141. https://doi.org/10.1371/journal.ppat.1009141

**Editor:** Ronald Swanstrom, University of North Carolina at Chapel Hill, UNITED STATES

**Received:** November 18, 2020

**Accepted:** March 24, 2021

**Published:** April 7, 2021

**Data Availability Statement:** All relevant data are within the manuscript and its Supporting Information files. Also, the integration site data will also be deposited in the Retroviral Integration site Database (RID https://rid.ncifcrf.gov), and accessible under the PubMed ID of this paper.

**Funding:** This research was supported by the Intramural Research Programs of the NIH, National Cancer Institute, and the Office of AIDS Research. JMC was a Research Professor of the American

## Abstract

HIV persists during antiretroviral therapy (ART) as integrated proviruses in cells descended from a small fraction of the CD4+ T cells infected prior to the initiation of ART. To better understand what controls HIV persistence and the distribution of integration sites (IS), we compared about 15,000 and 54,000 IS from individuals pre-ART and on ART, respectively, with approximately 395,000 IS from PBMC infected in vitro. The distribution of IS in vivo is quite similar to the distribution in PBMC, but modified by selection against proviruses in expressed genes, by selection for proviruses integrated into one of 7 specific genes, and by clonal expansion. Clones in which a provirus integrated in an oncogene contributed to cell survival comprised only a small fraction of the clones persisting in on ART. Mechanisms that do not involve the provirus, or its location in the host genome, are more important in determining which clones expand and persist.

## Author summary

In HIV-infected individuals, a small fraction of the infected cells persist and divide. This reservoir persists during fully suppressive ART and can rekindle the infection if ART is discontinued. Because the number of possible sites of HIV DNA integration is very large, each infected cell, and all of its descendants, can be identified by the site where the provirus is integrated (IS). To understand the selective forces that determine the fates of infected cells in vivo, we compared the distribution of HIV IS in freshly-infected cells to cells from HIV-infected donors sampled both before and during ART. We found that, as previously reported, integration favors highly-expressed genes. However, over time, the fraction of cells with proviruses integrated in highly-expressed genes decreases, implying

Cancer Society and was supported by the NCI through a Leidos subcontract, I3XS110 and research grant R35 CA200421. JWM was supported by the NCI through a Leidos subcontract, 12XS547. This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services or the National Institutes of Health, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. government. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** I have read the journal's policy and the authors of this manuscript have the following competing interests: JWM is a consultant to Gilead Sciences and Merck Co., Inc., and owns share options in Co-Crystal Pharmaceuticals, Inc. JMC is a member of the Scientific Advisory Board of ROME Therapeutics, Inc. All other authors declare no competing interests.

that they grow less well. There are exceptions to this broad negative selection. When a provirus is integrated in a specific region in one of seven genes, the proviruses affect the expression of the target gene, promoting growth and/or survival of the cell. Although this effect is striking, it is only a minor component of the forces that promote the growth and survival of the population of infected cells during ART.

## Introduction

Integration of viral DNA into the host cell genome is an essential step in the replication of HIV. In the course of untreated infection, the majority of infected cells die within a few days [1–3]. A small fraction of cells containing integrated proviral DNA persists for many years of suppressive antiretroviral therapy (ART) and a small fraction of the cells containing intact persistent proviruses is the source of the virus that re-emerges when therapy is interrupted [4]. Understanding the mechanism(s) by which the infected cells persist is central to developing and evaluating strategies to cure HIV infection. Recent studies describing the distribution of integration sites (IS) in cells obtained from infected people have provided important clues about the persistence of infected cells (reviewed in [5, 6]).

The distribution of proviruses in HIV-infected individuals on long-term ART derives from their initial distribution in newly infected cells, modified by selection that promotes their survival, preferential loss, or clonal expansion. In vitro experiments show that the IS for HIV proviruses are widely distributed in the human genome; however, the distribution is far from random, favoring the bodies of highly-expressed genes located in gene-dense regions [7–9]. Retroviral integration exhibits a modest preference for specific bases around the site of integration [10, 11], and there is no discernable preference for the integration of the viral DNA in either orientation relative to the direction of transcription of the target gene [7, 12, 13].

It is likely that integration follows the same rules in cells infected in culture and in a human host. In untreated chronic HIV infection, between one in 100 and one in 1000 CD4+ T cells are infected every day; almost all of the newly infected cells die. In untreated individuals, there are more than $10^8$ new integration events every day, a process that continues for many years. Starting soon after the initial infection, a small fraction of the infected cells survive, growing into clones that become large enough to detect ($>10^5$ cells) in as few as a 4 weeks [14]. Most of the surviving cells (>98%) carry defective proviruses [15, 16]. However, despite the fact that cells that carry intact infectious proviruses represent only a small fraction of the cells that survive, they still add up to a large number [16]. At least some of the infectious proviruses are in a poorly defined, but reversible, state of latency. Such cells persist for decades of ART and can produce progeny viruses that can rekindle the infection when ART is interrupted [17, 18]. Collectively, the intact infectious proviruses constitute the viral reservoir.

The reservoir comprises proviruses in cells that were infected prior to the initiation of ART (and their descendants), and is not maintained by persistent viral replication in blood or solid tissues [4, 14, 19, 20]. Studies of the distribution of IS in cells from infected individuals revealed that at least 40% of the long-lived, persistently infected cells were in a small number of clones each derived from a single infected cell. Clonally amplified cells can persist for at least 11 years [21–23]. Some clonally amplified CD4+ T cells carry infectious proviruses, and can occasionally express infectious viruses with the potential to rekindle the infection when ART is discontinued [22, 24–27].

Although we now know that there is extensive clonal expansion of infected cells in vivo, important questions remain: 1) Integration in at least three genes (*MKL2*, *BACH2*, *STAT5B*)

[21, 23, 28] can promote the survival and/or proliferation of HIV-infected cells. Are there other genes in which proviral DNA provides a selective advantage to the infected cell? 2) Are there other factors that help to reshape the initial distribution of IS in individuals on long term ART? 3) What is the relative importance of provirally-mediated expression of such genes and other factors (e.g., the immune response) in the clonal expansion and persistence of HIV-infected cells?

To address these questions, we prepared a large IS library from stimulated PBMCs infected in culture with HIV [29]. We then compared the distribution of the IS to the patterns of gene expression in these cells and to the distribution of IS obtained from cells from infected people both before ART and during suppressive ART. Our results support the conclusions that integration into oncogenes, while striking, is only a small factor in HIV persistence, and that clonal expansion must be driven largely by factors unrelated to the location of the provirus.

## Results

### Study participants

IS data were obtained from the donors described in the studies listed in S1 Table, and their detailed characteristics can be found in the references therein. All studies had the approval of the relevant institutional IRB, and informed consent was obtained for all the samples collected, as described in the cited studies.

### Integration site datasets

The distribution of IS in people living with HIV on long-term ART can only be properly understood in the context of the distribution of IS at the time the cells were initially infected. Unfortunately, it has not been possible to prepare sufficiently large IS libraries from newly-infected people [14]. We therefore analyzed the distribution of HIV IS in phytohemagglutinin (PHA)-stimulated CD8-depleted PBMC isolated from two HIV-negative individuals. PHA activation was used because HIV prefers to infect mitogen or antigen stimulated cells. Except for a ca. 8-fold decline in the number of IS obtained, there is very little difference in either gene expression or IS distribution in PHA stimulated and unstimulated cells; see (S6 Fig). The PBMC were infected in vitro with replication competent HIV-$1_{\mathrm{BAL}}$ (referred to hereafter as “PBMC”). We compared that distribution to the distribution of IS in cells obtained from infected individuals both pre-ART and on-ART using identical methods. The comparative analysis of the three datasets allowed us to assess the effects of various selective forces–both positive and negative–that modify the distribution of IS during suppressive antiretroviral ART.

Total DNA was extracted from the PBMC two days following infection, to allow time for integration, but not selection for, or against, proviruses integrated in specific sites or specific genes [29]. We used linker-mediated PCR [30] as previously described [21, 31] to identify IS. When we compared the gene distribution of the IS from the PBMC obtained from the two individuals, the correlation coefficient was 0.98 and, for the rest of the analyses, we combined the data from the two sets of IS. The combined dataset from the in vitro-infected PBMC contained 394,975 independent IS. It will be referred to as the “PBMC dataset” for the remainder of the paper.

By combining in vivo IS from several studies [14, 20, 21, 32] (S1 Table), we were able to assemble datasets comprising a total of 15,821 IS from pre-ART samples from 25 HIV-infected donors (hereafter referred to as “donors”) and 53,945 IS from on-ART samples from 36 donors (Tables 1 and S1). The introduction of a shearing step prior to linker ligation, PCR, and paired-end sequencing [30] creates random break points in the flanking host DNA, making it

**Table 1. Integration datasets used in this study.**

| **Set** [a] | **Description** | **Number of sites (% of total)** | | | | |
|---|---|---|---|---|---|---|
| **Genome DNA** | **From the RefSeq Database** [34] | **Total** [3080 Mb (100%)] [b] | **In genes** [1171 Mb (38%)][c] | **In expressed genes** [688 Mb (22%)] | **In exons** [64 Mb (2.1%)] | **In expressed exons** [44 Mb (1.4%)] |
| PBMC infected In vitro: | 2 uninfected donors | 384,975 | 320,918<br>(83%) | 315,803<br>(82%) | 24,863<br>(6.5%) | 24,445<br>(6.3%) |
| **Donor samples:** | All samples from various DRP studies (S1 Table) | | | | | |
| All pre-ART, all breakpoints | 25 Donors | 15,821<br>(1.21) [d] | 12,450<br>(79%) | 12,057<br>(76%) | 905<br>(5.7%) | 874<br>(5.5%) |
| All pre-ART, unique | 25 Donors | 13,283 | 10,520<br>(79%) [e] | 10,189<br>(76%) | 751<br>(5.6%) | 726<br>(5.4%) |
| All on-ART, all breakpoints | 36 Donors | 53,945<br>(1.62) [d] | 41871<br>(78%) | 40,519<br>(75%) | 2,298<br>(4.3%) | 2,187<br>(4.1%) |
| All on-ART, unique | 36 Donors | 33,336 | 25,577<br>(77%) [e] | 24,581<br>(74%) | 1,570<br>(4.7%) | 1,503<br>(4.5%) |
| Total in vivo, all breakpoints | 36 Donors | 69,720 | 54,321<br>(78%) | 52,576<br>(75%) | 3,203<br>(4.6%) | 3,061<br>(4.4%) |
| Total in vivo, unique | 36 Donors | 46,660 | 36,097<br>(77%) | 34,770<br>(75%) | 2,321<br>(4.3%) | 2,229<br>(4.7%) |

a Unique: Integration sites (IS) with more than 1 breakpoint have been counted as 1.

b Figures in brackets indicate the total length (in Mb) of the feature shown, and the percent of the full genome length

c Percentages relative to the values in column 3.

d Amplification ratio (average number of breakpoints/site)

e P values vs PBMC (Fishers Exact Test): Pre, 3.3x $10^{-41}$;on, 2.0 x $10^{-218}$

https://doi.org/10.1371/journal.ppat.1009141.t001

possible to distinguish identical IS that arose by clonal expansion of a single infected cell [21, 30, 31]. After collapsing the IS with multiple breakpoints to a unique site, the two libraries contained 13,283 unique IS pre-ART and 33,336 unique IS on-ART (Table 1).

## In vivo selection against proviruses integrated in highly-expressed genes

HIV preferentially integrates a DNA copy of its genome into highly-expressed genes in gene-rich regions [8, 9, 33]. However, the initial analyses were based on small datasets (<1 IS/gene), used methodology (restriction enzyme cleavage) which has the potential to introduce bias based on local base composition differences, and used cell lines rather than primary cells. To confirm the prior results and extend them to relevant target cells, we mapped the IS in the PBMC dataset into genes from the entire RefSeq gene database [34], slightly modified to remove gene overlaps, yielding a set of 20,207 genes. The overlaps were removed to prevent mapping single IS to two genes. To compare the sites of integration to the expression of the host genes, we performed an RNA-seq analysis using the same PBMC that were used to generate the PBMC IS dataset. Table 1 summarizes the results of this analysis. As expected, most of the IS (83%) were in genes, which comprise only 38% of the human genome, and most (82%) were in expressed genes, which comprise about 22% of the genome. Thus, only 18% of the HIV IS were found in the 78% of the genome that is not in expressed genes (defined as those with ≥0.5 transcripts per million reads, TPM). A similar preference was evident in the pre-ART data in which 79% of the IS were in genes and 76% were in expressed genes. In the on-ART dataset 77% of the unique IS were in genes and 74% of the pre-ART sites were in expressed genes. Although only 2.1% of the genome is in exons, exonic IS comprised 6.5% of the IS in PBMC and somewhat less (5.6% and 4.7%) of the unique IS in pre- and on-ART samples, respectively.

The differences between the percentages of the IS in genes in the in vitro PBMC (83%) and in the on-ART (79%) or pre-ART (76%) datasets, although not large, were highly significant (Table 1). There was an even larger difference in expressed genes (82% vs 76% and 74%). We compared the relationship of IS to gene expression in the three datasets using our non-overlapping gene set (available in S1 Data). All RefSeq genes were divided into 100 bins based on the RNA-seq data (as transcripts per million reads, TPM) obtained from the PHA-stimulated HIV-infected PBMC (Fig 1, green triangles) and also divided into 4 bins, again based on TPM. The mean integration site density (unique IS/Mb, normalized to the mean for each dataset) for the 3 IS datasets when the IS are divided into the 4 bins, is presented in Fig 1A. The same data, divided into 100 bins, are shown in S1 Fig. As expected [8, 9, 33], the results showed a strong correlation between integration site density and gene expression for all three datasets. There was a small, but highly significant, difference in the integration site density in the 3 datasets for the genes with the highest TPM. When the IS data were divided into 4 bins, the fraction of IS that were in the quartile of the most highly-expressed genes was about 10% lower in the pre-ART dataset than in the PBMC dataset, and about 18% lower in the on-ART dataset, compared to the PBMC dataset ($p = 3.3 \times 10^{-41}$). A smaller effect was seen in the second quartile, based

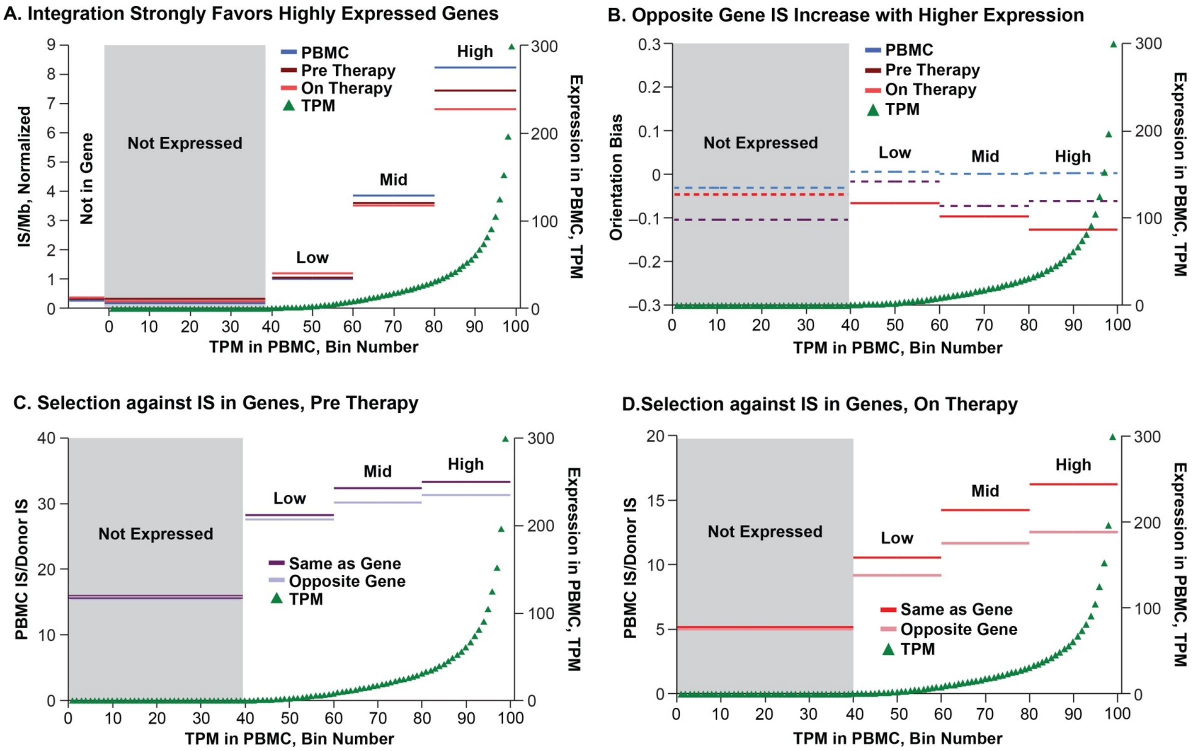

**Fig 1. Gene expression and integration.** The complete set of ca 22,000 RefSeq genes was slightly modified to remove overlaps (see Methods). The non-overlapping genes were divided into either 100 bins or 4 bins, based on transcripts per million reads (TPM) obtained from the RNA-seq analysis of the in vitro infected PBMC. The 100-bin RNA-seq data are shown in green triangles in all panels. The combined IS data are shown for the genes in each of the 4 bins in all panels for PBMC (blue), pre-ART donors (plum) and on-ART donors (red). Darker colors in C and D indicate proviruses oriented in the same direction as the host gene; lighter colors are in the opposite orientation. **A**. Total integration site (IS) density (sites/Mb) in each bin, normalized to the average for the whole genome (125 sites/Mb for PBMC, 4.28 sites/Mb pre-ART, and 10.7 sites/Mb on-ART). **B**. The orientation of the proviruses relative to the host gene was calculated for each bin as (proviruses oriented the same as the gene-proviruses opposite the gene)/(total proviruses). Dashed lines indicate p values (binomial) >0.05. **C** and **D**. Ratios of proviruses per bin for the pre-ART (**C**) or on-ART (**D**) samples. Note that the higher the ratio, the smaller the number of IS in the donor samples relative to the in vitro infected PBMC samples. S1 Fig shows the same results with the 100-bin IS data included.

https://doi.org/10.1371/journal.ppat.1009141.g001

on gene expression (p = 2.0 x $10^{-218}$). No similar effect was seen in the genes in the bins with two lowest levels of expression. This result implies the presence of selection, in vivo, against cells containing proviruses integrated in the most highly-expressed genes, and shows that the selection operated both pre-ART and on ART.

We and others previously reported a preference for proviruses integrated in an orientation opposite to the host gene in HIV-infected individuals and SIV-infected macaques on-ART (5% and 9%, respectively) that is not seen in vitro [21, 23, 29]. There was, as expected, no significant bias for the orientation of the proviruses, relative to the host gene, in the PBMC dataset (Fig 1B). However, enrichment for proviruses orientated opposite to the gene in the on-ART dataset, relative to the PBMC dataset, was evident (S2 Table, p = 1.7 x $10^{-63}$). We also observed a smaller, but significant, enrichment for proviruses in the opposite orientation to the gene in the pre-ART data (S2 Table, p = 0.002).

To understand the nature of the in vivo selection, we separated the IS data based on whether the provirus was oriented with the gene, or opposite the gene, and then plotted IS data relative to the level of expression of the host gene, dividing the genes into four bins based on their levels of RNA expression. There was no significant selection against proviruses that were in the same orientation as the host gene in genes that were not expressed in either the pre-ART or the on-ART dataset (Fig 1B). In expressed genes, there was a significant selection against proviruses oriented with the gene in the on-ART dataset [p values ranging from about 0.013 to 5 x $10^{-47}$ (S2 Table)]; the strongest selection was seen in the highly-expressed (>30 TPM) genes. A similar pattern was seen with the pre-ART data, although the effects were smaller and the effect for the genes that were expressed at a low level was not significant. While the difference in IS distribution between the PBMC and the in vivo samples could have reflected differences in initial integration preferences between the two conditions, the differences between pre- and on-ART samples are best explained by selective forces that affect the survival or proliferation of the infected cells. The on-ART data (Fig 1D) also showed that there is weaker selection against cells with proviruses integrated in highly-expressed genes in the opposite orientation to the gene (p = 2.3 x $10^{-48}$); again the strength of selection decreased with decreasing levels of gene expression. Similar, but much weaker, selection was seen in the pre-ART dataset (p = 0.011).

## Selection involving proviruses integrated in individual genes

To examine the contribution of proviruses in genes that are favored for HIV integration (hot spots) we ranked genes in both the PBMC and the on-ART datasets by IS density, as unique sites per kb (Table 2). For the most part, the IS site density in individual genes showed that the in vitro and in vivo data were in good agreement (Table 2 and Fig 2). We and others previously reported that proviruses integrated in certain introns of three genes (*BACH2*, *MKL2*, *STAT5B*) can confer a selective advantage for the clonal expansion and/or survival of infected cells [21, 23, 28]. The in vivo enrichment of IS in these genes is due to post-integration selection, an effect that has sometimes been misinterpreted in the literature as reflecting preferential "hot spots" for integration [35, 36].

To ask if there are other genes in which an HIV provirus could provide a selective advantage to the host cell in vivo, we compared, for each gene, the fraction of IS in the in vitro PBMC dataset to the fraction in the pre-ART and on-ART in vivo datasets (Fig 2). The top two panels show the same data using two different scales. Panel A shows the data for all of the genes. The inset shows the top 30 genes, showing that, for most genes, there is a strong correlation between the IS distribution in vitro and in vivo. There is one obvious exception to this correlation in the top 30 genes, *STAT5B* (at number 28 in PBMC), a gene in which HIV proviruses are known to confer a selective advantage on the host cell [28]. Panel B shows the top

**Table 2. Genes that are favored targets for integration in PBMC infected in vitro and the on-ART samples.**

| Gene | PBMC | | | On-ART, unique | | |
|---|---|---|---|---|---|---|
| | Sites | Sites/kb [a] | Rank [b] | Sites | Sites/kb | Rank |
| *SF1* | 382 | 26.8 | 1 | 14 | 0.98 | 10 |
| *MALAT1* | 209 | 23.8 | 2 | 27 | 3.08 | 2 |
| *PSMD13* | 357 | 22.1 | 3 | 14 | 0.87 | 14 |
| *PSMB9* | 125 | 22.0 | 4 | 3 | 0.53 | 62 |
| *RNPS1* | 287 | 18.8 | 5 | 15 | 0.98 | 11 |
| *SNHG1* | 24 | 18.0 | 6 | 2 | 1.50 | 5 |
| *NEAT1* | 393 | 17.3 | 7 | 41 | 1.80 | 3 |
| *NAA38* | 453 | 16.1 | 8 | 35 | 1.24 | 6 |
| *DDX17* | 362 | 15.8 | 9 | 12 | 0.52 | 65 |
| *ATP6V1G2-DDX39B* | 171 | 15.7 | 10 | 4 | 0.37 | 161 |
| *STAT5B* | 562 | 7.3 | 103 | 268 | 3.47 | 1 |
| *NDUFB10* | 21 | 8.5 | 68 | 4 | 1.63 | 4 |
| *ZNRD1* | 40 | 10.9 | 45 | 4 | 1.09 | 7 |
| TNF | 14 | 5.1 | 229 | 3 | 1.08 | 8 |
| *LSM2* | 104 | 10.8 | 46 | 10 | 1.04 | 9 |

a Number of unique IS per gene in PBMC divided by its length.

b All genes ranked in the top 10 by IS density in either PBMC or on-ART DNA are shown, ordered by PBMC rank.

https://doi.org/10.1371/journal.ppat.1009141.t002

2600 genes by IS frequency in PBMC. Although there is some scatter, the in vivo IS data closely mirror the PBMC data.

Despite the overall selection against proviruses in expressed genes, there were a few genes, circled in Fig 2B, in which there was a higher fraction of proviruses in vivo in the on-ART dataset than in vitro, implying in vivo selection for cells with HIV proviruses in these genes. As will be discussed below, proviruses in these genes show other evidence for a positive in vivo selective effect in the host cells–clustering of IS within the gene and a strong bias toward selection for proviruses in the same orientation as the gene (Table 3).

Panel C is a Venn diagram that shows the extensive overlap between the genes that are favored for integration in the 3 datasets. Of 4,247 genes with at least one IS in the pre-ART dataset, 4090 (96%) also had at least one IS in the PBMC dataset (there were 11,422 genes IS with at least one IS in PBMC). Similarly, of the 6,366 genes in the on-ART dataset, only 422 (6.6%) had no IS in either of the other two sets, and 5,099 (93%) of the genes with at least one IS in the on-ART dataset had at least one IS in PBMC. These comparisons validate the use of the in vitro PBMC data as a surrogate for the initial distribution of IS in vivo and show that the strongest determinant of the overall distribution of HIV proviruses in infected individuals, both pre-ART and on-ART, is the initial distribution at the time the cells (or their ancestors) were infected.

## Mapping and comparing the distribution of integration sites in vivo and in vitro

To help compare the distribution of IS in different datasets, we developed an Excel-based tool which can be used to view different sized regions of the host genome ranging from a selected portion of a gene to a whole chromosome. The tool, available as an Excel workbook in S1 Data, makes it possible to compare the IS site distribution and gene expression levels for up to 5 IS datasets at a time (S2 Fig). The workbook, including the visualization tool, gene list, and

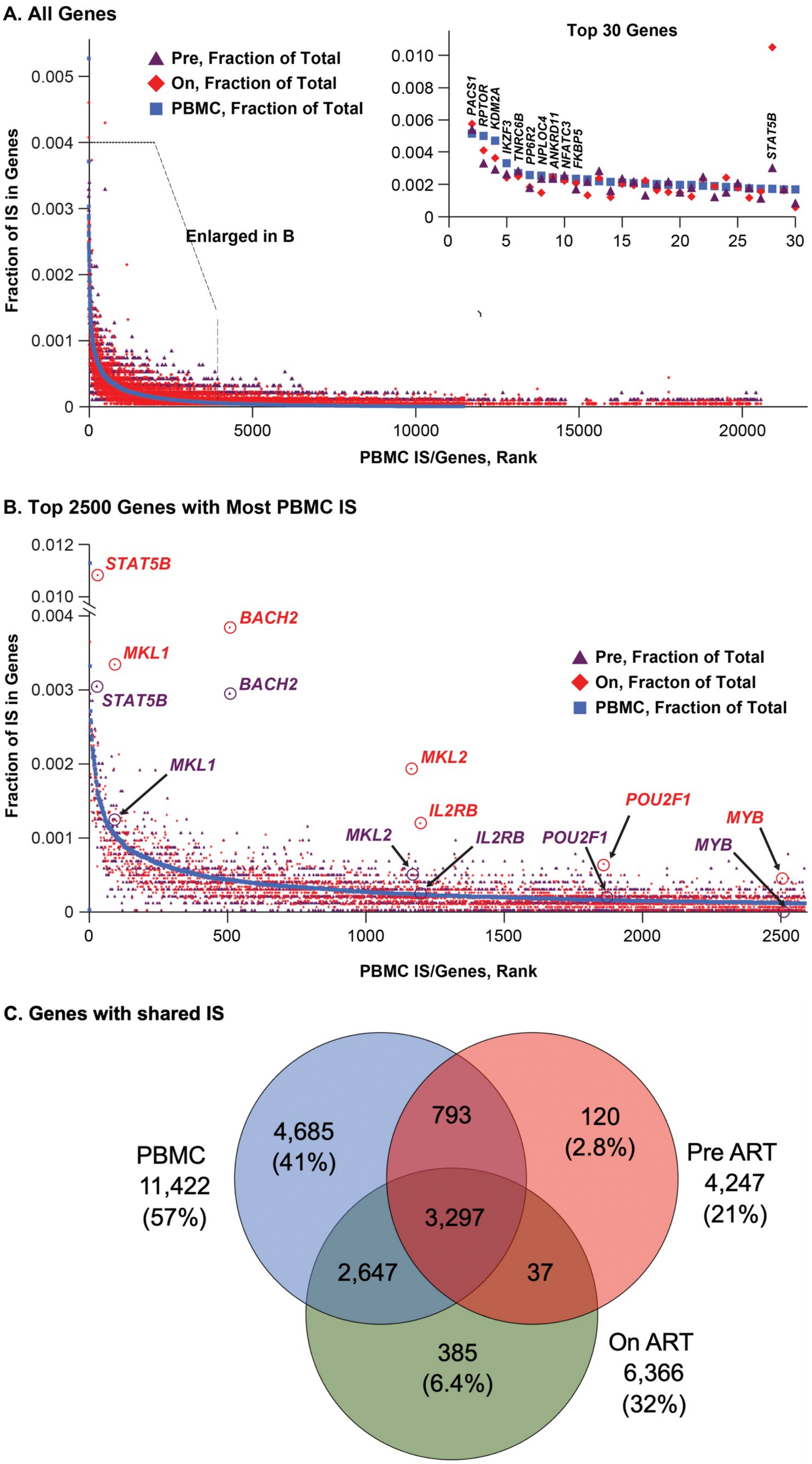
A. All Genes
Pre, Fraction of Total
On, Fraction of Total
PBMC, Fraction of Total
Fraction of IS in Genes
0.005
0.004
0.003
0.002
0.001
0
Enlarged in B
0
5000
10000
15000
20000
PBMC IS/Genes, Rank
Top 30 Genes
0.010
0.008
0.006
0.004
0.002
0
0
5
10
15
20
25
30
PACS1
RPTOR
KDM2A
IKZF3
TNRC6B
PP6R2
NPLOC4
ANKRD11
NFATC3
FKBP5
STAT5B
B. Top 2500 Genes with Most PBMC IS
Fraction of IS in Genes
0.012
0.010
0.004
0.003
0.002
0.001
0
STAT5B
BACH2
MKL1
STAT5B
BACH2
MKL1
MKL2
IL2RB
POU2F1
MYB
MKL2
IL2RB
POU2F1
MYB
Pre, Fraction of Total
On, Fracton of Total
PBMC, Fraction of Total
0
500
1000
1500
2000
2500
PBMC IS/Genes, Rank
C. Genes with shared IS
PBMC
11,422
(57%)
4,685
(41%)
793
120
(2.8%)
Pre ART
4,247
(21%)
3,297
2,647
37
385
(6.4%)
On ART
6,366
(32%)

**Fig 2. Comparative ranking of genes by integration frequency.** The non-overlapping RefSeq genes were ranked according to the number of unique IS in the PBMC in vitro dataset (blue squares), along with the number of sites in the same genes in the pre-ART (plum triangles) and on-ART (red diamonds) datasets. **A**. All 20,207 genes in the dataset are shown; a little over half have one or more IS in the in vitro PBMC dataset. Points with no IS have been removed for clarity, and the genes with no IS in PBMC are assigned a rank at random. The inset shows an expanded view of the 30 genes with the most IS. The top 10 are labeled, as is *STAT5B* (number 28). **B.** Expanded view of the 2500 genes (indicated by the box in **A**) with the most IS. The 7 genes in which an integrated provirus directly contributes to persistence and/or clonal expansion of the infected cell are circled and labeled. Some pre-ART points are indicated with arrows for clarity. **C.** Venn diagram showing the genes with at least one IS in the three datasets, and how those genes are shared among the datasets. Numbers indicate the number of genes in each category.

https://doi.org/10.1371/journal.ppat.1009141.g002

the RNA-seq integration data used in this study is available as S1 Data. The tool divides any chosen region of the genome into 250 bins. The number of IS per bin is shown as vertical bars; red indicates IS in the same orientation as the numbering of the chromosome, blue, the reverse orientation. The levels of RNA (TPM) in HIV-infected PBMC are shown as grey boxes. The locations of genes are indicated, as is their transcriptional orientation. If a whole gene is chosen to view, it is shown broken down into exons and introns,

Fig 3 shows a comparison of the distribution of IS in the on-ART dataset and the in vitro PBMC dataset. We chose to show the data for chromosome 16 because it contains *MKL2*, one of the genes in which an HIV provirus can provide positive selection for the host cell in vivo (see Fig 2B) [21]. The IS data are shown at several scales, from the whole chromosome (bin size ca 355,000 bp, Fig 3A) down to a fraction of a gene (bin size 150 bp, see Fig 4B). At the three largest scales, the pattern of IS is strikingly similar in the PBMC and on-ART datasets, and closely matches the distribution of expressed genes.

## Clustered IS in specific genes

The one exception to the strong similarity in the distribution of the proviruses in chromosome 16 in the PBMC and on-ART datasets is in the *MKL2* gene (also known as *MTRFB*), marked by an arrow in the lower panel of Fig 3A. This difference maps to the *MKL2* gene (Fig 3B and 3C). A map of the *MKL2* gene, showing the location of exons in addition to the distribution of the IS, is shown in Fig 4A. As Fig 3 shows, this gene is not favored for integration in PBMC. However, in the on-ART dataset, proviruses are found in 30 different sites within a 7 kb region, encompassing introns 4–6 of the *MKL2* gene (Fig 4B). Most of these proviruses were in clonally amplified cells (Fig 4B, compare the bottom 2 panels). The probabilities that the

**Table 3. Genes in which integrated provirus can be selected in vivo [a].**

| Gene name | Unique IS in genes on-ART | IS in PBMC | PBMC/on ART (Enrichment) [b] | Enrichment probability (Poisson)) | Provirus orientation: same as gene/opposite | Orientation probability (Binomial) | Selected IS relative to protein coding introns |
|---|---|---|---|---|---|---|---|
| All genes | 25,731 | 326,033 | 12.67 (1.0) | | 11,476/14,255 | $7.0 \times 10^{-60}$ | |
| *STAT5B* | 268 | 562 | 2.1 (6.0) | $1.4 \times 10^{-114}$ | 197/71 | $3.5 \times 10^{-15}$ | Upstream |
| *BACH2* | 98 | 132 | 1.3 (8.7) | $9.2 \times 10^{-56}$ | 71/20 | $3.6 \times 10^{-08}$ | Upstream |
| *MKL2* | 49 | 69 | 1.4 (8.5) | $5.7 \times 10^{-27}$ | 40/6 | $1.6 \times 10^{-07}$ | In between |
| *MKL1* | 85 | 331 | 3.9 3.2) | $34.7 \times 10^{-19}$ | 53/30 | $7.6 \times 10^{-03}$ | In between |
| *IL2RB* | 30 | 68 | 2.3 (4.8) | $1.8 \times 10–^{4}$ | 17/9 | $8.4 \times 10^{-02}$ | Upstream |
| *MYB* | 11 | 31 | 3.1 (4.1) | $2.3 \times 10^{-05}$ | 10/0 | $9.8 \times 10^{-04}$ | In between |
| *POU2F1* | 15 | 43 | 2.9 (4.4) | $7.5 \times 10^{-07}$ | 10/5 | $1.5 \times 10^{-01}$ | Upstream |

a From S2 Data.

b Column 2/Column 1*12.67.

https://doi.org/10.1371/journal.ppat.1009141.t003

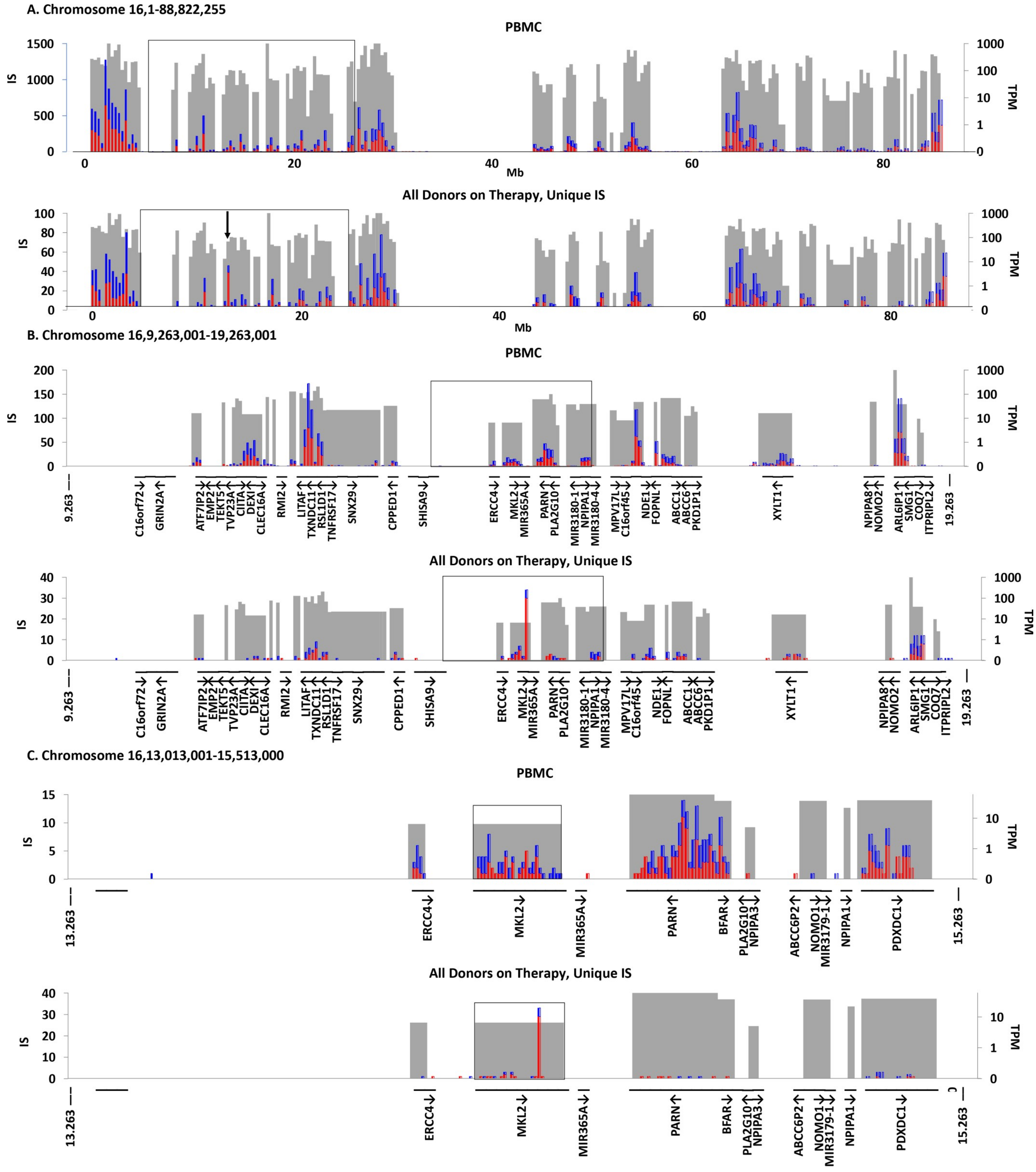

**Fig 3. Analysis of IS in MKL2.** Using the Excel-based application described in S2 Fig, a selected region of the genome of any size is divided into 250 bins. The IS in each bin are tabulated and the number of IS in the bin is shown as a bar, with the orientation of the provirus relative to the numbering of the chromosome indicated by color (red for the same as the chromosome and blue for the opposite). The grey boxes indicate the location and relative RNA level of the genes in each bin. RefSeq genes (from the hg19 sequence database) are shown at the bottom of each plot, with arrows indicating their orientation relative to the numbering of the chromosome. The figure shows the distribution of genes and IS on chromosome 16 at 3 different scales. In all panels, the top image shows the distribution of IS from PBMC infected in vitro; the bottom, the distribution of the unique IS data from on-ART donors. **A.** Entire chromosome 16

(ca 353,000 bp/bin); the arrow shows the region in which IS are enriched in the on-ART samples. The box shows the region enlarged in panel B. Because <10% of the 824 genes could be accommodated on the X-axis, chromosomal position is shown instead. **B.** A 10Mb region of chromosome 16, (40 kb/bin) centered on MKL2. Again, the box indicates the region expanded in panel C. **C.** A 2 Mb region of chromosome 16 (8 kb/bin) centered on *MKL2*. The boxed region is expanded in Fig 4A.

https://doi.org/10.1371/journal.ppat.1009141.g003

cluster would exist by chance or that almost all of the proviruses in the cluster would be in the same orientation are both extremely small (2.5 x $10^{-24}$ and 3 x $10^{-5}$, respectively). These data support our previous conclusion that this cluster of IS is the result of selection for cells in which a provirus integrated in a small intronic region of *MKL2* altered the normal expression and/or structure of the *MKL2* protein. Misexpression of the *MKL2* protein would, in turn, provide a selective advantage to cells carrying these proviruses.

In addition to *MKL2*, HIV proviruses integrated in two other genes, *BACH2* and *STAT5B* have been previously shown to contribute to the growth and persistence of infected cells in individuals on ART [21, 23, 28]. Proviruses in these genes share characteristics with the *MKL2* proviruses: enrichment in vivo relative to their frequency in PBMC infected in vitro; clustering in one or a few neighboring introns; and orientation of the proviruses is the same as the host gene (Fig 5 and Table 3). The current dataset provides strong statistical support for selection of proviruses integrated in *BACH2* and *STAT5B* on-ART (Table 3), as well as evidence of selection for proviruses in these two genes pre-ART (Fig 2).

It is possible that there are additional genes in which HIV proviruses can provide a selective advantage for the host cell. We analyzed the total on-ART IS dataset for the three hallmarks of selection for cells with proviruses in specific genes: 1) enrichment relative to the PBMC dataset; 2) enrichment for proviruses in the same orientation as the gene; and 3) clustering of the IS in one or a few nearby introns (Table 3).

As expected, when all 5,944 genes with IS in both the on-ART and the PBMC datasets were analyzed and ranked by the combined p value for the relative number of IS and for orientation bias (S2 Data), the three top ranking genes were *BACH2*, *STAT5B*, and *MKL2*. Four more genes were identified using the new on-ART dataset: *MKL1* (also known as *MRFTA*), a paralogue of *MKL2* [37]; *IL2RB*, encoding the β chain of the IL-2 receptor (https://www.ncbi.nlm.nih.gov/gene/3560); *MYB*, a known protooncogene originally identified in an avian retrovirus [38], and *POU2F1*, a ubiquitous OCT-1 motif binding transcription factor, whose misexpression is associated with a variety of cancers [39]. All of these genes have been linked directly to cell growth or survival (see Discussion). In all of these genes, the proviruses in the on-ART dataset displayed a strong bias for integration in the same orientation as the gene and the number of IS was significantly greater than what would have been predicted based on the PBMC dataset (Fig 5 and Table 3). In each of the seven genes, the IS in the on-ART datasets showed evidence of clustering in specific introns. In the current on-ART dataset, these seven genes were the only ones for which there was significant evidence for positive selection of cells with an integrated HIV provirus in vivo. Taken together, the 555 independent IS in these genes in the on-ART data comprised about 2.2% of the unique IS in genes. Because our on-ART dataset is large, we were able to identify genes for which the number of IS represented a very small fraction of the total IS (*MYB* is 10/33,336 unique IS, or 0.003%). There may be additional genes that we have not identified in which proviral integration could lead to positive selection; however, the fraction of the total IS in such genes must be very small (see Discussion).

In the analysis of the on-ART dataset, we identified a number of genes for which there was apparent enrichment of proviruses oriented opposite to the host gene. In 4 out of 6 cases, these genes had fewer total IS in the on-ART dataset than would be expected based on the PBMC dataset (S3 Table), consistent with stronger than average selection against cells with proviruses

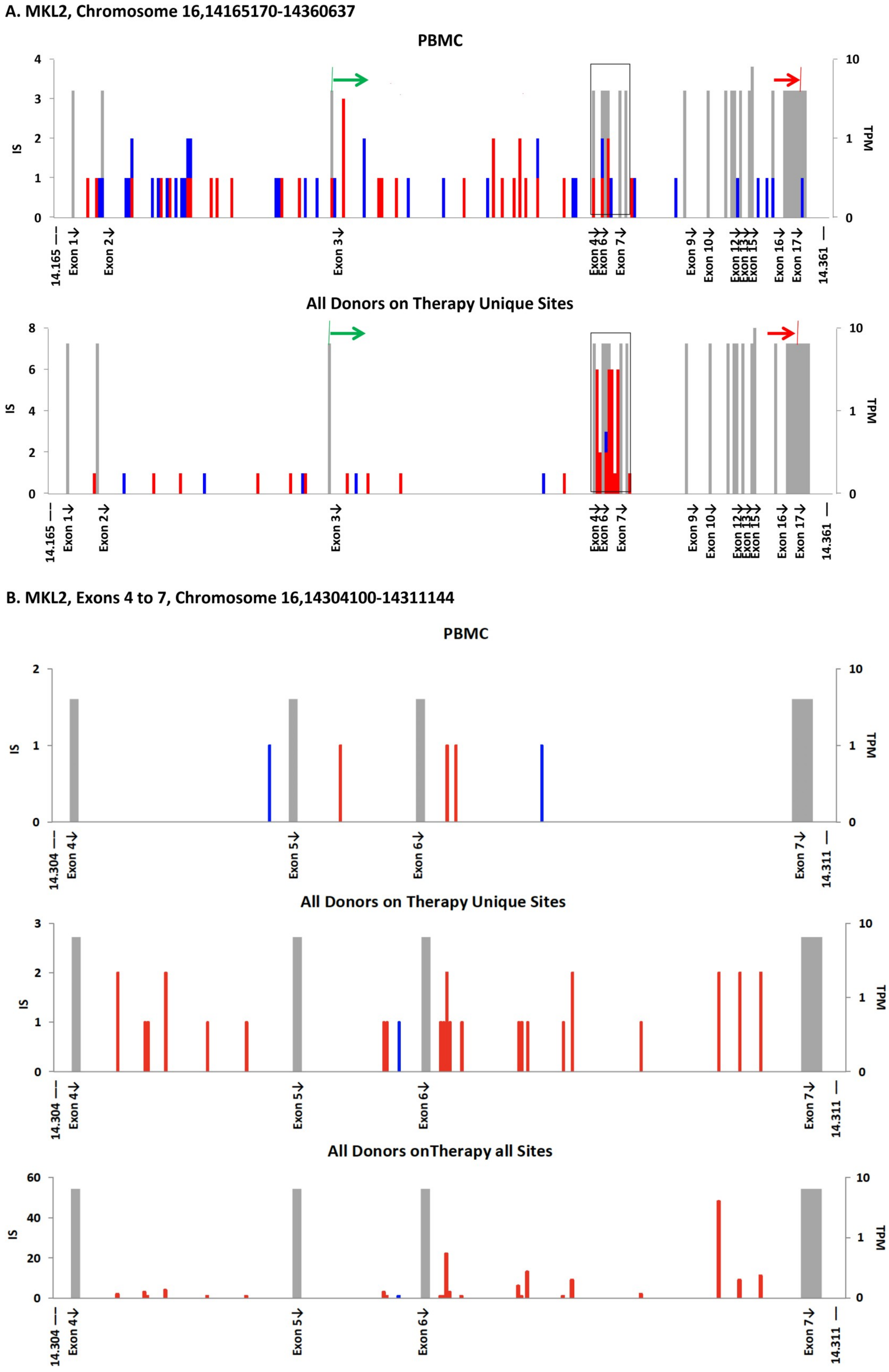

**Fig 4. IS in *MKL2* A.** The complete *MKL2* gene (308 bp/bin). The gray bars show the position of exons, with their height indicating the expression level of the gene in TPM, as in Fig 3. The protein coding region is indicated by the colored arrows

above the maps. The boxed region is enlarged in panel B. **B.** Cluster of *MKL2* IS on-ART in introns 4–6 (50bp/bin). Top, in vitro PBMC IS; middle, on ART unique sites; bottom on-ART, all sites.

https://doi.org/10.1371/journal.ppat.1009141.g004

integrated in the same orientation as the gene. Unlike the seven genes in which there was evidence of selection for a provirus in the same orientation as the gene, in the four genes with reduced same-sense integration, the overall distribution of IS within the genes appeared to be similar to that seen in the PBMC dataset, without obvious clustering (S3 Fig). One of the 6 genes listed in S3 Table (*SLC6A16*) had a statistically significant increase in the number of on-ART IS relative to the PBMC IS, suggesting that there might have been selection for cells with proviruses in the gene, despite their opposite-sense orientation. In both the PBMC and on-ART datasets, there was obvious clustering of IS in a large *SLC6A16* intron preceding the start of translation (S3C and S3D Fig) However, this gene, encoding a transporter involved in neurotransmission [40], is unlikely to affect T cell growth.

We also considered the possibility that there could be regions outside of genes where proviruses could provide a selective advantage to the infected cell in vivo. We used an unbiased scan to look for evidence of clustering of proviruses within 10-kb regions in the on-ART dataset (S4 Fig and S4 Table). With the exception of genes that had already been described, all of the clusters that were found in the on-ART dataset were also present in the PBMC dataset. Sites in which there was clustering in both the in vitro and in vivo dataset are regions that are favored for integration "hot spots"–this is in contrast to the clusters that arise in vivo from post-integration selection. Not surprisingly, clusters of proviruses were found in some of the genes that are the most favored targets for integration, including *PACS1* (S4F Fig), the gene which had the largest number of sites in the infected PBMC (Fig 2A, Inset), and the lncRNA genes *NEAT1* and *MALAT1* (S4E Fig), which are among the genes with the highest density of IS in vivo and in vitro (Table 2). One of the hot spot clusters identified in both the PBMC and on-ART datasets appeared to be intergenic (S4 Fig, panel H). However, analysis of the RNA-seq data revealed a transcription unit in this region which has not yet been annotated.

## Clonal amplification of HIV-infected cells

In the analyses presented thus far, we collapsed all the breakpoints associated with each IS to one, a step that simplified the comparisons of the in vitro and in vivo data. One of the most striking features of the distribution of HIV IS in vivo is the presence of identical IS that derive from clonally-amplified cells. The effect of clonal amplification on the IS data can be seen by comparing the distribution of IS in the *MKL2* cluster in Fig 4B before (bottom panel) and after (middle panel) the amplified sites were collapsed down to unique sites. In this case, the 30 unique IS were found a total of 144 different times–an average amplification ratio of nearly 5, ranging from 1 (no amplification) to more than 25. Overall, the amplification ratio for all sites across the entire genome was 1.6 for the on-ART samples, and 1.2 for the pre-ART samples (Fig 6). That the pre-ART ratio is >1 confirms that infected cells can grow into clones before ART is initiated [14] The amplification ratios were independent of whether the proviruses were inside or outside of genes, the orientation of the proviruses relative to the host genes, and of the level of expression of the host gene (Figs 6 and S4). The amplification ratios for the proviruses in the seven genes where proviruses could be selected, although quite variable, were, on average, not significantly different from all the other genes (Fig 6B). When the amplification ratios for each site were ranked and plotted in decreasing order, the sizes of the clones in the seven genes were scattered across the distribution (S5 Fig). Thus, the degree of clonal amplification cannot be taken as evidence of positive selection for a provirus in a specific location, and such proviruses are not a significant contributor to the overall level of clonal expansion.

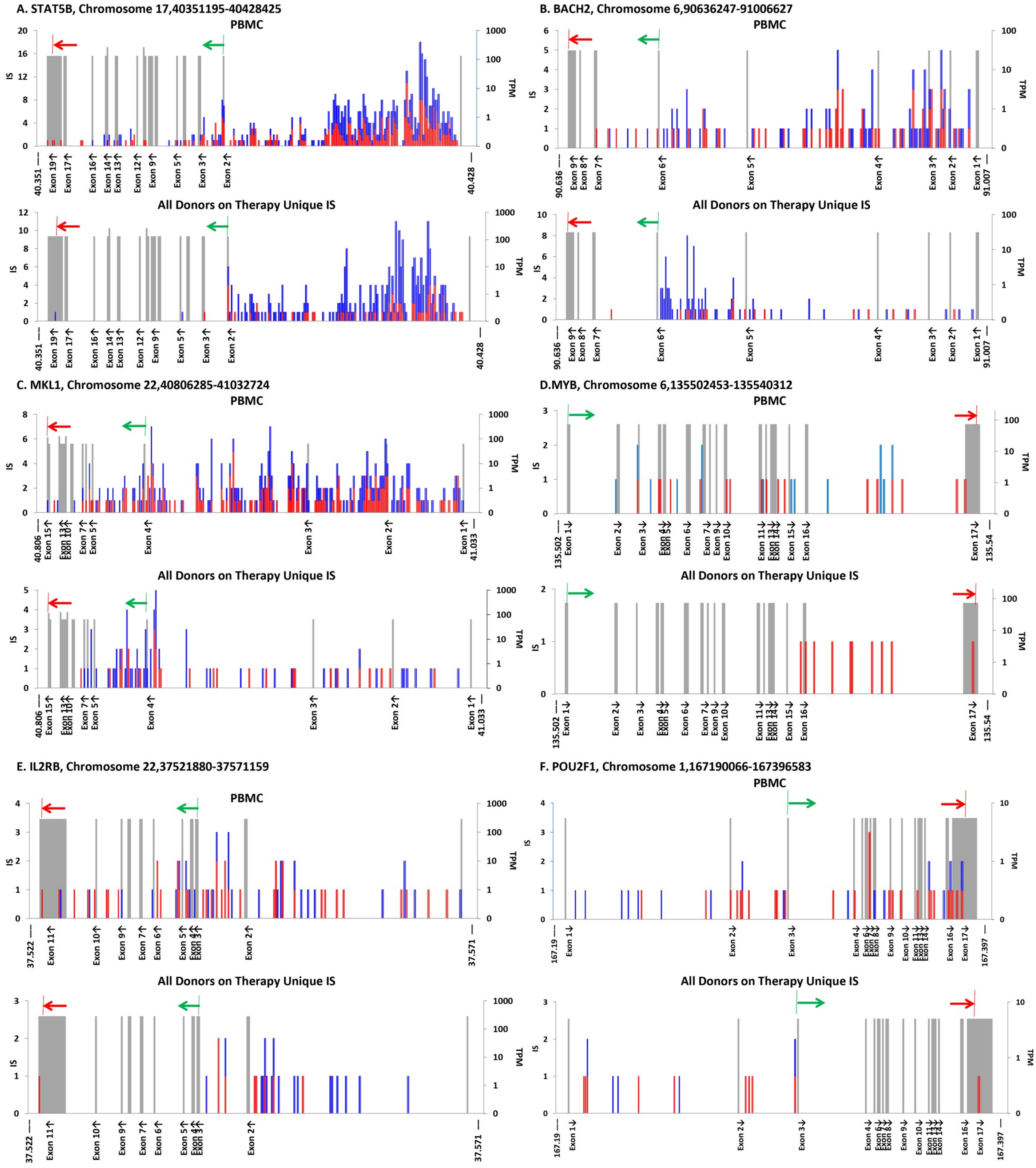

**Fig 5. HIV IS in other genes in which proviruses can give the host cell a selective advantage.** Maps were generated for each of the genes in Table 3 as described in Fig 4. For each pair of panels, the map at the top shows the IS distribution for the PBMC infected in vitro, and the on-ART distribution is at the bottom **A.** *STAT5B* (309 bp/bin). **B.** *BACH2* (1.48 kb/bin). **C.** *MKL1* (906 bp/bin). **D.** *MYB* (151 bp/bin). **E.** *IL2RB* (197 bp/bin) F. *POU2F1* (828 kb/bin).

https://doi.org/10.1371/journal.ppat.1009141.g005

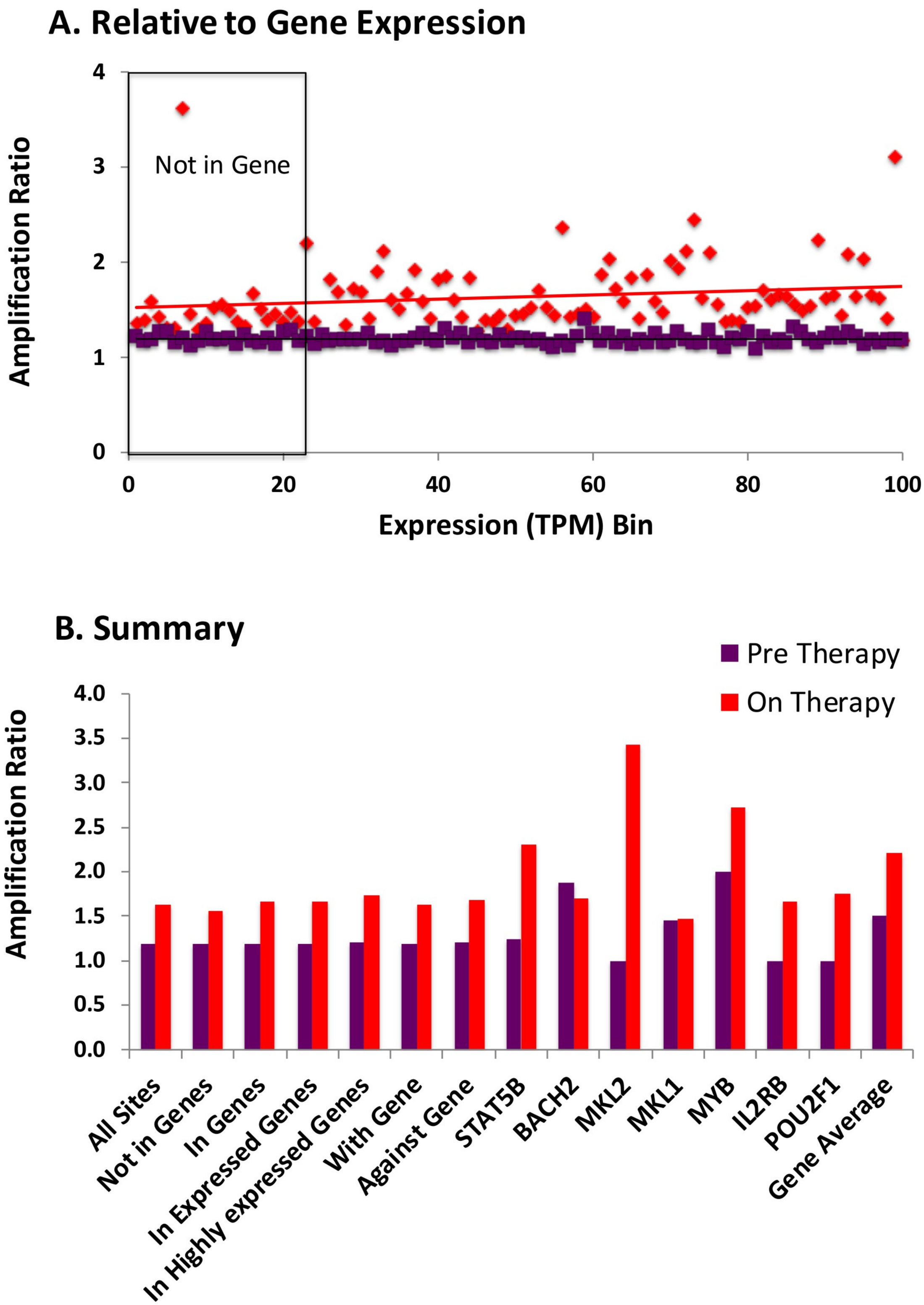

**Fig 6. IS distribution, RNA level, and clonal amplification.** The amplification ratio for all integrated proviruses was calculated using (total number of IS)/(number of unique IS). **A.** Each RefSeq gene in the non-overlapping dataset was assigned to one of 100 bins (ca. 200 genes in each bin) based on the level of RNA for the gene (as TPM in the in vitro infected PBMC), and the overall amplification ratio was calculated for the sites in each bin. Non-expressed genes were assigned to bins at random. The regression lines shown have a slope of 0.0008 and 0.0013 (p = 0.88 an 0.12) for pre-ART and on-ART bins, respectively. **B** shows the clonal amplification ratios pre-ART and on-ART. Color coding is the same as in previous figures: Plum, pre-ART; red, on-ART.

https://doi.org/10.1371/journal.ppat.1009141.g006

**Table 4. IS detected in centromeric repeat DNA[a].**

| | Genome (Mb) | PBMC[c] unique sites | Sites on therapy[d] | | |
|---|---|---|---|---|---|
| | | | Total | Unique | Amp ratio |
| All | 3,080 | 216,054 | 64,283 | 37,967 | 1.69 |
| Outside genes[b] (%) | 524 (17%) | 36,729 (17%) | 14,142 (22%) | 8,732 (23%) | 1.62 |
| In centromeres (%) | 216 (7%) | 178 (0.082%) | 129 (0.20%) | 76 (0.20%) | 1.70 |

a Data from additional assays of some samples (S1 Table) were analyzed for IS in centromeric repeat DNA.

b Recalculated using the proportions in Table 1 to estimate the numbers of total and unique IS for the different-sized samples.

c A portion of the data from PBMC infected in vitro. Because "amplification" is due only to DNA replication immediately following infection, it is without significance here and is not reported.

d From the donors (S1 Table), including additional samples that were taken after those reported in the rest of the manuscript.

https://doi.org/10.1371/journal.ppat.1009141.t004

## Integration in centromeric repeats

The complex nature of the alphoid repeat sequences precludes precise mapping of IS in centromeres, and, for this reason, they are not annotated in either the hg 19 or hg38 genome database and were not detected using our standard pipeline to identify and locate IS, even though the centromeres comprise approximately 7% of the genome [41]. [See, for an example, the large gene-free area around 40 Mb of chromosome 16 (Fig 3)]. It has recently been reported that in rare HIV-infected individuals who maintain very low levels of viremia in the absence of ART (known as elite controllers) the intact proviruses are preferentially found within centromeric sequences [42]. We used a manual protocol to search specifically for centromeric IS in a portion of the data from the PBMC and on-ART datasets. Although finding IS in centromeres is not trivial, and we are not confident that we identified all the centromeric IS in the datasets, Table 4 shows that there are detectable IS in the centromeric repeats in both the PBMC and on-ART datasets. The data show that the amplification ratio of the on-ART proviruses (1.70) is not significantly different from the proviruses found outside genes (1.62, P = 0.77). Nor is the amplification ratio in centromeric proviruses different from proviruses that are integrated in genes. Proviruses integrated in centromeric regions do not appear to be expressed and there are no known genes in centromeres. Thus, these results support our conclusion that factors other than the misexpression of genes with integrated proviruses are the primary factors that control the clonal expansion of infected cells.

## Discussion

HIV persists for decades on fully suppressive ART, primarily as poorly expressed proviruses in infected cells that are long-lived, and continue to divide. Many (perhaps all) of these long-lived cells are in expanded clones. The present study was undertaken to better understand the factors that affect which provirus-containing cells survive in vivo. We compared a large IS dataset obtained from PBMC infected in vitro to pre-ART and on-ART datasets compiled from a number of in-house studies of HIV-infected individuals. The approach we used [21, 30, 31] provided IS data that are unbiased by the distribution of restriction enzyme sites in the human genome. The data were extensively screened to remove cross-contamination and any PCR artifacts or other errors.

There are three factors that modify the initial distribution of proviruses in vivo: 1) negative selection against cells with proviruses in highly-expressed genes; 2) positive selection for cells with proviruses in specific regions of certain genes; and 3) stochastic effects, often due to immune signaling [43], that lead to clonal amplification of some infected cells. The primary determinant of the pre- and on-ART distribution of IS is their initial distribution at the time the cells were infected. In the experiments we report here, that distribution was modeled using the distribution of IS in donor PBMC infected in vitro. In general, the three IS datasets are very similar. The excellent agreement, for most genes, in the distribution of the IS in the PBMC data and the two in vivo datasets supports our use of the PBMC IS dataset to accurately reflect the initial distribution of IS in vivo. The greater similarity between the distribution of IS in PBMC and the distribution of IS in the pre-ART samples compared to the on-ART samples is consistent with the pre-ART dataset comprising a mixture of newly infected, short-lived cells and cells that have been infected for longer times. Differences between the PBMC and on-ART IS datasets can provide insights into the selective forces that reshape the initial distribution of IS in vivo.

As has been known for many years, HIV DNA integration strongly favors the bodies of transcribed genes and the orientation of the provirus is independent of the chromosomal orientation of the host gene [7–9]. The strong association of HIV integration with expressed genes is shown in Figs 1 and 2 and Table 1, where the normalized integration density (IS/Mb) in expressed genes was more than 15-fold greater than in the rest of the genome. Integration density in PBMC was 8-fold greater in genes with high expression levels than in those expressed at low levels.

Many highly-expressed genes are good targets for HIV integration; however, high levels of gene expression do not always predict high levels of integration. This effect may be mediated by the distribution of LEDGF on chromatin [9, 44–46]. For example, in S2 Fig, which shows three adjacent *STAT* genes, all of which produce similar levels of RNA in PBMC, *STAT5B* is a good target for HIV integration, *STAT3*, less so, and *STAT5A* has very few IS. These discrepancies may reflect the presence, in the CD8-depleted PBMCs used for the RNA-seq analysis, of differentially expressed genes in cells that are and are not targets for HIV infection, or genes that express very stable transcripts, so that the steady state RNA levels measured by RNA-seq may not be indicative of the level of transcription, and there could be other explanations as well.

Almost all regions of the host genome that are not expressed are poor targets for HIV integration. This result was consistently seen both in vitro and in vivo. An unbiased search for clustered IS turned up one apparent intergenic region, on chromosome 6, in which there was a cluster of IS but no annotated gene. However, the RNA-seq data show that this region is transcribed, although not yet annotated as a gene encoding a protein or a known long noncoding RNA. This result suggests that clusters of HIV IS might be a useful way to check genome databases for genes that have not been annotated. Very rarely, we found annotated genes that were good targets for integration both in vivo and in vitro, in the absence of detectable levels of RNA in PBMC. *HORMAD2*, whose protein product is involved in meiosis, is a good example (S4 Fig), in that it had the largest number of proviral ISs of any non-expressed gene in both the in vitro-infected PBMC and on-ART datasets. Interestingly, but perhaps coincidentally, a provirus in this gene (in the form of a solo LTR) is in one of the largest clones we have seen in any on-ART donor [21, 47].

Although the distributions of IS in genes in the PBMC, pre-ART, and on-ART datasets are very similar, some clear differences emerged when the frequency of the IS was plotted against the level of gene expression (Fig 1). The normalized fractions of IS in the three datasets were identical for genes that were poorly expressed; however, the fraction of the IS in highly-expressed genes was higher in the in vitro PBMC dataset than in either the pre-ART or the on-ART datasets. The in vivo datasets also showed that there is selection against proviruses in either orientation in highly-expressed genes. Although an overall opposite-the-gene orientation bias has been previously reported for on-ART samples [21, 29], the fact that the magnitude of the selective effect depends on the level of expression of the genes had not been reported. Nor was it previously known that there is a weaker selection against proviruses in highly-expressed genes that are oriented opposite to the gene.

It has been proposed that latent, intact, proviruses within genes may be more likely to be expressed than those in extragenic regions, leading to death of the cell and accounting for a bias for defectiveness for proviruses in genes [24], although the net effect is not large and counterexamples have been reported [48]. Although the vast majority of proviruses are defective, that type of selection could also apply to cells with defective proviruses if the defective proviruses expressed epitopes that are recognized by CTL, for which there is conflicting evidence [49–51]. Given the recent results showing that only a small fraction of cells, containing either intact or defective proviruses, express unspliced viral RNA [25], we think it more likely that the selection against cells with proviruses in highly-expressed genes is due to the effects of the provirus on the expression of the host gene. In the case of proviruses in the same orientation as the gene, these effects would involve the insertion of viral sequences containing signals for transcription initiation, splicing, polyadenylation, etc., interfering with its expression. Certain endogenous proviruses of mice are also known to affect the expression of the host gene in which they reside in both ways [52, 53]. The effect on gene expression is limited to the allele in which the provirus is inserted, possibly explaining why the strongest selection is seen for proviruses that are integrated in the most highly-expressed genes. Proviruses integrated in the opposite orientation to a gene could also reduce its expression by transcriptional interference; i.e., collisions between RNA polymerases transcribing in opposite directions [54]. However, this effect on gene expression requires LTR-driven transcription and is therefore likely to be weaker for (largely nonexpressed) proviruses in the opposite orientation to the gene. In support of the interpretation that the effect is on the level of expression of the host gene, the effect is due to a weak selection on a large number of genes, rather than strong selection in a small number of genes. The observation that there is a much weaker relationship between gene expression and provirus loss in the pre-ART than in the on-ART samples (compare Fig 1, panels C and D) also supports this conclusion.

It has been known for nearly 40 years that modification of gene structure and expression by retroviral DNA integration can cause cancer in animals [55, 56]. Proviruses are known to induce oncogenic modification of genes by the insertion of a promoter or an enhancer, which leads to the overexpression of a host oncogene and/or a truncation of the encoded protein [57]. Similar mechanisms likely affect the expression of the three genes (*BACH2*, *MKL2*,and *STAT5B*) previously identified as targets for HIV provirus mediated preferential expansion or survival of infected T cells in humans [21, 23, 28, 36, 58]. In the present study, we did a comprehensive search for additional genes in which an HIV provirus insertion could cause similar posititve selection for infected T cells. Host genes in which a provirus could help the cell proliferate or survive were identified by comparing the IS frequencies in individual genes in the on-ART dataset and the PBMC dataset. Genes were identified based on finding a large increase in the integration frequency in vivo, as well as selection for proviruses in the same orientation as the gene, and clustering of the proviruses in one or a few introns. A complete list of genes with at least one IS in any of the datasets, sorted by thes criteria can be found in S2 Data.

Using the same criteria, we found three additional genes in which proviruses contributed to the growth and survival of the cell: *MKL1*, *IL2RB*, and *MYB*. A fourth gene, *POU2F1*, an OCT-1 binding transcrition factor, may also belong to this group. Because we began with a large on-ART dataset, this set of seven likely represents a complete, or nearly complete, set of the host genes which are good targets for HIV DNA integration (>10 IS in our PBMC dataset), and in which enhancement of growth and/or survival of the host cell by an integrated HIV provirus is likely to play an important role in HIV persistence. As described in Results, there were only 10 (out of 33,336) unique ISs in *MYB* in the on-ART dataset, all in the same orientation as the gene. While it is possible that there are other poorly-targeted genes in which HIV proviral integration would have similar effects, any gene that has not yet been identified must be a very poor target for HIV integration. Of the 20,207 genes in our modified database, only about half (11,422) had one or more IS in the PBMC dataset, and only about one quarter (5,159) had more than 10. Any gene we might have missed would have had even fewer activating proviruses than *MYB* and *POU2F1*. If either *MYB* or *POU2F1* was excluded from the group, there would be only a small decrease in the number of cells in which a provirus was selected (10 out of the 398–2.5%) proviruses integrated in the seven genes that are oriented in the same direction as the gene). Similarly, adding a gene to the list that had fewer than 10 proviruses would make only a very minor difference in the numbers, and would not have affected the conclusions.

Proviruses integrated in one of these seven genes comprised, together, approximately 1.7% of the total unique proviruses. Many of the proviruses in these genes are defective (Hill, S. and Maldarelli, F, unpublished observations); there are no published reports of an intact provirus in any of these genes. If clonally expanded cells are counted, the cells with proviruses in the seven genes comprise about 2.2% of the total infected cells. Thus, it is unlikely that the cells that have expanded because they have a provirus that drives their proliferation and/or survival are an important part of the viral reservoir, which is, by definition, composed of intact proviruses.

We and others have previously pointed out that a large fraction of the HIV proviruses in on-ART samples are integrated in genes/oncogenes that play important roles in the proliferation and survival of cells [21, 23, 59]. However, as the data we present here clearly show, this preference is a result of the propensity of HIV proviruses to integrate in expressed genes, and is not due to post-integration selection for cells that have proviruses integrated in oncogenes other than the ones that have already been discussed. For all other oncogenes, the fraction of the IS in the gene does not increase when the PBMC data and the on-ART data are compared, showing the absence of strong post-integration selection in vivo for proviruses in other oncogenes. It is noteworthy that selection for integration in STAT3 was not observed (S2 Fig), although such integration has been found to promote growth of T cells in vitro [60].

As previously mentioned, proviruses of oncogenic retroviruses can modify the expression of host genes [56, 57], and it is likely that HIV proviruses can use similar mechanisms to alter host gene expression. In *BACH2*, *STAT5B*, and *IL2RB*, the selected proviruses are in introns upstream of the start site for translation, and probably provide a new promoter for the gene, and a new upstream leader sequence that is derived from the provirus [28]. Proviruses integrated in *MKL1* and *MKL2* interrupt the protein coding region, which, if the mechanism of activation is promoter insertion, would lead to the expression of truncated proteins that lack amino terminal domains present in the normal proteins. In the case of *MYB*, the proviruses are in a large 3' intron and are likely to cause premature polyadenylation, resulting in a protein product that lacks the correct 40 C-terminal amino acids. The inserted proviruses would also lead to the synthesis of an mRNA lacking the long 3' noncoding region of *MYB* mRNA, which has been reported to contain signals that promote rapid RNA degradation [61]. These are all mechanisms that oncogenic retroviruses use to modify expression of host genes [56].

The cells in which an HIV provirus has integrated into one of the seven oncogenes have a growth and/or survival advantage, but they are not cancer cells. However, as with cancer, proviral induced misexpression of the host target gene product may promote cell division that would override the normal growth controls. Alternatively, the altered protein may provide a survival advantage, by, for example, suppressing apoptotic signaling, which is a normal part of the T cell immune response. Of the seven genes in which proviruses led to positive selection, only three were in cells that were clonally amplified to a greater extent than average for all cells in the on-ART samples (Fig 5C). It is possible that this difference reflects two different types of positive selection: Cells in which the provirus is in *STAT5B*, *MKL2*, and *MYB* may have increased potential for growth and cells with proviruses in *BACH2*, *MKL1*, *IL2RB*, *and POU2F1* may have better long-term survival due to a reduction in apoptotic signaling.

Clonal amplification is the most striking feature of persistent proviruses on-ART and is responsible for the majority–perhaps all–of the infected cells that persist on long-term ART, whether they contain intact or defective proviruses. We show here that provirus-driven clonal expansion accounts only for a small minority of the clonally expanded cells, which means that forces independent of the location and structure of the provirus cause the majority of the clonal expansion that characterizes infected cells in individuals on ART. Thus, in most cases, the proviruses only provide useful markers for the cells in expanded clones. In support of this conclusion, we found that the extent of clonal expansion is independent of the location of the provirus, whether it is in a gene or intergenic, the expression level of the host gene, and the orientation of the provirus relative to the gene (Fig 5). Even when integration is in centromere-associated alphoid repeat DNA, which is believed to be highly inimical to proviral expression [42, 62], the extent of amplification was not different from what was found for all of the proviruses taken together. We therefore conclude that proviral-driven expression of "cancer genes" plays only a minor role in the clonal expansion of infected cells, and that other mechanisms, probably antigen-driven or homeostatic expansion, account for the proliferation and persistence of the large majority of clones of infected cells in individuals on ART, especially those with intact proviruses.

The differences in the distributions of the proviruses in the PBMC infected in vitro and in the pre- and on-ART samples, although highly significant, are small (Fig 2). This result implies that the location of the proviruses in the host genome has only a small effect on the survival of an infected cell. Because cells that carry expressed proviruses are preferentially lost on ART, we conclude that the location of a provirus in the host genome has little effect on the expression of the provirus in long-lived cells. We show that there is a modest loss of cells with proviruses in highly expressed genes, and the loss is greater for the provirus in the same transcriptional orientation as the gene. However, as we have already discussed, this selection is more likely to be due to the effects of the proviruses on the expression of the host genes than the other way around. Turning the problem around, if the cells that carry proviruses that are not expressed have a survival advantage, and if the location of the proviruses in the genome was a substantial factor in whether the proviruses were expressed, we would have expected to see a larger shift in the distribution of the IS on ART. Although there is, with time on ART, a very slow and erratic selective loss of relatively intact or inducible proviruses [49, 63, 64] which leads to an enrichment in the fraction of defective proviruses [47], there is no evidence that the location of the provirus plays a role in this selection. Our conclusion that the location of an HIV provirus in the genome has little, if any, effect on its expression in infected individuals is supported by data, obtained from cells that were infected with HIV-based vectors in vitro, that the distributions, in the host genome, of the vector proviruses that are and are not expressed are quite similar [62].

There are limitations on the interpretations that can be made based on the data we have presented. For one, the analysis of the behavior of HIV infected cells and the proviruses they carry are limited by the depth of sampling that can be performed. A 100 ml blood sample contains, on average, about $10^8$ CD4+ target cells, about $10^{-4}$ of all the CD4+ T cells present in the body. Assuming that the cells are uniformly distributed, and approximately 1000 IS are obtained from such a sample, we can expect to recognize a clone of infected cells if it contains $>\sim10^5$ cells [14]. A clone of this size would be at least 16 generations removed from its originally infected progenitor cell. Although it is possible, by obtaining more IS, to identify clones smaller than $10^5$ cells, as a practical matter it is not possible to identify clones much smaller than $10^4$ cells. Although it is likely that the large majority–if not all–of the infected cells in patients on long-term ART are in clones of cells that have divided multiple times since they were initially infected [65], our hypothesis cannot be demonstrated experimentally.

Second, our assays monitor only IS, not the structure of the appended proviruses, putting a limitation on our ability to ask whether the results we have presented would differ if we had separately analyzed only the intact, potentially infectious, proviruses. It is now possible, to a limited extent, to determine both the integration site and the structure of the linked provirus [24, 26]. There are reports of differences in the behavior of cells with intact and defective proviruses during long-term ART in the form of a preferential loss of cells that carry intact proviruses, and that intact proviruses are more commonly outside of genes [24, 66–68]. Testing of these proposals will require analysis of much larger numbers of proviruses and their linked IS than is possible with current technology.

In conclusion, the distribution of HIV proviruses, even after years of complete virological suppression, in which the majority of the infected surviving cells are many generations removed from the originally infected cells, still closely resembles the distribution of the proviruses at the time the cells were initially infected, represented by the proviral distribution in PBMC acutely infected in vitro. In addition to a modest modification to the initial distribution by a positive selection for proviruses integrated in the seven genes we have listed, the distribution is also modified by selection against cells with proviruses integrated in highly-expressed genes. There is also extensive clonal amplification; however, more than 98% of the amplification results from factors in which the proviruses and their locations in the genome play no apparent role in the behavior, growth, or survival of the infected cells.

## Methods

### Ethics statement

PBMC were isolated from blood drawn from HIV-infected donors participating in one of a number of clinical studies under the aegis of the HIV DRP, the University of Pittsburgh, and Stellenbosch University as described in the references to S1 Table. Both donors provided written informed consent and the donation protocol was approved by the University of Pittsburgh Institutional Review Board.

**Donors and cells.** PBMC were isolated from 120 mL of blood from two anonymous healthy, HIV negative donors and depleted of CD8+ T cells using Dynabeads (ThermoFisher). Approximately 30 x $10^6$ cells were placed in 30ml of R10 medium, were treated with or without 0.5ug/mL of phytohemagglutinin (PHA) [69], and were incubated at 37˚C for 2 days.

Aliquots of $10^7$ cells, were pelleted and resuspended in 2mL of fresh RPMI-1640 medium, with or without PHA, supplemented with 10% FBS and containing $10^7$ IU/ml HIV-$1_{BAL}$ [70] for a multiplicity of infection of 1 infectious unit/cell as assayed in GHOST cells [71]. After 2 hours, the cells were washed and incubated at 37˚C for a further 2 days. $10^7$ cells from each flask were pelleted, resuspended in 1ml fresh cryopreservation medium, and stored at -80˚C. Except for a ca. 8-fold decline in the number of IS obtained, there was very little difference in expression of the genes or in the distribution of the IS in the PHA stimulated and unstimulated cells; (see S6 Fig).

**Integration site analysis.** The distribution of HIV IS in all PBMC samples was determined by linker-mediated nested PCR of fragmented DNA using LTR and linker-specific primers and paired-end Illumina sequencing as described [21]. Raw data were analyzed using the bioinformatic pipeline described by Wells et al [31], which includes steps designed to remove artifacts due to mispriming, cross contamination, etc. such as those found in Cohn et al. [31, 35]. When we compared the gene-by-gene distribution of the PBMC IS from the two donors (using the Excel CORREL function), the correlation coefficient (r) was 0.98, and, for all analyses, the two datasets were combined, giving 384,295 total IS. All IS data have been posted on the Retroviral Integration Database (RID) at https://rid.ncifcrf.gov/ [72], and can be found using the PubMed ID of this paper. All data are also present in the Microsoft Excel document that can be found in the "ISA" sheet of S1 Data, along with all the IS and RNAseq data used in this study, The PBMC dataset has also been used as a comparator for a study on SIV IS distribution in the macaque model [29].

The IS analysis pipeline used for this study reports the site (at the end of the 3' LTR) in the hg19 human sequence database, the orientation of each provirus relative to the host chromosome, and the number of differently sheared flanking host DNA sequences (or breakpoints) associated with each distinct IS. The breakpoint count indicates the relative number of clonally-expanded cells containing the same provirus. For most analyses, the clonally expanded IS data were collapsed so that only a single unique IS was used. Since our analysis involves simultaneous amplification of IS using primers for both 3' and 5' LTRs, it is possible that we are slightly overcounting the number of clonally amplified proviruses. This overcount would not affect any of our conclusions, since most analyses (Figs 1–5) used collapsed data and, in the analysis shown in Fig 6, it would have affected all counts equally, leaving the conclusion intact.

Although most current human genome data are now reported in the hg38 format, most of the integration site data, including the published patient data used in this study, were collected from many early studies using hg19. To be consistent, we chose hg19 for all the integration site analysis and hence the RNAseq data also in hg19. We are in the process of transferring the data in S1 Data to hg38, and will make it available when we do so.

To map IS to centromeric regions, fastq files generated from the Illumina platform were processed and analyzed using our in-house standard pipelines for HIV integration sites analysis, with specific modifications to accommodate centromere mapping. Reads properly trimmed and filtered were subjected to centromere alignment using BLAT without -ooc filter to GRChg38 (released Dec. 2013) centromere assembly obtained from the University of California Santa Cruz genome browser (https://genome.ucsc.edu) as a reference. Reads mapped to centromeres were further merged and filtered to identify real integration sites. Manual inspections were used to finalize the mapping results.

Gene locations were obtained from the RefSeq set [34]. For counting purposes, in cases of overlapping genes, the 5'-most gene was truncated, and genes entirely inside other genes were removed so that all sequences in host genes were assigned to a single gene. This correction was necessary to avoid double counting of IS where there are overlapping genes.

The maps shown in Figs 3–5 and S3–S4, were generated using a custom-made Microsoft Excel workbook, as described in S2 Fig, which also contains all the data used in this paper and allows the user to select up to 5 datasets at a time, and choose a specific gene, chromosome, or chromosomal region to be viewed. The program and all the IS and RNAseq data used herein are available in S1 Data.

P values for most comparisons were calculated using a binomial or Poisson distribution or Fisher's exact test, as indicated.

**RNA-seq analysis.** Total nucleic acid was extracted from one million cells each of the CD8-depleted PBMC infected with HIV from both donors with or without PHA stimulation. RNAseq was carried out using an Illumina TruSeq mRNA library kit with 200ng of poly(A) selected RNA as input. Libraries were pooled and sequenced on Illumina NextSeq500 with TruSeq V2 chemistry paired end reads x 75bp, yielding ~100 million reads per sample. Sequencing was performed at NCI CCR Sequencing Core Facility. RNAseq data analysis was performed using the standard RNAseq workflow in CLC Genomics Workbench V12. Reads were mapped to human genome hg19 and Ensembl genes annotation v74 was used to calculate gene expression values as transcripts per million reads (TPM). Values obtained from the two donors were averaged. Gene expression levels in the PHA-stimulated, infected PBMC ranged from 0.5 to 12725 TPM. Genes with < 0.5 TPM are referred to as "unexpressed". Only protein coding genes and lncRNA genes were included for combined gene expression and IS analysis.

## Supporting information

**S1 Table. Sources of Donor IS Data.**
(PDF)

**S2 Table. Distribution of IS as a function of gene expression (TPM)[a].**
(PDF)

**S3 Table. Genes in which proviruses oriented opposite to the host gene were strongly selected in the on-ART dataset.**
(PDF)

**S4 Table. Genes with the largest clusters in 10 kb windows in on-ART samples.**
(PDF)

**S1 Fig. Gene expression and integration.**
(PDF)

**S2 Fig. Application for visualization of IS and expression data.**
(PDF)

**S3 Fig. Genes with high frequency of proviruses integrated in the opposite transcriptional orientation.**
(PDF)

**S4 Fig. Clusters of IS not associated with cell growth or survival.**
(PDF)

**S5 Fig. Integration into a non-expressed gene.**
(PDF)

**S6 Fig. IS Ranked by Clonal Amplification.**
(PDF)

**S7 Fig. Effect of PHA stimulation on transcription and integration.**
(PDF)

**S1 Data. Integration-Transcription Maps Data and Worksheet.**
(XLSB)

**S2 Data. Complete list of IS in genes in the PBMC and on-ART datasets.**
(XLSX)

## Acknowledgments

The authors would like to thank Teresa Burdette for help in preparing the manuscript, Allen Kane for help with the Figs, and Anne Arthur for editorial help. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services or the National Institutes of Health, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. government.

## Author Contributions

**Conceptualization:** John M. Coffin, John W. Mellors, Stephen H. Hughes.

**Data curation:** Daria Wells, Xiaolin Wu.

**Formal analysis:** John M. Coffin, Michael J. Bale, Brian Luke, Twan Sia, Wei Shao, Frank Maldarelli, Mary F. Kearney.

**Funding acquisition:** John M. Coffin, Stephen H. Hughes.

**Investigation:** Daria Wells, Shuang Guo, Jennifer M. Zerbato, Michele D. Sobolewski, Frank Maldarelli.

**Methodology:** John M. Coffin, Michael J. Bale, Xiaolin Wu.

**Resources:** Frank Maldarelli, John W. Mellors, Stephen H. Hughes.

**Software:** John M. Coffin, Michael J. Bale, Xiaolin Wu.

**Supervision:** Xiaolin Wu, Frank Maldarelli, Mary F. Kearney, John W. Mellors, Stephen H. Hughes.

**Visualization:** John M. Coffin, Michael J. Bale, Stephen H. Hughes.

**Writing – original draft:** John M. Coffin, Stephen H. Hughes.

**Writing – review & editing:** Michael J. Bale, Jennifer M. Zerbato, Frank Maldarelli, Mary F. Kearney, John W. Mellors.

## References

1. Coffin JM. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. Science. 1995; 267(5197):483–9. https://doi.org/10.1126/science.7824947 PMID: 7824947.
2. Ho DD, Neumann AU, Perelson AS, Chen W, Leonard JM, Markowitz M. Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. Nature. 1995; 373(6510):123–6. Epub 1995/01/12. https://doi.org/10.1038/373123a0 PMID: 7816094.
3. Wei X, Ghosh S, Taylor ME, Johnson VA, Emini EA, Deutsch P, et al. Viral dynamics in human immunodeficiency virus type 1 infection. Nature. 1995; 373:117–22. https://doi.org/10.1038/373117a0 PMID: 7529365
4. Joos B, Fischer M, Kuster H, Pillai SK, Wong JK, Boni J, et al. HIV rebounds from latently infected cells, rather than from continuing low-level replication. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105(43):16725–30. https://doi.org/10.1073/pnas.0804192105 PMID: 18936487; PubMed Central PMCID: PMC2575487.
5. Cohn LB, Chomont N, Deeks SG. The Biology of the HIV-1 Latent Reservoir and Implications for Cure Strategies. Cell Host Microbe. 2020; 27(4):519–30. https://doi.org/10.1016/j.chom.2020.03.014 PMID: 32272077.
6. Hughes SH, Coffin JM. What Integration Sites Tell Us about HIV Persistence. Cell Host Microbe. 2016; 19(5):588–98. https://doi.org/10.1016/j.chom.2016.04.010 PMID: 27173927; PubMed Central PMCID: PMC4900157.
7. Craigie R, Bushman FD. HIV DNA integration. Cold Spring Harbor perspectives in medicine. 2012; 2 (7):a006890. https://doi.org/10.1101/cshperspect.a006890 PMID: 22762018; PubMed Central PMCID: PMC3385939.
8. Schröder AR, Shinn P, Chen H, Berry C, Ecker JR, Bushman F. HIV-1 integration in the human genome favors active genes and local hotspots. Cell. 2002; 110(4):521–9. https://doi.org/10.1016/s0092-8674(02)00864-4 PMID: 12202041.
9. Singh PK, Plumb MR, Ferris AL, Iben JR, Wu X, Fadel HJ, et a. LEDGF/p75 interacts with mRNA splicing factors and targets HIV-1 integration to highly spliced genes. Genes Dev. 2015; 29(21):2287–97. https://doi.org/10.1101/gad.267609.115 PMID: 26545813; PubMed Central PMCID: PMC4647561.
10. Holman AG, Coffin JM. Symmetrical base preferences surrounding HIV-1, avian sarcoma/leukosis virus, and murine leukemia virus integration sites. Proc Natl Acad Sci U S A. 2005; 102(17):6103–7. https://doi.org/10.1073/pnas.0501646102 PMID: 15802467; PubMed Central PMCID: PMC1087937.

11. Wu X, Burgess SM. Integration target site selection for retroviruses and transposable elements. Cell Mol Life Sci. 2004; 61(19–20):2588–96. https://doi.org/10.1007/s00018-004-4206-9 PMID: 15526164.

12. Kirk PD, Huvet M, Melamed A, Maertens GN, Bangham CR. Retroviruses integrate into a shared, non-palindromic DNA motif. Nat Microbiol. 2016; 2:16212. https://doi.org/10.1038/nmicrobiol.2016.212 PMID: 27841853.

13. Serrao E, Krishnan L, Shun MC, Li X, Cherepanov P, Engelman A, et al. Integrase residues that determine nucleotide preferences at sites of HIV-1 integration: implications for the mechanism of target DNA binding. Nucleic Acids Res. 2014; 42(8):5164–76. https://doi.org/10.1093/nar/gku136 PMID: 24520116; PubMed Central PMCID: PMC4005685.

14. Coffin JM, Wells DW, Zerbato JM, Kuruc JD, Guo S, Luke BT, et al. Clones of infected cells arise early in HIV-infected individuals. JCI Insight. 2019; 4(12). Epub 2019/06/21. https://doi.org/10.1172/jci.insight.128432 PMID: 31217357; PubMed Central PMCID: PMC6629130.

15. Bruner KM, Murray AJ, Pollack RA, Soliman MG, Laskey SB, Capoferri AA, et al. Defective proviruses rapidly accumulate during acute HIV-1 infection. Nat Med. 2016. https://doi.org/10.1038/nm.4156 PMID: 27500724.

16. Ho YC, Shan L, Hosmane NN, Wang J, Laskey SB, Rosenbloom DI, et al. Replication-competent noninduced proviruses in the latent reservoir increase barrier to HIV-1 cure. Cell. 2013; 155(3):540–51. https://doi.org/10.1016/j.cell.2013.09.020 PMID: 24243014; PubMed Central PMCID: PMC3896327.

17. De Scheerder MA, Vrancken B, Dellicour S, Schlub T, Lee E, Shao W, et al. HIV Rebound Is Predominantly Fueled by Genetically Identical Viral Expansions from Diverse Reservoirs. Cell Host Microbe. 2019; 26(3):347–58 e7. https://doi.org/10.1016/j.chom.2019.08.003 PMID: 31471273.

18. Palmer S, Maldarelli F, Wiegand A, Bernstein B, Hanna GJ, Brun SC, et al. Low-level viremia persists for at least 7 years in patients on suppressive antiretroviral therapy. Proc Natl Acad Sci U S A. 2008; 105(10):3879–84. Epub 2008/03/12. https://doi.org/10.1073/pnas.0800050105 PMID: 18332425; PubMed Central PMCID: PMC2268833.

19. Kearney MF, Spindler J, Shao W, Yu S, Anderson EM, O'Shea A, et al. Lack of Detectable HIV-1 Molecular Evolution during Suppressive Antiretroviral Therapy. PLoS pathogens. 2014; 10(3):e1004010. https://doi.org/10.1371/journal.ppat.1004010 PMID: 24651464; PubMed Central PMCID: PMC3961343.

20. McManus WR, Bale MJ, Spindler J, Wiegand A, Musick A, Patro SC, et al. HIV-1 in lymph nodes is maintained by cellular proliferation during antiretroviral therapy. The Journal of clinical investigation. 2019; 130:4629–42. https://doi.org/10.1172/JCI126714 PMID: 31361603; PubMed Central PMCID: PMC6819093.

21. Maldarelli F, Wu X, Su L, Simonetti FR, Shao W, Hill S, et al. HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. Science. 2014; 345(6193):179–83. https://doi.org/10.1126/science.1254194 PMID: 24968937; PubMed Central PMCID: PMC4262401.

22. Simonetti FR, Sobolewski MD, Fyne E, Shao W, Spindler J, Hattori J, et al. Clonally expanded CD4+ T cells can produce infectious HIV-1 in vivo. Proceedings of the National Academy of Sciences of the United States of America. 2016; 113(7):1883–8. https://doi.org/10.1073/pnas.1522675113 PMID: 26858442; PubMed Central PMCID: PMC4763755.

23. Wagner TA, McLaughlin S, Garg K, Cheung CY, Larsen BB, Styrchak S, et al. HIV latency. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. Science. 2014; 345 (6196):570–3. https://doi.org/10.1126/science.1256304 PMID: 25011556; PubMed Central PMCID: PMC4230336.

24. Einkauf KB, Lee GQ, Gao C, Sharaf R, Sun X, Hua S, et al. Intact HIV-1 proviruses accumulate at distinct chromosomal positions during prolonged antiretroviral therapy. The Journal of clinical investigation. 2019; 129(3):988–98. https://doi.org/10.1172/JCI124291 PMID: 30688658; PubMed Central PMCID: PMC6391088.

25. Musick A, Spindler J, Boritz E, Perez L, Crespo-Velez D, Patro SC, et al. HIV Infected T Cells Can Proliferate in vivo Without Inducing Expression of the Integrated Provirus. Front Microbiol. 2019; 10:2204. Epub 2019/10/22. https://doi.org/10.3389/fmicb.2019.02204 PMID: 31632364; PubMed Central PMCID: PMC6781911.

26. Patro SC, Brandt LD, Bale MJ, Halvas EK, Joseph KW, Shao W, et al. Combined HIV-1 sequence and integration site analysis informs viral dynamics and allows reconstruction of replicating viral ancestors. Proceedings of the National Academy of Sciences of the United States of America. 2019; 116 (51):25891–9. https://doi.org/10.1073/pnas.1910334116 PMID: 31776247; PubMed Central PMCID: PMC6925994.

27. Halvas EK, Joseph KW, Brandt JLD, Guo S, Sobolewski MD, Jacobs JL, et al. HIV-1 Viremia Not Suppressible By Antiretroviral Therapy Can Originate from Large T-Cell Clones Producing Infectious Virus J Clin Inv. 2020;In Press.

28. Cesana D, Santoni de Sio FR, Rudilosso L, Gallina P, Calabria A, Beretta S, et al. HIV-1-mediated insertional activation of STAT5B and BACH2 trigger viral reservoir in T regulatory cells. Nat Commun. 2017; 8(1):498. https://doi.org/10.1038/s41467-017-00609-1 PMID: 28887441; PubMed Central PMCID: PMC5591266.

29. Ferris AL, Wells DW, Guo S, Del Prete GQ, Swanstrom AE, Coffin JM, et al. Clonal expansion of SIV-infected cells in macaques on antiretroviral therapy is similar to that of HIV-infected cells in humans. PLoS pathogens. 2019; 15(7):e1007869. https://doi.org/10.1371/journal.ppat.1007869 PMID: 31291371; PubMed Central PMCID: PMC6619828.

30. Berry CC, Gillet NA, Melamed A, Gormley N, Bangham CR, Bushman FD. Estimating abundances of retroviral insertion sites from DNA fragment length data. Bioinformatics. 2012; 28(6):755–62. https://doi.org/10.1093/bioinformatics/bts004 PMID: 22238265; PubMed Central PMCID: PMC3307109.

31. Wells DW, Guo S, Shao W, Bale MJ, Coffin JM, Hughes SH, et al. An analytical pipeline for identifying and mapping the integration sites of HIV and other retroviruses. BMC Genomics. 2020; 21(1):216. Epub 2020/03/11. https://doi.org/10.1186/s12864-020-6647-4 PMID: 32151239; PubMed Central PMCID: PMC7063773.

32. Van Zyl GU, Katusiime MG, Wiegand A, McManus WR, Bale MJ, Halvas EK, et al. No evidence of HIV replication in children on antiretroviral therapy. J Clin Invest. 2017; 127(10):3827–34. Epub 2017/09/12. https://doi.org/10.1172/JCI94582 PMID: 28891813; PubMed Central PMCID: PMC5617669.

33. Lewinski MK, Yamashita M, Emerman M, Ciuffi A, Marshall H, Crawford G, et al. Retroviral DNA integration: viral and cellular determinants of target-site selection. PLoS pathogens. 2006; 2(6):e60. https://doi.org/10.1371/journal.ppat.0020060 PMID: 16789841; PubMed Central PMCID: PMC1480600.

34. Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research. 2005; 33(Database issue):D501–4. https://doi.org/10.1093/nar/gki025 PMID: 15608248; PubMed Central PMCID: PMC539979.

35. Cohn LB, Silva IT, Oliveira TY, Rosales RA, Parrish EH, Learn GH, et al. HIV-1 integration landscape during latent and active infection. Cell. 2015; 160(3):420–32. https://doi.org/10.1016/j.cell.2015.01.020 PMID: 25635456; PubMed Central PMCID: PMC4371550.

36. Ikeda T, Shibata J, Yoshimura K, Koito A, Matsushita S. Recurrent HIV-1 integration at the BACH2 locus in resting CD4+ T cell populations during effective highly active antiretroviral therapy. J Infect Dis. 2007; 195(5):716–25. Epub 2007/01/31. https://doi.org/10.1086/510915 PMID: 17262715.

37. Kalita K, Kuzniewska B, Kaczmarek L. MKLs: co-factors of serum response factor (SRF) in neuronal responses. Int J Biochem Cell Biol. 2012; 44(9):1444–7. https://doi.org/10.1016/j.biocel.2012.05.008 PMID: 22626970.

38. Leprince D, Saule S, de Taisne C, Gegonne A, Begue A, Righi M, et al. The human DNA locus related to the oncogene myb of avian myeloblastosis virus (AMV): molecular cloning and structural characterization. EMBO J. 1983; 2(7):1073–8. PMID: 6194989; PubMed Central PMCID: PMC555237.

39. Vazquez-Arreguin K, Tantin D. The Oct1 transcription factor and epithelial malignancies: Old protein learns new tricks. Biochim Biophys Acta. 2016; 1859(6):792–804. Epub 2016/02/16. https://doi.org/10.1016/j.bbagrm.2016.02.007 PMID: 26877236; PubMed Central PMCID: PMC4880489.

40. Pramod AB, Foster J, Carvelli L, Henry LK. SLC6 transporters: structure, function, regulation, disease association and therapeutics. Mol Aspects Med. 2013; 34(2–3):197–219. Epub 2013/03/20. https://doi.org/10.1016/j.mam.2012.07.002 PMID: 23506866; PubMed Central PMCID: PMC3602807.

41. Podgornaya OI, Ostromyshenskii DI, Enukashvily NI. Who Needs This Junk, or Genomic Dark Matter. Biochemistry (Mosc). 2018; 83(4):450–66. Epub 2018/04/09. https://doi.org/10.1134/S0006297918040156 PMID: 29626931.

42. Jiang C, Lian X, Gao C, Sun X, Einkauf KB, Chevalier JM, et al. Distinct viral reservoirs in individuals with spontaneous control of HIV-1. Nature. 2020; 585(7824):261–7. Epub 2020/08/28. https://doi.org/10.1038/s41586-020-2651-8 PMID: 32848246.

43. Simonetti FR, Zhang H, Soroosh GP, Duan J, Rhodehouse K, Hill AL, et al. Antigen-driven clonal selection shapes the persistence of HIV-1 infected CD4+ T cells in vivo. J Clin Invest. 2020. Epub 2020/12/11. https://doi.org/10.1172/JCI145254 PMID: 33301425.

44. Ciuffi A, Llano M, Poeschla E, Hoffmann C, Leipzig J, Shinn P, et al. A role for LEDGF/p75 in targeting HIV DNA integration. Nat Med. 2005; 11(12):1287–9. https://doi.org/10.1038/nm1329 PMID: 16311605.

45. Engelman AN, Singh PK. Cellular and molecular mechanisms of HIV-1 integration targeting. Cell Mol Life Sci. 2018; 75(14):2491–507. https://doi.org/10.1007/s00018-018-2772-5 PMID: 29417178; PubMed Central PMCID: PMC6004233.

46. Ferris AL, Wu X, Hughes CM, Stewart C, Smith SJ, Milne TA, et al. Lens epithelium-derived growth factor fusion proteins redirect HIV-1 DNA integration. Proceedings of the National Academy of Sciences of

the United States of America. 2010; 107(7):3135–40. https://doi.org/10.1073/pnas.0914142107 PMID: 20133638; PubMed Central PMCID: PMC2840313.

47. Anderson EM, Simonetti FR, Gorelick RJ, Hill S, Gouzoulis MA, Bell J, et al. Dynamic Shifts in the HIV Proviral Landscape During Long Term Combination Antiretroviral Therapy: Implications for Persistence and Control of HIV Infections. Viruses. 2020; 12(2). Epub 2020/01/30. https://doi.org/10.3390/v12020136 PMID: 31991737; PubMed Central PMCID: PMC7077288.

48. Halvas EK, Joseph KW, Brandt LD, Guo S, Sobolewski MD, Jacobs JL, et al. HIV-1 viremia not suppressible by antiretroviral therapy can originate from large T cell clones producing infectious virus. J Clin Invest. 2020. Epub 2020/10/06. https://doi.org/10.1172/JCI138099 PMID: 33016926.

49. Antar AA, Jenike KM, Jang S, Rigau DN, Reeves DB, Hoh R, et al. Longitudinal study reveals HIV-1-infected CD4+ T cell dynamics during long-term antiretroviral therapy. J Clin Invest. 2020; 130(7):3543–59. Epub 2020/03/20. https://doi.org/10.1172/JCI135953 PMID: 32191639; PubMed Central PMCID: PMC7324206.

50. Imamichi H, Dewar RL, Adelsberger JW, Rehm CA, O'Doherty U, Paxinos EE, et al. Defective HIV-1 proviruses produce novel protein-coding RNA species in HIV-infected patients on combination antiretroviral therapy. Proc Natl Acad Sci U S A. 2016; 113(31):8783–8. https://doi.org/10.1073/pnas.1609057113 PMID: 27432972; PubMed Central PMCID: PMC4978246.

51. Pollack RA, Jones RB, Pertea M, Bruner KM, Martin AR, Thomas AS, et al. Defective HIV-1 Proviruses Are Expressed and Can Be Recognized by Cytotoxic T Lymphocytes, which Shape the Proviral Landscape. Cell Host Microbe. 2017; 21(4):494–506 e4. Epub 2017/04/14. https://doi.org/10.1016/j.chom.2017.03.008 PMID: 28407485; PubMed Central PMCID: PMC5433942.

52. Jenkins NA, Copeland NG, Taylor BA, Lee BK. Dilute (d) coat colour mutation of DBA/2J mice is associated with the site of integration of an ecotropic MuLV genome. Nature. 1981; 293(5831):370–4. https://doi.org/10.1038/293370a0 PMID: 6268990.

53. Stoye JP, Fenner S, Greenoak GE, Moran C, Coffin JM. Role of endogenous retroviruses as mutagens: the hairless mutation of mice. Cell. 1988; 54(3):383–91. https://doi.org/10.1016/0092-8674(88)90201-2 PMID: 2840205.

54. Cullen BR, Lomedico PT, Ju G. Transcriptional interference in avian retroviruses—implications for the promoter insertion model of leukaemogenesis. Nature. 1984; 307(5948):241–5. https://doi.org/10.1038/307241a0 PMID: 6363938.

55. Neel BG, Hayward WS, Robinson HL, Fang J, Astrin SM. Avian leukosis virus-induced tumors have common proviral integration sites and synthesize discrete new RNAs: oncogenesis by promoter insertion. Cell. 1981; 23(2):323–34. https://doi.org/10.1016/0092-8674(81)90128-8 PMID: 6258798.

56. Rosenberg N, Jolicoeur P. Retroviral Pathogenesis. In: Coffin JM, Hughes SH, Varmus HE, editors. Retroviruses. Cold Spring Harbor (NY)1997.

57. Jern P, Coffin JM. Effects of retroviruses on host genome function. Annu Rev Genet. 2008; 42:709–32. https://doi.org/10.1146/annurev.genet.42.110807.091501 PMID: 18694346.

58. Lange UC, Bialek JK, Walther T, Hauber J. Pinpointing recurrent proviral integration sites in new models for latent HIV-1 infection. Virus Res. 2018; 249:69–75. https://doi.org/10.1016/j.virusres.2018.03.007 PMID: 29550509.

59. Liu R, Yeh YJ, Varabyou A, Collora JA, Sherrill-Mix S, Talbot CC Jr., et al. Single-cell transcriptional landscapes reveal HIV-1-driven aberrant host gene transcription as a potential therapeutic target. Sci Transl Med. 2020; 12(543). Epub 2020/05/15. https://doi.org/10.1126/scitranslmed.aaz0802 PMID: 32404504; PubMed Central PMCID: PMC7453882.

60. Yoon JK, Holloway JR, Wells DW, Kaku M, Jetton D, Brown R, et al. HIV proviral DNA integration can drive T cell growth ex vivo. Proc Natl Acad Sci U S A. 2020; 117(52):32880–2. Epub 2020/12/16. https://doi.org/10.1073/pnas.2013194117 PMID: 33318172; PubMed Central PMCID: PMC7777207.

61. Cesi V, Casciati A, Sesti F, Tanno B, Calabretta B, Raschella G. TGFbeta-induced c-Myb affects the expression of EMT-associated genes and promotes invasion of ER+ breast cancer cells. Cell Cycle. 2011; 10(23):4149–61. https://doi.org/10.4161/cc.10.23.18346 PMID: 22101269.

62. Sherrill-Mix S, Lewinski MK, Famiglietti M, Bosque A, Malani N, Ocwieja KE, et al. HIV latency and integration site placement in five cell-based models. Retrovirology. 2013; 10:90. https://doi.org/10.1186/1742-4690-10-90 PMID: 23953889; PubMed Central PMCID: PMC3765678.

63. Gandhi RT, Cyktor JC, Bosch RJ, Mar H, Laird GM, Martin A, et al. Selective Decay of Intact HIV-1 Proviral DNA on Antiretroviral Therapy. J Infect Dis. 2020. Epub 2020/08/22. https://doi.org/10.1093/infdis/jiaa532 PMID: 32823274.

64. Siliciano JD, Kajdas J, Finzi D, Quinn TC, Chadwick K, Margolick JB, et al. Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4+ T cells. Nat Med. 2003; 9(6):727–8. Epub 2003/05/20. https://doi.org/10.1038/nm880 PMID: 12754504.

**65.** Reeves DB, Duke ER, Wagner TA, Palmer SE, Spivak AM, Schiffer JT. A majority of HIV persistence during antiretroviral therapy is due to infected cell proliferation. Nat Commun. 2018; 9(1):4811. https://doi.org/10.1038/s41467-018-06843-5 PMID: 30446650; PubMed Central PMCID: PMC6240116.

**66.** Anderson JA, Archin NM, Ince W, Parker D, Wiegand A, Coffin JM, et al. 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. Journal of virology. 2011; 85(10):5220–3. https://doi.org/10.1128/JVI.00284-11 PMID: 21367910; PubMed Central PMCID: PMC3126162.

**67.** Crooks AM, Bateson R, Cope AB, Dahl NP, Griggs MK, Kuruc JD, et al. Precise Quantitation of the Latent HIV-1 Reservoir: Implications for Eradication Strategies. J Infect Dis. 2015; 212(9):1361–5. https://doi.org/10.1093/infdis/jiv218 PMID: 25877550; PubMed Central PMCID: PMC4601910.

**68.** Peluso MJ, Bacchetti P, Ritter KD, Beg S, Lai J, Martin JN, et al. Differential decay of intact and defective proviral DNA in HIV-1-infected individuals on suppressive antiretroviral therapy. JCI Insight. 2020; 5 (4). https://doi.org/10.1172/jci.insight.132997 PMID: 32045386; PubMed Central PMCID: PMC7101154.

**69.** Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, Salazar MG, et al. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. Proc Natl Acad Sci U S A. 2008; 105(21):7552–7. https://doi.org/10.1073/pnas.0802203105 PMID: 18490657; PubMed Central PMCID: PMC2387184.

**70.** Gartner S, Markovits P, Markovitz DM, Kaplan MH, Gallo RC, Popovic M. The role of mononuclear phagocytes in HTLV-III/LAV infection. Science. 1986; 233(4760):215–9. Epub 1986/07/11. https://doi.org/10.1126/science.3014648 PMID: 3014648.

**71.** Morner A, Bjorndal A, Albert J, Kewalramani VN, Littman DR, Inoue R, et al. Primary human immunodeficiency virus type 2 (HIV-2) isolates, like HIV-1 isolates, frequently use CCR5 but show promiscuity in coreceptor usage. Journal of virology. 1999; 73(3):2343–9. https://doi.org/10.1128/JVI.73.3.2343-2349.1999 PMID: 9971817; PubMed Central PMCID: PMC104479.

**72.** Shao W, Shan J, Kearney MF, Wu X, Maldarelli F, Mellors JW, et al. Retrovirus Integration Database (RID): a public database for retroviral insertion sites into host genomes. Retrovirology. 2016; 13(1):47. https://doi.org/10.1186/s12977-016-0277-6 PMID: 27377064; PubMed Central PMCID: PMC4932684.