HIV infection results in clonal expansions containing integrations within pathogenesis-related biological pathways

Kevin G. Haworth[1], Lauren E. Schefter[1], Zachary K. Norgaard[1], Christina Ironside[1], Jennifer E. Adair[1, 2], Hans-Peter Kiem[1,2,3].

[1]Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA; [2]Department of Medicine, University of Washington, Seattle, Washington, USA. [3]Department of Pathology, University of Washington, Seattle, Washington, USA.

Short title: Expansion of HIV infected cells in unique pathways

Corresponding Author: Hans-Peter Kiem; Fred Hutchinson Cancer Research Center; 1100 Fairview Ave N, Mail Stop D1-100, PO Box 19024, Seattle, WA 98109-1024.

Phone: 206-667-4425; Fax: 206-667-6124; Email: hkiem@fredhutch.org.

ABSTRACT

The genomic integration of human immunodeficiency virus (HIV) into cells results in long-term persistence of virally infected cell populations. This integration event acts as a heritable mark that can be tracked to monitor infected cells which persist over time. Previous reports have documented clonal expansion in people and linked them to proto-oncogenes; however, their significance or contribution to the latent reservoir has remained unclear. Here we demonstrate that a directed pattern of clonal expansion occurs in vivo, and specifically in gene pathways important for viral replication and persistence. These biological processes include cellular division, transcriptional regulation, RNA processing, and post-translational modification pathways. This indicates preferential expansion when integration events occur within genes or biological pathways beneficial for HIV replication and persistence. Additionally, these expansions occur quickly during unsuppressed viral replication in vivo, reinforcing the importance of early intervention for individuals to limit reservoir seeding of clonally expanded HIV infected cells.

INTRODUCTION

Human immunodeficiency virus (HIV) is the causative agent of acquired immunodeficiency syndrome (AIDS) and was first characterized in humans over three decades ago (1, 2). It is estimated that over 35 million people have died from HIV infection with an additional 36.7 million people worldwide currently infected with the virus (Joint United Nations Programme on HIV/AIDS [UNAIDS] www.unaids.org). Potent drug regimens administered as combination antiretroviral therapy (cART) are able to suppress viral replication to undetectable levels in most individuals; however, viral rebound occurs rapidly following any disruption in treatment (3-5). These reactivations increase the risk of selecting for mutations in the viral population which are resistant to previously administered antiretroviral drugs (6, 7). Like all retroviruses, HIV reverse-transcribes its genome and permanently inserts itself into selected chromosomal locations within the infected cell, leading to long-term viral persistence (8, 9). Viral rebound occurs when an infected cell containing an integrated provirus reactivates transcription and emerges from its latent state (10). The process behind site selection of integration is still not entirely understood, but HIV is known to preferentially integrate into regions of open chromatin and active gene transcription (11). Once HIV has successfully integrated within a chromosomal locus, all subsequent cells arising through cellular division will contain the identical viral integration site (IS), functioning as a unique marker for each independent viral infection event. This results in a unique and heritable viral integration signature which can be sequenced and tracked over time using integration site analysis (ISA).

Previous reports have characterized IS within HIV infected individuals while they were suppressed on cART (12, 13). These studies established that clonal expansion of HIV infected cells contribute on some level to viral persistence (14). In agreement with historical data, they observed the majority of integrations occurred within gene transcripts (approximately 80% of IS), and about 12.5% of these genes have been associated with cancer development (13). It remains unclear what link, if any, there is between IS site selection of HIV and the potential for oncogenic

development. These reports suggested that the persistence and clonal expansion of infected cells may be influenced by the specific gene harboring the viral IS. Despite providing important initial insights into the viral integration landscape of individuals infected with HIV, several important questions remain. First, the driving force behind these clonal expansions and their potential clinical relevance remains to be determined. Additionally, it has also been reported that the majority of infected and clonally expanded cells contain a defective provirus, implying these expansions are not driven by viral integration (15). However, it is still possible that despite having a defective provirus, intact LTR regions could still alter local gene expression through recruitment of transcriptional factors (16). Therefore, even dead proviral elements might provide important insight into clonal expansion.

These observed clonal expansions raise some interesting possibilities. After HIV infection, the virus represses cellular division by using accessory genes to mediate several intracellular signaling pathways (17). As a result, clonal expansion of HIV-infected cells would only occur under certain circumstances: i) if the integrated provirus became defective, ii) the provirus silenced expression resulting in latency, iii) an infected cell underwent antigen stimulated expansion overwhelming viral anti-expansion signals, or iv) the integration itself initiated cellular proliferation due to insertional transformation. The two circumstances of highest clinical interest are situations in which the virus transitioned to a latent infection, and those occurring due to integration induced expansion. While HIV has not been shown to directly cause oncogenic transformation, the occurrence of lymphomas is higher in AIDS patients (18, 19). These AIDS-related lymphomas are thought to arise due to systemic immune dysregulation caused by HIV-mediated $CD4^+$ depletion (20). Another possibility for clonal expansion occurs when integrated proviral elements become dormant and transition into latently infected cells. It is thought that these cells contribute to viral rebound after cART withdrawal since they are refractory to treatment either due to low or no viral gene expression (21). It is still unclear exactly what causes an infected cell to become latent and subsequently where these cells reside in virally suppressed individuals (22).

One hypothesis is that latency is a randomly occurring event in a rare subset of infected cells due to changes in cellular transcription or chromatin status (23). Another possibility is the specific site of integration could drive cells to latency due to either local gene expression patterns or integration into a temporarily transcriptionally silent region of the genome.

In this manuscript, we use a preclinical HIV animal model to link, for the first time, integration site locations within specific gene families to their subsequent clonal expansion and persistence over time. We can then compare these IS identified using our in vivo model system to IS observed during in vitro infection conditions. While they may appear comparable in terms of total chromosomal distribution, there are significant differences in specific enrichment and expansion locations of integrations. We identify significantly expanded clones primarily during in vivo infection, occasionally in known proto-oncogenes, and also determine that cellular expansions are specifically driven by IS occurring within genes from biological pathways the virus would benefit from manipulating. Using a dataset of almost 240,000 IS, we demonstrate that HIV infected cells preferentially expand when integrated within genes linked to cellular processes of transcription, protein modification, cellular replication, and other viral processes. Additionally, we also observed significant clonal expansion of IS within two known proto-oncogenes involved in cellular proliferation pathways which has not been previously reported. This data confirms that clonal expansion occurs early in infection and appears specific to processes important for viral replication or persistence. This reiterates the importance of integration site tracking moving forward for all HIV treatment protocols and trials as an essential benchmark for measuring the effect on and elimination of the latent viral reservoir.

RESULTS

Generating in vivo HIV infected samples for integration site analysis. In order to generate in vivo HIV infected samples for integration site (IS) analysis, we adapted a previously characterized mouse model of infection (24). Humanized mice were generated by engrafting human fetal liver $CD34^+$ hematopoietic stem and progenitor cells (HSPCs) into neonatal non-obese diabetic (NOD)-severe combined immunodeficiency (SCID)-common γ chain-/- (NSG) mice (25, 26). Cells from two unique donors were engrafted in two different litters of mice for a total of 13 animals. Beginning at week 8, blood samples were collected every other week to monitor engraftment and lineage development. Mice were challenged with HIV 16 weeks post engraftment once peripheral levels of human $CD45^+$ and $CD3^+$ T-cells stabilized. At time of challenge, average peripheral human engraftment was 70.1%. This level of engraftment resulted in 23.3% of total blood cells being human $CD3^+$ T-cells, of which 15.4% were $CD4^+$, resulting in a CD4:CD8 ratio of 1.96. Three animals were unchallenged controls, and two animals that were challenged with HIV did not successfully initiate an infection as determined by quantitative viral load PCR and were excluded from the remainder of this study. After HIV challenge, all eight mice that were successfully infected demonstrated a rapid loss of peripheral human $CD45^+$ cells (**Figure 1A**), which corresponded to a loss of $CD3^+$ (**Figure 1B**) and $CD4^+$ cells (**Figure 1C**) when compared to uninfected mock controls. Viral load in mice which successfully initiated infection ranged between 10^5–10^7 viral copies/mL throughout the experiment (**Supplemental Figure 1**). After 28–30 weeks post-engraftment, 12–14 weeks post-infection, mice were sacrificed and lymphoid tissues analyzed for human engraftment and lineage contribution. While all mice exhibited significant depletion in peripheral human $CD45^+$ engraftment, human cells were still detected in bone marrow, spleen, and thymus (**Figure 1D**). Both bone marrow and spleen samples were analyzed for proviral integration site analysis. A summary of all samples used for integration site analysis throughout the manuscript is provided (**Supplemental Table 1**).

Integration sites preferentially cluster on specific chromosomes. A total of 6,259 unique HIV integration sites were identified in bone marrow and spleen samples analyzed from the eight mice (**Table 1**). An average of 80.51% of all IS fell within known gene transcripts and 6.81% within known oncogenes. We first wanted to determine the distribution of these integration sites located on each individual chromosome of the human genome. Assuming an entirely random distribution, IS would occur more frequently on larger chromosomes than smaller chromosomes for any set number of unique sites. When the proportion of observed to expected IS were calculated based on the specific length of each individual chromosome, three chromosomes exhibited a 2.5- to 4-fold enrichment for HIV integrations (**Figure 2A**). It has previously been documented that HIV preferentially integrates into locations of active gene transcription at time of infection (11), and these three chromosomes (chromosome 16, 17, and 19) also have a higher gene density content (27). As a positive control for our in vivo IS database, we also analyzed IS in several in vitro infections of either primary human $CD4^+$ cultured cells, or a well-characterized Jurkat cell line (**Supplemental Figure 2**). These in vitro infections were propagated for up to 28 days post-challenge. When the same analysis was performed for the proportion of integrations relative to expectation in these in vitro infections, the results matched the in vivo data set (**Figure 2B**). The same pattern was also consistent for in vitro infections when using either a *CXCR4* or *CCR5* tropic virus (**Supplemental Figures 3 and 4, and Supplemental Table 2**) although the *CXCR4* infection datasets only represented an n=1 in these studies.

When analyzed either together as one group (All HIV) or broken down by individual condition, the pattern was the same. This suggests that regardless of cell type or model system used, the IS distribution, at least on the chromosomal level, is comparable. This trend was consistent when we analyzed integration site orientation frequency in reference to either the chromosome, or transcriptional direction of IS within genes (**Table 2**). There was essentially a 50/50 split of IS in the forward and reverse orientations for each data group. However, there was

a slight increase, though not statistically significant, in frequency for IS in the forward direction relative to transcript orientation for those IS found within genes. Nevertheless, the HIV in vivo dataset exhibited the greatest increase of integrations in frame with gene transcripts (3.58%) indicating an increased preference for the same transcriptional orientation as transcribed genes.

Specific regions of integration site enrichment present on several chromosomes. We next wanted to transition from a bulk chromosomal view to a higher resolution analysis to determine what differences existed between the in vivo and in vitro integration site datasets. Each chromosome was broken down into a series of consecutive 25kB bins stretching the entire length, and the number of unique integration sites identified within each bin was calculated (**Figure 3A**). Similar to the total chromosomal view, the in vivo integration sites from the mice exhibited an almost identical pattern to the in vitro integration sites despite a large difference in the total number of sites between these data sets, 6,229 and 233,684, respectively. A handful of identified IS (30 for in vivo and 822 for in vitro) could not reliably be mapped to a unique chromosomal location and were excluded from this analysis. Interestingly, the distribution of viral integration sites across each individual chromosome varied greatly, with some individual bins containing almost 10x more integrations than the immediately surrounding bins. These specific 'hot-spot' regions are discussed in more detail later in this manuscript. Despite several of these enriched bins occurring on either end of chromosomes, no distinct pattern of integrations were observed (**Figure 3B**).

In order to ensure the bioinformatics pipeline our lab developed for identifying and aligning integration sites to the genome was not artificially skewing our results, we performed an identical analysis comparing HIV integration sites to two completely unrelated IS data sets: i) human non-hematopoietic based cell lines and ii) human $CD34^+$ HSPCs and lineage progeny both transduced using VSVG-pseudotyped lentiviral vectors (**Supplemental Figure 5 and Supplemental Table 3**). There was still a high degree of similarity between HIV integration sites and lentiviral integration sites in primary HSCPs; however, there was a large difference in chromosomal

integration site distribution when compared to the non-hematopoietic cell populations. This indicates that the high degree of similarity is not a result of data skewing by our bioinformatics analysis but, rather, may arise due to similar transcription profiles in cells of hematopoietic origin. We also analyzed the proportional frequency of integrations across each individual chromosome for each of these data sets (**Supplemental Figure 6**). While some chromosomes exhibited slightly higher integration frequencies at unique locations, most indicated a consistent distribution. One exception to this trend was found at the midpoint of chromosome 11 which indicated a highly-enriched location for integrations in both HIV datasets and in lentivirus-transduced hematopoietic cells. This could represent a region of chromatin that is consistently open and expressed in hematopoietic cells resulting in a hotspot of viral integrations.

Specific genomic regions are enriched for HIV integrations in vivo. Since the distribution of IS across individual bins within each chromosome appeared nearly identical, we next wanted to determine if there were any enriched bins which contained a significantly higher proportion of integration sites occurring specifically in vivo. To achieve this, each bin across the genome that contained integration sites was assigned a frequency for the total number of unique IS occurring in that bin relative to the entire dataset. The HIV in vivo dataset was directly compared to a dataset containing all IS identified in vitro. To determine which individual bins exhibited a differential frequency of IS falling outside the expected normal distribution, the upper and lower 99% confidence intervals were estimated from simulated standard deviations using a bootstrap method. This resulted in a total of 685 bins exhibiting an enrichment of IS within the HIV in vivo dataset and 144 bins within the in vitro dataset (**Figure 4A**). Almost all of these enriched bins within both the in vivo and in vitro dataset contain a gene transcript (93.9% and 97.2%, respectively). The top 10 bins exhibiting the highest enrichment frequency all contained gene transcripts within their boundaries (**Table 3**). The highest enriched gene bin was found to contain the gene *NEAT1*, a long non-coding RNA transcript previously hypothesized to play a role in HIV

infection (28). Other top genes identified, including *MINK1*, *SMARCC1*, *SEPT9*, and *POLR2A*, encode for genes important in cellular signaling cascades or messenger RNA transcription. When all significantly enriched bins in either in vivo or in vitro datasets were analyzed for their specific chromosomal locations, different patterns of enrichment were observed (**Figure 4B**). While there was a relatively equal chromosomal distribution of bins identified in the in vivo dataset, the in vitro dataset was concentrated on chromosomes 16, 17, 19, and 22. Other chromosomes exhibited only sporadic enrichment or none at all. Since this analysis is a measure of unique genomic regions enriched for integrations, the results indicate there is a higher propensity for HIV integrations to occur in specific locations during in vivo infection, hence more regions of enrichment, while in vitro ISs are more equally distributed across the genome, resulting in fewer locations of enrichment.

Expanded clones occur more frequently during in vivo infection. We next analyzed the breakdown of individual clones identified within each experimental infection group to determine both the frequency of expanded clones and the degree of expansion. Our IS protocol utilizes randomized acoustic shearing to fragment the DNA prior to linker ligation. This typically results in two different cells containing identical integration sites to shear at different locations, yielding various lengths of intervening genomic sequence. Counting these various fragment lengths provides a minimum estimate for the total number of cells containing the same integration site, indicating an expanded clone. These clones were broken down into three groups: i) unexpanded clones found in one cell, ii) low-level expanded clones found in two to four cells, or iii) high-level expanded clones found in five or more cells. A cell and its progeny must have undergone at least four cellular divisions to yield five or more identical clones. Since HIV typically shuts down cellular replication machinery during infection, we set a clonal expansion of five as our threshold to be considered an expanded clone. When each group was analyzed for clonal expansion, the in vivo infection dataset exhibited the highest frequency of expansion at both low-level (23.90% of

clones) and high-level (1.92% of clones) expansion (**Figure 5 and Table 4**). The total frequency of highly expanded clones in either of the in vitro primary or cell line infections was 0.01% and 0.04% respectively, despite having a significantly higher number of total integration sites detected. The top expanded clones for the in vivo infection dataset also exhibited higher total cell numbers (36-72 clones) compared to either in vitro primary cells (4-5 clones) or the cell line (6-5 clones).

Two of the top expanded clones found within the in vivo infection dataset occurred in known oncogenes, *JAK2* (**Figure 6A**) and *SEPT9* (**Figure 6B**). *SEPT9* was also the gene identified containing an enriched frequency of integrations during in vivo infections. While there were several other low-level expanded clones occurring in both of these genes, there was only one in *JAK2* and two in *SEPT9* (circled red arrows) which were classified as highly expanded. Additionally, these two clones which exhibited the highest expansion, 72 and 50 respectively, were both in the same transcriptional orientation as the gene. Both expanded integrations in *SEPT9* occurred in what appears to be a hotspot of integrations in intron 2 of the full-length gene transcript. The remaining two expanded clones identified over 50 times did not occur in known oncogenes. We also detected numerous integrations within two other genes which have previously been reported in the literature to contain expanded clones in HIV infected individuals (*MKL2* and *BACH2*) (12, 13); however, we did not detect clones expanded over three cells in any of our data sets (**Supplemental Figure 7**).

Clonally expanded cells are enriched in specific biological pathways. In order to determine if there is a correlation between the individual genes demonstrating the highest level of clonal expansion in each dataset, we utilized the DAVID Bioinformatics Database through the National Institute of Allergy and Infectious Disease (NIAID) (29, 30). We first calculated the average clonal expansion for integrations occurring in each gene identified by integration site analysis, then sorted for the top 1000 genes in each dataset. These genes were analyzed using the biological

processes output tool in DAVID and graphed in a heat map distribution according to the significance of gene enrichment (**Figure 7 and Supplemental Table 4**). A total of 44 pathways were very significantly enriched (p-value < 0.005) within the HIV in vivo dataset specifically, while only one or three pathways were identified in either the in vitro primary or cell line dataset, respectively. In addition to analyzing the HIV integration sites, we also included in the analysis datasets from both the lentiviral-transduced hematopoietic and non-hematopoietic cells and a randomly generated integration site library. There was very little overlap of enriched pathways between the six datasets, with each one containing a unique cluster of several pathways. For example, the processes identified for lentiviral transduced non-hematopoietic cells were primarily involved in metabolic pathways. Within the HIV in vivo dataset, the most significantly enriched pathway contained genes involved in viral processes (p-value 3.63×10^{-8}) indicating a significant preferential expansion of clones containing integration sites within these genes. In addition to the viral process pathway, numerous other relevant pathways that encompass biological processes HIV would benefit from manipulating, such as RNA splicing (p-value 3.39×10^{-5}), protein phosphorylation (p-value 8.31×10^{-4}), transcriptional regulation (p-value 8.12×10^{-4}), and cell division (p-value 2.65×10^{-3}) were significantly enriched.

DISCUSSION

The use of viral integration sites as an important facet for understanding HIV pathogenesis and persistence in clinical studies is becoming more apparent (31-33). Gaining a better understanding of the processes behind integration site selection, how the transition to latent viral infection occurs, and if clonal expansion of infected cells might contribute to the latently infected pool of cells is essential. Here we document for the first time targeted clonal expansions of HIV infected cells in genes significantly enriched for specific biological pathways, despite only representing a small

fraction of total IS. The number of IS demonstrating clonal expansion in this study is most likely due to both the shorter timeline of infection (3 months) and the lack of ART suppression, which would lead to higher levels of viral induced cytotoxicity compared to humans with HIV on treatment. Analyzing samples in this manner lets us determine what the IS profile looks like early during infection when the latent reservoir is most likely seeded, as opposed to after long term treatment before significant clonal outgrowths are expected. Furthermore, these IS frequently occur in pathways the virus would benefit from mediating at the genetic level. These include multiple pathways involved in transcription, translation, and post-translational modification of gene products, as well as pathways involved in mediating cell cycle progression and cellular division. We also observe expanded clones containing an intact LTR sequence in several previously undocumented proto-oncogenes known to play a role in hematopoietic malignant development (34, 35). These findings only occurred during in vivo HIV infection of humanized mice and not during in vitro cell culture infections. This further emphasizes the need for validated pre-clinical models of HIV infection designed for treatment and cure related strategies in order to accurately recapitulate an infection setting.

Viral integrations were consistently detected in genes represented in numerous biological pathways, and these integrations exhibited specific clonal expansion only during in vivo infection. The accessory genes of HIV are known to broadly alter the internal composition of infected cells by hijacking normal phosphorylation (36) and ubiquitin processes (37), mediating viral gene transcription (38), and suppressing immune surveillance and detection (39). This data indicates that cells containing viral integrations within genes relevant to these pathways have a higher frequency of expansion, especially during in vivo infection, indicating preferential IS expansion. Thus, after a random integration event, preferential expansion occurs within the infected pool of cells through manipulation of specific signaling pathways. This ultimately might provide a selective advantage for the virus, such as altering transcription/translation or cell cycle regulation. While these studies are not able to determine whether or not the provirus is intact and capable of

creating infectious viral particles, the presence of an LTR promoter sequence at these loci could be enough to alter transcriptional regulation of these genes (40). Further investigations will be required to determine if the transcriptional profile of these types of integrations actually alter gene expression, especially if they are observed in human clinical patient data.

Recent publications by several groups have documented that clonal expansion does occur in people during viral suppression on cART (12, 13, 41). Furthermore, these expansions often are associated with specific genes and frequently occur in the same transcriptional orientation and intronic location. Specific viral integration events within unique genes have also been linked to increasing cellular proliferation and survival of infected cells (42). While we did observe significantly expanded clonal populations in several genes, including numerous oncogenes, we did not observe expansion in two previously reported genes by other research groups (12, 13, 43). Both *BACH2* and *MKL2* genes were identified in our dataset as containing viral integrations, but none were expanded beyond two cells. We also did not observe the phenomenon of a high frequency identical IS within the same intronic region or same transcriptional orientation, implying that such expansions might only occur during long-term ART suppression. If such a restriction of IS persistence is observed during treatment through multiple-round sampling over time, then this would provide strong selective evidence for targeted clonal expansion. The pattern of equal orientation distribution held true in all datasets, with frequencies of IS occurring in either the same or opposite orientation as gene expression at approximately 50%. The only exception to this was for HIV integrations within genes occurring in vivo, where a slight preference was observed for IS in the same orientation as gene transcription.

Another publication also documented clonal expansion occurring early during acute HIV infection and demonstrated that expansion was more likely to occur when the provirus was integrated in proximity to genes linked to cell activation and chromatin regulation (44). However, their dataset was limited to 18,104 IS, most of which were obtained from in vitro infections. In this manuscript, we document almost 240,000 IS, including over 6,000 from in vivo samples. Similarly,

we observe a clonal expansion of HIV-infected cells that occurs within 12 weeks of infection. A handful of viral IS were greatly expanded (detected in over 50 individual cells) and were obtained from only a fraction of the total tissue or peripheral reservoir. One explanation for clonal expansion of infected cells is through T-cell receptor antigen stimulated cell division. However, these humanized mice are maintained in an enhanced pathogen-free barrier within the mouse facility and all consumable nutrients are sterilized. Therefore, pathogen-mediated antigen stimulation should occur relatively infrequently, although cannot be completely ruled out. Using a dataset from 13 individuals including 6,719 unique IS, another publication demonstrated that most clonally expanded HIV infected cells contained a defective provirus (15). These cells most likely would be detected in assays to measure total proviral content, artificially inflating the frequency of latent cells capable of reinitiating an infection. However, while they might not be functionally able to reactivate, the presence of proviral elements, especially the transcriptionally active LTR sequences, could alter local gene expression through recruitment of positive or negative regulators of transcription.

These data demonstrate that unsuppressed replication during acute infection stages can seed a larger reservoir of clonally expanded cells which may ultimately lead to persistence of these infected cells over time. Additionally, these early clonal expansions are significantly enriched in biological processes beneficial to viral replication and persistence. Several pathways involved in histone regulation, mitotic progression, or cell division were also enriched, supporting an intriguing possibility of IS selection being a determinant for latency. It would also be of great interest to analyze a time course of infection including IS associated with acute, chronic, and treated stages of infection. Such an analysis could provide further details of how ART treatment or other interventions shift the IS landscape and enable direct comparisons between early clonal expansion and the latent reservoir population.

While the mouse model of HIV infection utilized here is frequently used for testing different treatment strategies, we cannot rule out that findings are limited by this fact. In this study, we

present data generated from unsuppressed viral replication in humanized mice, while most IS data generated to date have occurred in ART suppressed, chronically infected humans. Additionally, human data have primarily been collected from peripheral blood samples, while we analyzed viral IS in the lymphoid tissue compartments of bone marrow and spleen due to low volume of blood which can reliable be collected from mice. This does limit the potential interpretations that can be made and could result in any discrepancies between these data and previously published work. Additional studies are certainly required to further investigate these questions and to determine how similar viral IS and clonal expansions observed in animal models are when compared to people infected with HIV. These studies will undoubtedly prove invaluable in better understanding viral persistence and inform treatment options aimed at reducing or eliminating the reservoir and associated clonal expansions of virally infected cells.

Together, our data demonstrate that although clonal expansion can and does occur in all infection settings, clonal outgrowth is statistically correlated to relevant gene pathways during in vivo infections in an animal model of HIV infection, reinforcing the importance of analyzing such pre-clinical model systems for any experimental treatment protocol. Indeed, new methodologies for either treating HIV infection or attempting to reduce the latent reservoir need to perform ISA to quantify the effect these treatments have on this population of cells. These findings also open new possibilities for developing protocols of therapeutic interventions during HIV treatment to mediate identified pathways and aid in the elimination of latently infected cells which persist through standard patient care.

MATERIAL AND METHODS

Human CD34 processing and isolation. Human $CD34^+$ cells were isolated from fetal liver tissue purchased from Advanced Bioscience Resources (Alameda, CA). Tissues were first manually broken down with scalpels and incubated in RPMI medium (Thermo Fisher Scientific, Waltham,

MA) under gentle agitation for 1 hour at 37°C in the presence of 25 µg/mL DNAse (Sigma-Aldrich, St. Louis, MO) and 5 µg/mL Liberase TH (Roche, Basal, Switzerland). Cell suspension was then filtered through a 70 µm filter (BD Bioscience, San Jose, CA) and lysed using hemolytic solution. $CD34^+$ cells were then labeled and isolated using the CD34 MicroBead UltraPure Kit (Miltenyi Biotec, Bergisch Gladbach, Germany) according to manufacturer's protocol.

Mice. NOD.*Cg-Prkdc*^{scid}*Il2rg*^{tm1Wjl}/Szj (NOD SCID gamma$^{-/-}$, [NSG]) mice were purchased from The Jackson Laboratory (Bar Harbor, ME) or bred in-house under approved protocols and in pathogen-free housing conditions. Neonatal mice between one and three days post birth received 150 cGy of radiation, followed 3-4 hours later by a single intrahepatic injection of 1×10^6 $CD34^+$ cells resuspended in 30 µL of PBS containing 1% heparin (Abraxis BioScience, Los Angeles, CA). Since mice were injected as neonates, sex was not taken into account for this study.

NSG mouse sample collection and processing. Blood collection began 8 weeks after transplant and continued every other week through retro-orbital puncture using glass capillary pipettes and collected into EDTA Microtainers (BD Bioscience). A maximum of 200 µL of blood was collected at each time point and diluted 1:1 with PBS. Blood was then centrifuged and serum collected for RNA isolation. Cellular portion was then used for antibody staining and flow cytometry to determine engraftment levels and lineage contribution. At time of necropsy, tissues harvested included bone marrow, spleen, and thymus, in addition to peripheral blood. Tissue samples were passed through a 70 µm filter (BD Bioscience) and washed with PBS. Blood and tissue samples were stained with appropriate fluorescently conjugated antibodies for florescence-activated cell sorting (FACS) for 15 minutes at room temperature. Red blood cells were removed by incubation in BD FACS Lysing Solution (BD Bioscience), which was diluted out using PBS prior to analysis. All cells were acquired on a FACS Canto II (BD Bioscience) and analyzed using FlowJo software v10.1 (BD Bioscience). Up to 20,000 viable cells from blood and 100,000 viable

cells from tissues were acquired when possible. Gates were established using full minus one (FMO)-stained controls. Samples were stained at a 1:20 dilution using human CD45-PerCP (Clone 2D1), mouse CD45.1/CD45.2-V500 (Clone 30-F11), CD3-FITC or allophycocyanin (APC) (Clone UCHT1), CD4-V450 (Clone RPA-T4), CD8-APC-Cy7 (Clone SK1), CD20-phycoerythrin (PE) (Clone 2H7), CD14-APC or PE-Cy7 (Clone M5E2), and CD34-APC (Clone 581). All antibodies were acquired from BD Bioscience. Once mice where infected with HIV, all samples were fixed using 10% neutral-buffered formalin solution (Sigma-Aldrich) for 10 minutes after antibody staining and red blood cell lysis.

HIV virus preparation and mouse challenge. Mice were challenged by a single intraperitoneal injection of 200 µL of HIV-1 virus containing 2.5×10^5 infections units. Viral strains used included HIV-1 BaL, NL4.3, and JRCSF. All three viral strains were obtained through the National Institute of Health (NIH) AIDS Reagent Program (ARP) Division of AIDS, National Institute of Allergy and Infectious Diseases (NIAID). The BaL strain (Cat# 510) was deposited by Dr. Suzanne Garter, Dr. Mikula Popovic, and Dr. Robert Gallo (45). The NL4.3 strain (Cat# 114) was deposited by Dr. Malcolm Martin (46). The JRCSF strain (Cat# 2708) was deposited by Dr. Irvin SY Chen and Dr. Yoshio Koyanagi (47). BaL virus was propagated in PM1 cells (Cat# 3038) obtained through the NIH ARP, Division of AIDS, NIAID, and deposited by Dr. Marbin Reitz (48). NL4.3 and JRCSF were obtained as proviral clones, and transfected into 293T cells using TurboFect Transfection Reagent (Thermo Fisher) according to manufacturer's protocols. In all cases, supernatant was collected from cells and filtered through 0.22 mm filters (EMD Millipore, Burlington, MA) and tittered on GHOST cells (Cat# 3942) obtained through the NIH ARP, and deposited by Dr. Vineet N. KewalRamani and Dr. Dan R. Littman (49) according to provided protocols.

In vitro HIV infections. Jurkat reporter cells (JLTRG) were obtained from the NIH ARP, Division of AIDS, NIAID (Cat # 11587) and deposited by Dr. Olaf Kutsch (50, 51). Primary $CD4^+$ cells were

isolated from adult human apheresis product purchased from the Hematopoietic Cell Processing and Repository facility at Fred Hutchinson Cancer Research Center. Cells were enriched using the CD4 MicroBead kit (Miltenyi) according to manufacturer's protocols. Cells were infected with HIV at an MOI ranging between 0.1–0.001 diluted in RPMI (Thermo Fisher) containing 1% pen/strep (Thermo Fisher) for 4 hours at 10×10^6 cells per mL. Cells were then washed with PBS to remove unbound viral particles, and resuspended at 1×10^6 cells per mL in culture media (RMPI, 1% pen/strep, 10% fetal bovine serum [FBS, Atlas Biologicals, Fort Collins, CO]). $CD4^+$ cells were cultured in the same culture media supplemented with 50 µg/mL human recombinant interleukin-2 (IL2) and 5 µg/mL phytohaemagglutinin (PHA). Cells were propagated for up to 3 weeks and split 1:2 every 3–4 days with fresh culture media. Cells were pelleted and genomic DNA extracted for IS analysis. Infection in JLTRG cells was tracked by flow cytometry for GFP expression after collection and cellular fixation using 10% neutral-buffered formalin solution (Sigma-Aldrich) for 10 minutes. Gates were set using uninfected cell controls to contain <1% GFP^+ cells.

Quantitative viral load PCR. Viral RNA was extracted from mouse serum or tissue culture supernatant using the QIAamp Viral RNA Mini Kit (QIAGEN, Hilden, Germany) as previously described (52). Briefly, viral RNA was then analyzed using the TaqMan RNA-to-Ct 1-Step Kit (Thermo Fisher) using primer and probes specific to the long terminal repeat (LTR) region (F: 5'-GCCTCAATAAAGCTTGCCTTGAG-3', R: 5'-GGCGCCACTGCTAGAGATTTTC-3', Probe: FAM 5'-AAGTAGTGTGTGCCCGTCTGTTRTKTGACT-3' TAMARA). Plates were analyzed on an ABI TaqMan 7500 real-time PCR system (Thermo Fisher).

Lentivirus-transduced cell populations. Human hematopoietic cells containing integrated lentiviral vectors were compiled from historical data, some of which have previously been published (52, 53), using integration site analysis from either clinical or preclinical samples. For a

non-hematopoietic IS library, HeLa (obtained through ATCC) or GHOST (obtained through NIH ARP) cells were transduced at an MOI of 10 using a SIN LV (pRSC-hPGK.eGFP) produced with a third-generation split packaging system and pseudotyped by the vesicular stomatitis virus G protein (VSVG). These vectors were produced by our institutional Vector Production Core (Director: Hans-Peter Kiem) at the Fred Hutchinson Cancer Research Center. Infectious titer was determined by flow cytometry evaluating EGFP expression following titrated transduction of HT1080 human fibrosarcoma-derived cells.

Integration site processing. Integration site analysis was performed on spleen and bone marrow samples from mice and cell culture samples as previously described (53, 54), with the following modifications: DNA was extracted from cells using the DNeasy Blood and Tissue Kit (QIAGEN), and up to 3 µg was randomly sheared using an M220 focused ultrasonicator (Covaris, Woburn, MA). Fragmented DNA was purified, polished (End-It DNA End Repair Kit; Epicenter, Madison, WI), and ligated to modified linker cassettes containing known primer binding sites. This product was amplified using sequential nested exponential PCR. Product from first PCR was purified, and eluted DNA was diluted prior to a second nester PCR which added both barcodes and sequences required for compatibility with the next-generation sequencing MiSeq platform (Illumina, San Diego, CA). Sequencing was performed by the Genomics Core Facility at the Fred Hutchinson Cancer Research Center. Sequences for all primers and linker cassettes used are provided in the supplement (**Supplemental Table 5**). IS were identified using a bioinformatics platform as previously described in detail (52).

Transduction filter methods. Crossover integration sites appearing in distinct samples originating from unique transduction events were present in the data. Theoretically, this should never be observed and is likely the result of contamination, barcode swapping, or other errors in processing. In some cases, it is possible to determine which sample the IS originates from by

comparing the number of genomically aligned sequence reads representing the IS in each sample. When examining collisions, the genomically aligned sequence counts were used instead of normalized frequencies to avoid biases introduced by low capture frequency in samples with few genomically aligned reads because the log-base ten-fold difference between the most (131,641) and fewest (25) genomically aligned reads across samples was large (3.72).

Using a custom python script, a list of all collisions was generated. Each transduction event was parsed for observations of IS in the collision list. For each transduction event in which a collision IS was detected, the mean count of the IS for samples in which it was detected was recorded. For example, an IS at chr9:5,069,731 was observed in four transduction events. In the first transduction event, it was observed in two samples where it was represented by 174 and 77 genomically aligned sequence reads. It was observed in one sample from each of the other transduction events where it was represented by a single genomically aligned read. The mean count of the IS in the first transduction event was 125.5 and one for all other transduction events.

The ratio of mean counts from each transduction event was compared to the maximum mean count from a single transduction event. If a transduction event had a mean count greater than or equal to one half of the maximum mean count for the IS, the IS was discarded from the dataset; otherwise, the IS was kept for the transduction event in which it had the highest count and removed from the other samples. In other words, if the ratio of the maximum mean count to the next highest mean count was greater than 1:1 (½ or 0.5), the IS was discarded. If the ratio was less than 1:1, the IS was retained in the transduction event where it had the highest count and removed from all others. Returning to the previous example, 1: 125.5 is 0.008. In this case, all non-maximum genomically aligned read counts fall below 0.5, and the IS is retained in the first transduction event dataset and removed from all other transduction event datasets. Overall, from an initial set of 377,643 unique IS, 6,072 collisions (1.6%) were detected and 2,166 (0.57%) were unresolvable (removed from all datasets).

Significant bin comparison. For each sample group, unique integration sites were converted to C-Start positions on a concatenation of the entire genome to ensure the maximum number of equally sized bins. An integration site's new position, *L*, is given by the equation:

$$L = (\sum_{c=0}^{m-1} S_c) + l,$$

where *c* is a chromosome number, *m* is an integration site's chromosome number, S_c is the size of chromosome *c*, and *l* is the C-Start of an integration site. For this equation, the size of chromosome zero is assumed to be zero, and chromosomes X and Y are coded as chromosomes 23 and 24, respectively. Chromosome sizes were obtained from the Genome Reference Consortium (https://www.ncbi.nlm.nih.gov/grc/human/data). Integration sites that mapped to unincorporated contigs were not included in this analysis. As an example, in the HIV in vitro data set, there is an integration site at chr4: 55,428,089. To determine its new position, *L,* we sum the sizes of all preceding chromosomes (chr1, chr2, chr3) and add the C-Start, or

$$L = (\sum_{c=0}^{4-1} S_c) + 55{,}428{,}089,$$

$$L = (S_1 + S_2 + S_3) + 144{,}267{,}527\,,$$

$$L = (248{,}956{,}422 + 242{,}193{,}529 + 198{,}295{,}559) + 55{,}428{,}089,$$

$$L = 744{,}873{,}599\,.$$

The bin number, *B*, in which the new position falls is given by the equation

$$B = \lceil L \div w \rceil,$$

where *w* is the window size (or number of base pairs in each bin). Using our previous example and a window size of 25kb, the bin number, *B,* is the ceiling of the quotient obtained from dividing the new position, *L*, by the window size, *w*, or

$$B = \lceil 744{,}873{,}599 \div 25{,}000 \rceil,$$

$$B = \lceil 29{,}794.94 \rceil,$$

$$B = 29{,}795.$$

The number of integration sites falling within each bin of the concatenated genome, was then counted for each sample group. The normalized frequency of integrations for a given bin *B*, *f(B)*, was calculated using the equation

$$f(B) = k\_B \div \sum_{B=1}^{n} k\_B,$$

where k_B is the number of integration sites falling within a given bin *B*, and *n* is the number of bins in the linearized genome. For example, in the bin identified previously, *B* = 29,795, there were a total of three unique integration sites, a total of 123,531 bins, and 7,295 unique integration sites fell within those bins, so

$$f(29{,}795) = 3 \div 7{,}295,$$

$$f(29{,}795) = 0.0004.$$

Once the normalized frequencies were calculated for the two samples of interest, a bootstrap analysis was conducted to estimate which bins have differential frequencies falling outside the expected distribution. First, any bins with zero integration sites in both samples were removed from the analysis because their inclusion can cause slight differences to appear significant if the number of bins with integrations is small relative to the total number of bins in the concatenated genome (data not shown). The differential frequency, d, or distance between a given bin in the first sample, B_1, and a given bin in the second sample, B_2, is given by the equation

$$d = f(B_1) - f(B_2).$$

The upper and lower limits, D, of the expected distribution were calculated as the points falling three standard deviations above or below the mean differential frequency (where 99% of frequencies should fall) and is given by the equation

$$D = \{\underline{d} \pm 3 \times \sigma(d)\}.$$

For the actual bootstrapping, the *boot()* function in **R** was used (https://cran.r-project.org/web/packages/boot/boot.pdf). The *boot()* function allowed for the construction of simulated comparisons. A simulated data set was constructed for each sample by randomly sampling from actual bin frequencies, *f(B)*, for each sample until a full set of bin frequencies was

obtained. The synthetic sets of bin frequencies were used to calculate new estimates of the upper and lower limits, *D*, of the expected distribution of differential frequencies. Using the *boot()* function, 2,000 simulations were conducted for both the upper and lower limits. The limits reported in the figure(s) represent the mean value of the 2,000 simulations for the upper and lower limits. For the figures, only differential bin values falling above or below those extremes are plotted.

Annotating integration sites. We compared the genomic locations of all HIV integration sites to the latest human genome (hg38 – Genome Reference Consortium Human Build 38) refseq gene list available from UCSC genome browser. This was achieved using a custom python script. Several attributes of each gene in the refseq list were memorized: the NCBI gene name, chromosome, strand, transcription start site, transcription end site, exon count, exon and intron positions, and alternate gene name. For each sample group, the genomic location of each integration site was compared to the data obtained from the refseq file. Integrations were annotated with distance to the nearest transcription start site, NCBI and alternate gene names for the gene with the nearest transcription start site, and whether the integration site falls within a gene or not. If the integration site falls within a gene, some additional information is recorded as well: the NCBI and alternate gene names for the gene(s) in which the integration site falls within, from which strand (forward or reverse) the gene(s) is(are) transcribed, whether the integration site is within an intron or exon, and which intron or exon the integration falls within.

DAVID analysis. For each sample group the annotated integrations were combined and used as input in a custom Java script to identify and tally all integration sites that occurred within genes. For each gene, basic statistics were calculated from the integrations falling within the gene: the highest, average, and median number of genomically aligned reads and read fragment lengths; the percentage of integrations located on the forward and reverse strands; and, the percentage of integrations falling within introns and exons. COSMIC database was used to identify oncogenic

genes. Each output file was sorted on highest average read fragment length and the top 1000 genes were used as input gene lists in the DAVID Bioinformatics tool (https://david.ncifcrf.gov/summary.jsp) (29, 30). Gene ontology information concerning significant biological pathways was attained through the Functional Annotation Tool.

Circos plot generation. A custom python script was used to split each chromosome in the human genome into consecutive 25 kb regions. Additionally, for each sample group, the number of integration events that occurred in each bin was recorded. The output of this script was then used as input to Circos (http://circos.ca/) to create dual histogram plots. An additional Circos plot was created to visualize the presence or absence of enriched bins identified using the 'bin comparison method'.

Statistics. Statistical analysis was performed as described in detail in previous methods sections. For the transduction filter applied to samples, statistics were performed in **R**, for significant bin comparisons, statistics were performed in python scripts, and for DAVID analysis, statistics were performed within the publically available analysis package and provided as p-values.

Study approval. All animal studies were carried out in compliance with approved protocol number 1864 by the Institutional Animal Care and Use Committee (IACUC) at the Fred Hutchinson Cancer Research Center.

AUTHOR CONTRIBUTIONS

H.P.K. is the principal investigator of the study and coordinated the overall execution of the project. K.G.H. conceived, designed, coordinated, and executed the experiments. L.E.S. and Z.K.N. wrote all bioinformatics scripts used in data analysis, L.E.S. created visual representation

of data including Circos plot outputs, and Z.K.N. created the bin comparison method. C.I. processed mouse and culture samples for analysis and prepared samples for sequencing. J.E.A. provided critical feedback for experimental design and data analysis and provided lab space and reagents for sample processing. K.G.H. wrote the manuscript, which was critically reviewed by J.E.A. and H.P.K.

ACKNOWLEDGMENTS

We thank Helen Crawford for help preparing and formatting this manuscript. We also thank Sarah Weitz, Melissa Comstock, and Don Parrilla for assistance with performing many of the mouse procedures and general colony maintenance, Biswajit Paul for assistance with the in vitro HIV infections, Cristina McAllister for sample processing and preparation, Daniel Humphrys for assistance with integration site analysis processing, and Martin Wohlfahrt and Donald Gisch for producing the lentiviral vector controls used in this manuscript. We also thank all former Kiem lab members who contributed to integration site processing and analysis from hematopoietic derived cells that was used in this manuscript, and the Comparative Medicine staff at Fred Hutchinson Cancer Research Center for support with vivarium maintenance. Hans-Peter Kiem is a Markey Molecular Medicine Investigator and received support as the inaugural recipient of the Jose Carreras/E. Donnall Thomas Endowed Chair for Cancer Research and the Endowed Chair for Cell and Gene Therapy. This work was supported by Hans-Peter Kiem's endowment fund.

CONFLICTS OF INTEREST

The authors have declared that no conflict of interest exists.

REFERENCES

1. Gallo RC, Sarin PS, Gelmann EP, Robert-Guroff M, Richardson E, Kalyanaraman VS, Mann D, Sidhu GD, Stahl RE, Zolla-Pazner S, et al. Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS). *Science.* 1983;220(4599):865-7.
2. Barre-Sinoussi F, Chermann JC, Rey F, Nugeyre MT, Chamaret S, Gruest J, Dauguet C, Axler-Blin C, Vezinet-Brun F, Rouzioux C, et al. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science.* 1983;220(4599):868-71.
3. Fischer M, Hafner R, Schneider C, Trkola A, Joos B, Joller H, Hirschel B, Weber R, Gunthard HF, and Swiss HIVCS. HIV RNA in plasma rebounds within days during structured treatment interruptions. *AIDS.* 2003;17(2):195-9.
4. Harrigan PR, Whaley M, and Montaner JS. Rate of HIV-1 RNA rebound upon stopping antiretroviral therapy. *AIDS.* 1999;13(8):F59-62.
5. Taylor S, Boffito M, Khoo S, Smit E, and Back D. Stopping antiretroviral therapy. *AIDS.* 2007;21(13):1673-82.
6. Henrich TJ, Hanhauser E, Marty FM, Sirignano MN, Keating S, Lee TH, Robles YP, Davis BT, Li JZ, Heisey A, et al. Antiretroviral-free HIV-1 remission and viral rebound after allogeneic stem cell transplantation: report of 2 cases. *Annals of Internal Medicine.* 2014;161(5):319-27.
7. Peterson CW, Haworth KG, Polacino P, Huang ML, Sykes C, Obenza WM, Repetto AC, Kashuba A, Bumgarner R, DeRosa SC, et al. Lack of viral control and development of combination antiretroviral therapy escape mutations in macaques after bone marrow transplantation. *AIDS.* 2015;29(13):1597-606.
8. Craigie R, and Bushman FD. HIV DNA integration. *Cold Spring Harb Perspect Med.* 2012;2(7):a006890.

9. Mullins JI, and Frenkel LM. Clonal expansion of human immunodeficiency virus-infected cells and human immunodeficiency virus persistence during antiretroviral therapy. *J Infect Dis.* 2017;215(Suppl_3):S119-S27.
10. Siliciano RF, and Greene WC. HIV latency. *Cold Spring Harb Perspect Med.* 2011;1(1):a007096.
11. Schroder AR, Shinn P, Chen H, Berry C, Ecker JR, and Bushman F. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell.* 2002;110(4):521-9.
12. Maldarelli F, Wu X, Su L, Simonetti FR, Shao W, Hill S, Spindler J, Ferris AL, Mellors JW, Kearney MF, et al. HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science.* 2014;345(6193):179-83.
13. Wagner TA, McLaughlin S, Garg K, Cheung CY, Larsen BB, Styrchak S, Huang HC, Edlefsen PT, Mullins JI, and Frenkel LM. HIV latency. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. *Science.* 2014;345(6196):570-3.
14. Hughes SH, and Coffin JM. What integration sites tell us about HIV persistence. *Cell Host Microbe.* 2016;19(5):588-98.
15. Cohn LB, Silva IT, Oliveira TY, Rosales RA, Parrish EH, Learn GH, Hahn BH, Czartoski JL, McElrath MJ, Lehmann C, et al. HIV-1 integration landscape during latent and active infection. *Cell.* 2015;160(3):420-32.
16. Lusic M, Marcello A, Cereseto A, and Giacca M. Regulation of HIV-1 gene expression by histone acetylation and factor recruitment at the LTR promoter. *EMBO Journal.* 2003;22(24):6550-61.
17. Malim MH, and Emerman M. HIV-1 accessory proteins--ensuring viral survival in a hostile environment. *Cell Host Microbe.* 2008;3(6):388-98.
18. Goedert JJ. The epidemiology of acquired immunodeficiency syndrome malignancies. *Semin Oncol.* 2000;27(4):390-401.

19. Grogg KL, Miller RF, and Dogan A. HIV infection and lymphoma. *J Clin Pathol.* 2007;60(12):1365-72.
20. Epeldegui M, Vendrame E, and Martinez-Maza O. HIV-associated immune dysfunction and viral infection: role in the pathogenesis of AIDS-related lymphoma. *Immunol Res.* 2010;48(1-3):72-83.
21. Alexaki A, Liu Y, and Wigdahl B. Cellular reservoirs of HIV-1 and their role in viral persistence (Review). *Current HIV Research.* 2008;6(5):388-400.
22. Dahabieh MS, Battivelli E, and Verdin E. Understanding HIV latency: the road to an HIV cure. *Annu Rev Med.* 2015;66(407-21.
23. Wang GP, Ciuffi A, Leipzig J, Berry CC, and Bushman FD. HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Research.* 2007;17(8):1186-94.
24. Holt N, Wang J, Kim K, Friedman G, Wang X, Taupin V, Crooks GM, Kohn DB, Gregory PD, Holmes MC, et al. Human hematopoietic stem/progenitor cells modified by zinc-finger nucleases targeted to CCR5 control HIV-1 in vivo. *Nature Biotechnology.* 2010;28(8):839-47.
25. Lepus CM, Gibson TF, Gerber SA, Kawikova I, Szczepanik M, Hossain J, Ablamunits V, Kirkiles-Smith N, Herold KC, Donis RO, et al. Comparison of human fetal liver, umbilical cord blood, and adult blood hematopoietic stem cell engraftment in NOD-scid/gammac-/-, Balb/c-Rag1-/-gammac-/-, and C.B-17-scid/bg immunodeficient mice. *Hum Immunol.* 2009;70(10):790-802.
26. Patton J, Vuyyuru R, Siglin A, Root M, and Manser T. Evaluation of the efficiency of human immune system reconstitution in NSG mice and NSG mice containing a human HLA.A2 transgene using hematopoietic stem cells purified from different sources. *J Immunol Methods.* 2015;422(13-21.

27. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. Initial sequencing and analysis of the human genome [erratum appears in Nature 2001 Aug 2;412(6846):565 Note: Szustakowki, J [corrected to Szustakowski, J]. *Nature.* 2001;409(6822):860-921.
28. Zhang Q, Chen CY, Yedavalli VS, and Jeang KT. NEAT1 long noncoding RNA and paraspeckle bodies modulate HIV-1 posttranscriptional expression. *MBio.* 2013;4(1):e00596-12.
29. Huang DW, Sherman BT, and Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols.* 2009;4(1):44-57.
30. Huang DW, Sherman BT, and Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research.* 2009;37(1):13-Jan.
31. Kiselinova M, De Spiegelaere W, and Vandekerckhove L. The use of HIV-1 integration site analysis information in clinical studies aiming at HIV cure. *J Virus Erad.* 2016;2(3):175-6.
32. Maldarelli F. The role of HIV integration in viral persistence: no more whistling past the proviral graveyard. *J Clin Invest.* 2016;126(2):438-47.
33. Maldarelli F. HIV-infected cells are frequently clonally expanded after prolonged antiretroviral therapy: implications for HIV persistence. *J Virus Erad.* 2015;1(4):237-44.
34. Mahon FX. JAK the trigger. *Oncogene.* 2005;24(48):7125-6.
35. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, and Stratton MR. A census of human cancer genes (Review). *Nature Reviews Cancer.* 2004;4(3):177-83.
36. Francis AC, Di Primio C, Allouch A, and Cereseto A. Role of phosphorylation in the nuclear biology of HIV-1. *Curr Med Chem.* 2011;18(19):2904-12.
37. Arora S, Verma S, and Banerjea AC. HIV-1 Vpr redirects host ubiquitination pathway. *J Virol.* 2014;88(16):9141-52.

38. Karn J, and Stoltzfus CM. Transcriptional and posttranscriptional regulation of HIV-1 gene expression. *Cold Spring Harb Perspect Med.* 2012;2(2):a006916.
39. Kamp W, Berk MB, Visser CJ, and Nottet HS. Mechanisms of HIV-1 to escape from the host immune surveillance. *Eur J Clin Invest.* 2000;30(8):740-6.
40. Geiss GK, Bumgarner RE, An MC, Agy MB, van 't Wout AB, Hammersmark E, Carter VS, Upchurch D, Mullins JI, and Katze MG. Large-scale monitoring of host cell gene expression during HIV-1 infection using cDNA microarrays. *Virology.* 2000;266(1):8-16.
41. Kok YL, Vongrad V, Shilaih M, Di Giallonardo F, Kuster H, Kouyos R, Gunthard HF, and Metzner KJ. Monocyte-derived macrophages exhibit distinct and more restricted HIV-1 integration site repertoire than CD4(+) T cells. *Sci Rep.* 2016;6(24157.
42. Ali ASM, Mowbray C, Lanz M, Stanton A, Bowen S, Varley CL, Hilton P, Brown K, Robson W, Southgate J, et al. Targeting deficiencies in the TLR5 mediated vaginal response to treat female recurrent urinary tract infection. *Sci Rep.* 2017;7(1):11039.
43. Ikeda T, Shibata J, Yoshimura K, Koito A, and 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.
44. Satou Y, Katsuya H, Fukuda A, Misawa N, Ito J, Uchiyama Y, Miyazato P, Islam S, Fassati A, Melamed A, et al. Dynamics and mechanisms of clonal expansion of HIV-1-infected cells in a humanized mouse model. *Sci Rep.* 2017;7(1):6913.
45. Gartner S, Markovits P, Markovitz DM, Kaplan MH, Gallo RC, and Popovic M. The role of mononuclear phagocytes in HTLV-III/LAV infection. *Science.* 1986;233(4760):215-9.
46. Adachi A, Gendelman HE, Koenig S, Folks T, Willey R, Rabson A, and Martin MA. Production of acquired immunodeficiency syndrome-associated retrovirus in human and nonhuman cells transfected with an infectious molecular clone. *Journal of Virology.* 1986;59(2):284-91.

47. Koyanagi Y, Miles S, Mitsuyasu RT, Merrill JE, Vinters HV, and Chen IS. Dual infection of the central nervous system by AIDS viruses with distinct cellular tropisms. *Science.* 1987;236(4803):819-22.
48. Lusso P, Cocchi F, Balotta C, Markham PD, Louie A, Farci P, Pal R, Gallo RC, and Reitz MS, Jr. Growth of macrophage-tropic and primary human immunodeficiency virus type 1 (HIV-1) isolates in a unique CD4+ T-cell clone (PM1): failure to downregulate CD4 and to interfere with cell-line-tropic HIV-1. *J Virol.* 1995;69(6):3712-20.
49. Morner A, Bjorndal A, Albert J, KewalRamani VN, Littman DR, Inoue R, Thorstensson R, Fenyo EM, and Bjorling E. 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.
50. Ochsenbauer-Jambor C, Jones J, Heil M, Zammit KP, and Kutsch O. T-cell line for HIV drug screening using EGFP as a quantitative marker of HIV-1 replication. *Biotechniques.* 2006;40(1):91-100.
51. Kutsch O, Levy DN, Bates PJ, Decker J, Kosloff BR, Shaw GM, Priebe W, and Benveniste EN. Bis-anthracycline antibiotics inhibit human immunodeficiency virus type 1 transcription. *Antimicrob Agents Chemother.* 2004;48(5):1652-63.
52. Haworth KG, Ironside C, Norgaard ZK, Obenza WM, Adair JE, and Kiem HP. In vivo murine-matured human CD3+ cells as a preclinical model for T cell-based immunotherapies. *Mol Ther Methods Clin Dev.* 2017;6(17-30.
53. Adair JE, Beard BC, Trobridge GD, Neff T, Rockhill JK, Silbergeld DL, Mrugala M, and Kiem HP. Extended survival of glioblastoma patients after chemoprotective HSC gene therapy. *Science Translational Medicine.* 2012;4(133):133ra57.
54. Adair JE, Johnston SK, Mrugala MM, Beard BC, Guyman LA, Baldock AL, Bridge CA, Hawkins-Daarud A, Gori JL, Born DE, et al. Gene therapy enhances chemotherapy tolerance and efficacy in glioblastoma patients. *J Clin Invest.* 2014;124(9):4082-92.

FIGURES AND FIGURE LEGENDS

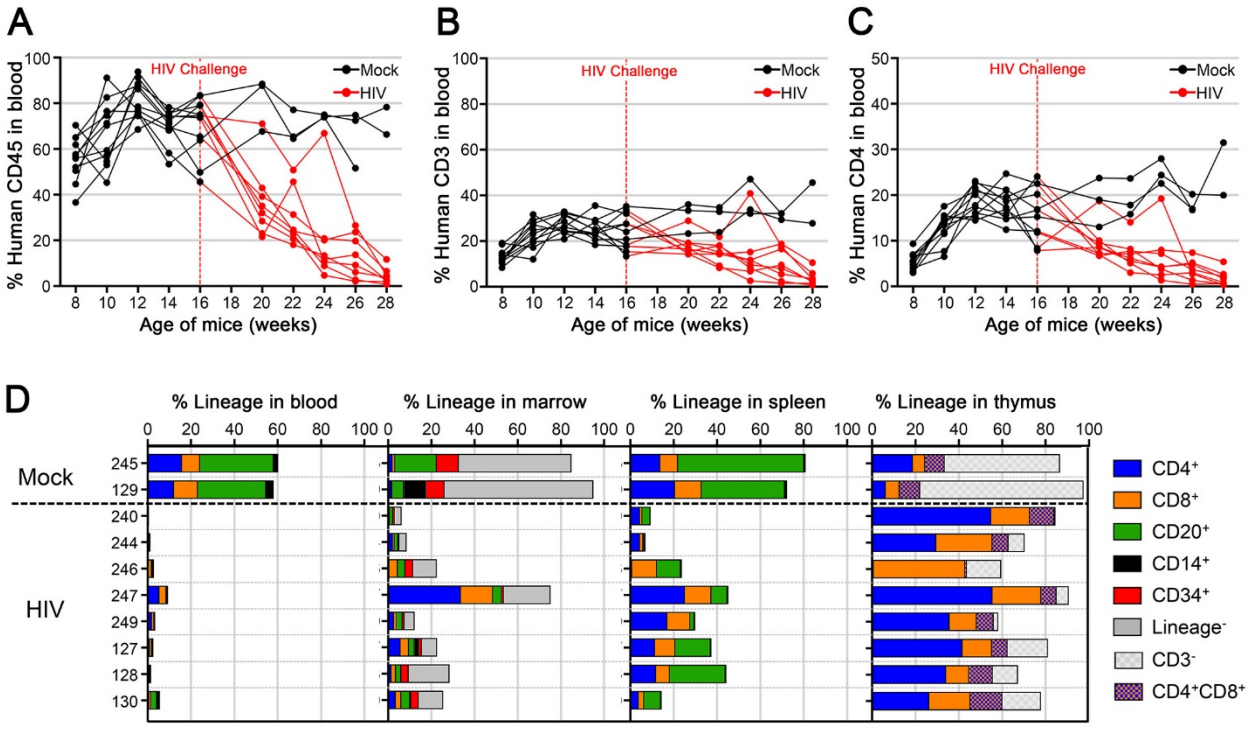


Figure 1. Depletion of circulating human CD4⁺ cells in periphery of infected mice. Neonate NSG mice were engrafted with human CD34⁺ cells at birth and challenged with HIV at 16 weeks of age after development of CD4⁺ T-cells. (**A**) Peripheral human CD45⁺ engraftment in blood samples over time for two cohorts of mice. Each line represents a single mouse and is colored red after infection at 16 weeks, indicated by vertical dashed red line. (**B**) CD3⁺ and (**C**) CD4⁺ T-cell development over time for individual mice, represented as percent of cells in total blood. Mock mice represented as black lines (n=3) and HIV infected mice represented as red lines (n=8). (**D**) Human engraftment at time of necropsy in peripheral blood, bone marrow, spleen, and thymus of mice as indicated on upper x-axis. Bar length corresponds to total human CD45⁺ engraftment, and is broken down into stacked boxes representing specific cell lineages. Individual animal numbers listed on y-axis and horizontal dashed black line separates mock from HIV infected mice. One mock mouse died early and was excluded from necropsy analysis.

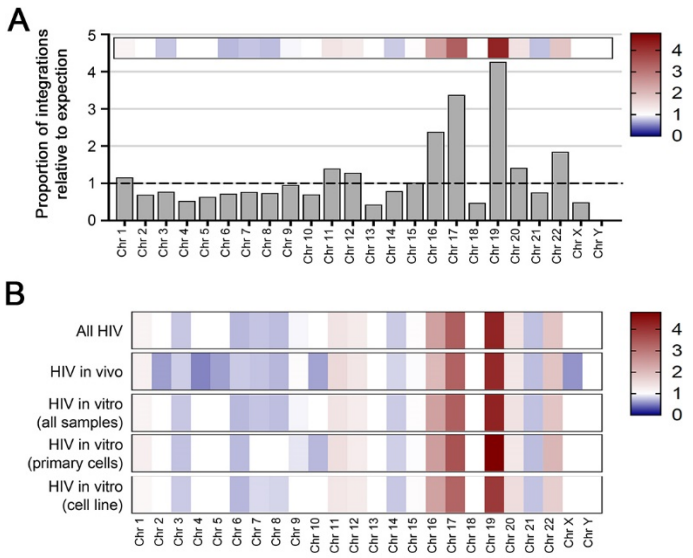


Figure 2. Distribution of HIV integration sites across all human chromosomes. Visual representation of the frequency for HIV viral integration events on each individual chromosome. (**A**) All HIV integrations from both in vivo or in vitro datasets were combined and plotted based on frequency of integrations relative to unbiased expectation of chromosome size with larger chromosomes expected to contain more unique integrations when compared to smaller ones. The y-axis is proportion of integrations observed relative to expectation based on chromosome size with horizontal dashed black line at a value of 1. Overlaid heat map at top of histogram corresponds to height of each bar and is color coded with higher frequencies than expected as darker red, lower frequency as darker blue, and at expectation as white. (**B**) Heat map representation of different HIV datasets broken down by infection system or cell type, and color code is same as previously described.

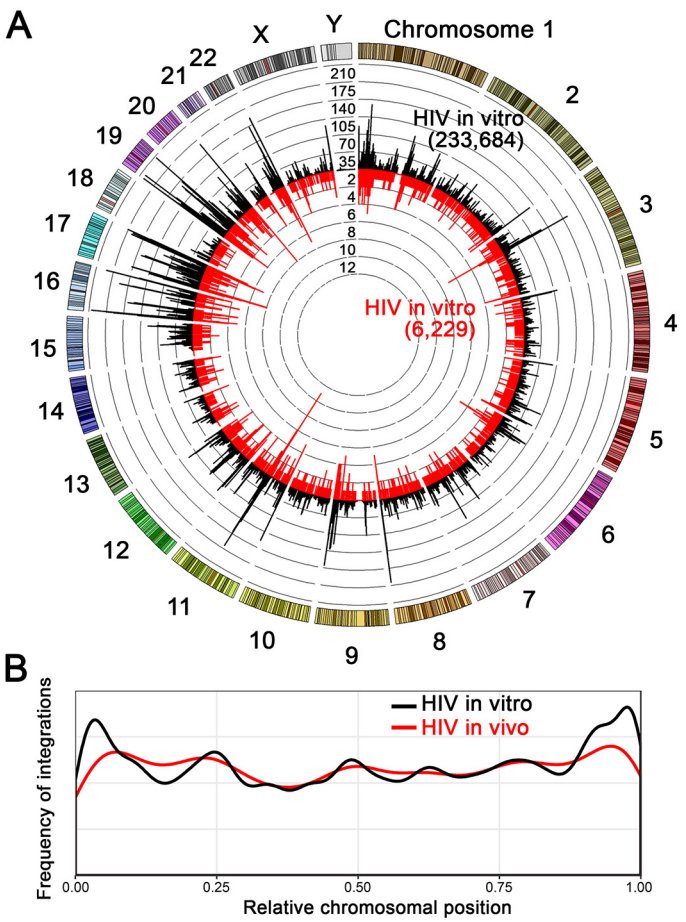


Figure 3. HIV integration sites cluster in specific chromosomal regions. Integration sites identified in either in vivo NSG mouse experiments or in vitro tissue culture infections were analyzed. (**A**) Circos plot depicting IS sites across the genome. Each chromosome is represented on the exterior of the ring and is broken down into sequential bins each 25kB in size. The total number of unique integrations occurring within each bin is represented by the height of histogram bars with black bars radiating outward depicting integrations found in vitro (233,684 IS) and red bars radiating inward depicting those found in vivo (6,229 IS). Concentric rings function as a y-axis and have incremental values of 25 for the black histogram and 2 for the red histogram. (**B**) Specific location of integration relative to chromosomal length is plotted for either in vitro infections (black line) or in vivo infections (red line). The x-axis represents relative chromosomal position with 0.25, meaning 25% of the distance from beginning of chromosome and y-axis represents the frequency of observing an integration at each position throughout the chromosome.

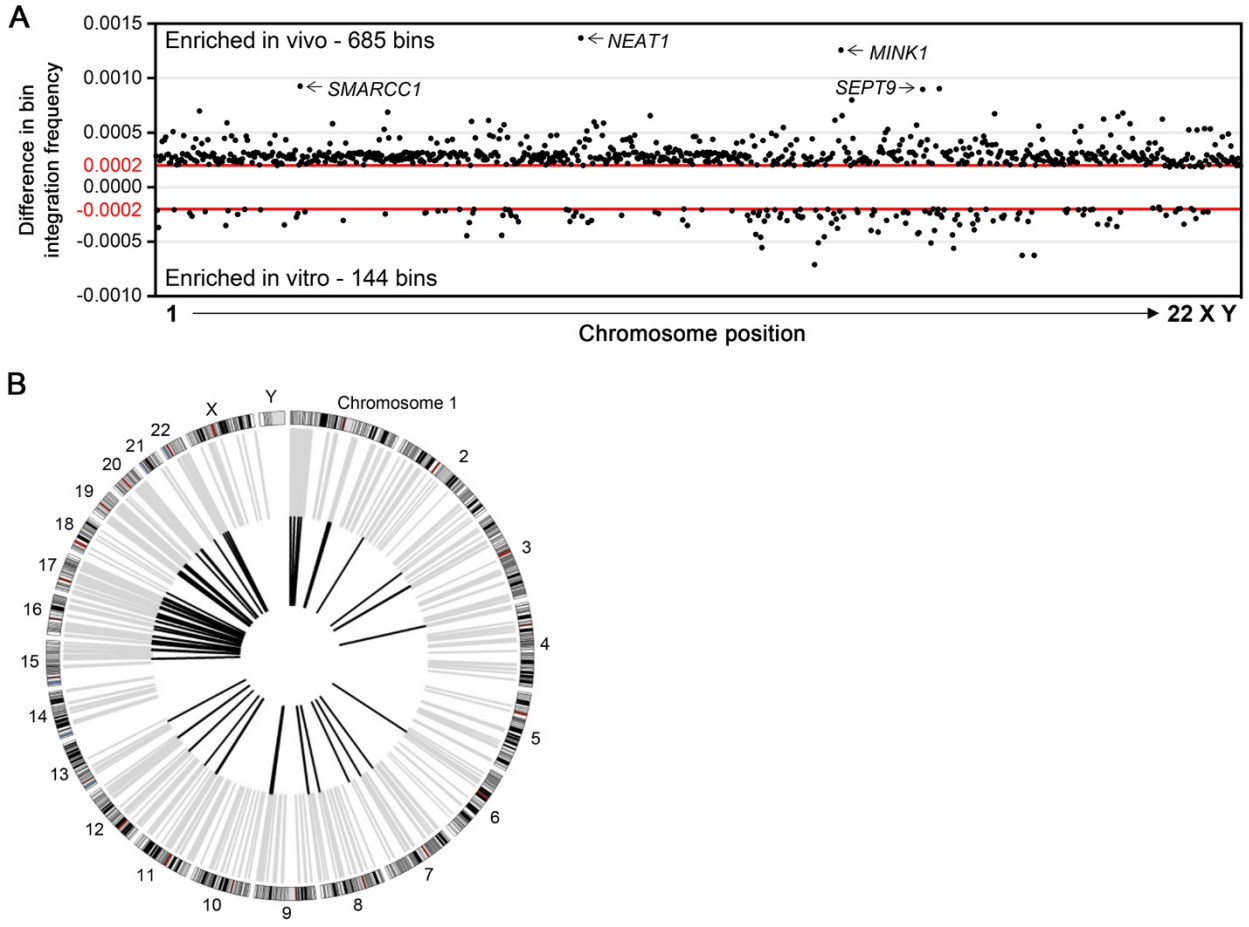


Figure 4. Specific chromosomal regions are enriched for viral integration sites. The HIV in vivo IS dataset was directly compared to in vitro IS to determine if there were specific locations within the genome that were significantly enriched for IS during in vivo infection. (**A**) Dot plot representing significantly enriched 25kB bin segments in one dataset compared to the other. Bins in the positive direction on y-axis were enriched for during in vivo infection (685 bins), while bins in the negative direction were enriched for during in vitro infections (144 bins). X-axis represents relative chromosomal position, but is not normalized for chromosome length. Top four bins enriched in vivo are labeled with the genes contained at that locus. Horizontal red lines indicate the 99th percentile of simulated data, similar to a p-value of 0.01. (**B**) Circos plot highlighting the bin locations for each enriched dataset. Bins containing integrations enriched during in vivo infection are plotted in gray and those enriched during in vitro infection in black. Chromosomes represented on exterior of ring.

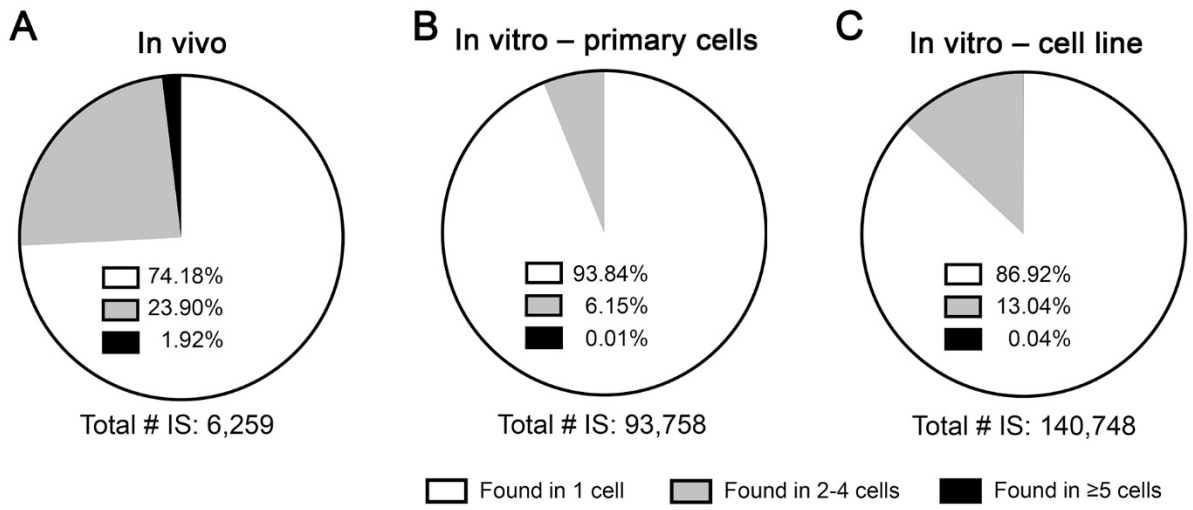


Figure 5. Expanded clones occur more frequently during in vivo infection. All IS identified were classified into three groups of expansions and plotted in pie charts for (**A**) HIV in vivo integrations, (**B**) HIV in vitro primary cell integration, and (**C**) HIV in vitro cell line integrations. Proportion of IS found in one cell for each group represented by white area, two to four cells represented with gray area, and five or more cells represented as black area. Actual percentages for each category are listed in each pie chart.

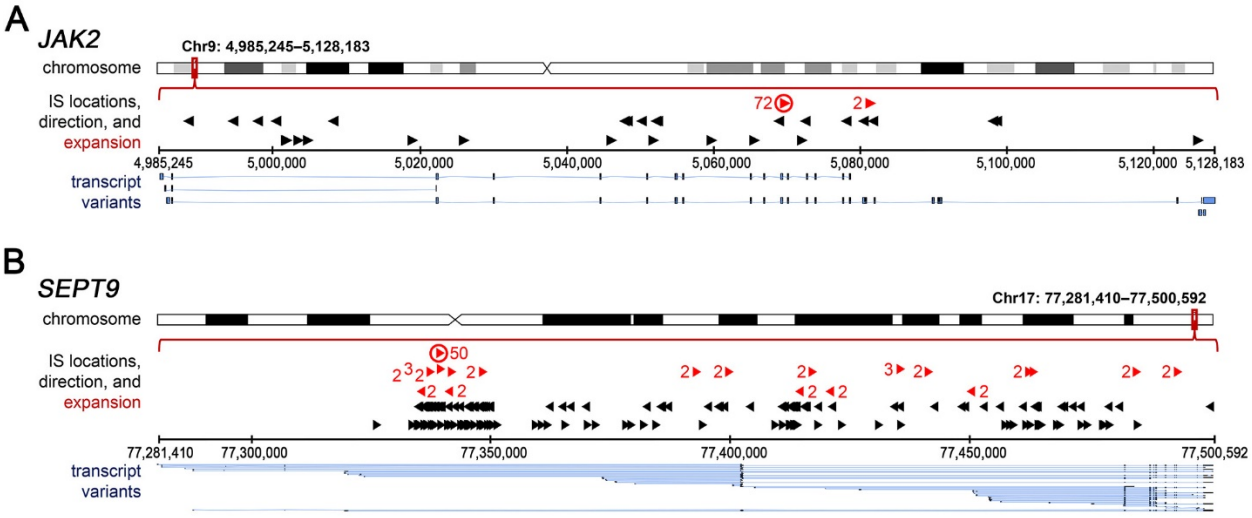


Figure 6. Clonal expansion observed in known oncogenes. Individual gene plots indicating the location of all detected integrations within two oncogenes identified as containing significantly expanded clones (**A**) *JAK2* and (**B**) *SEPT9*. Gene name is listed at top left for each group and specific location on chromosome is highlighted by red box. Gene transcript is expanded below chromosome, each arrow indicates a unique IS, and arrow direction depicts orientation. Black arrows represent an IS found in one cell and red arrows represent an IS found in two or more, with the number of clones detected noted next to each arrow. Circled red arrows indicate expanded clones over five cells. Blue lines at bottom of each graph represent all known transcript variants of each gene.

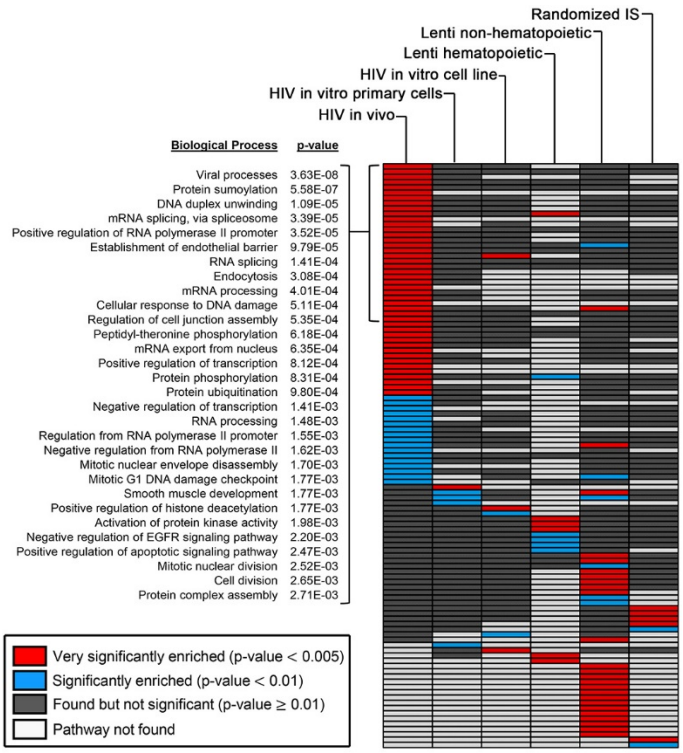


Figure 7. Clonally expanded cells are significantly enriched in specific gene pathways. Heat map of biological processes identified using publically available DAVID database. The top 1000 genes containing expanded clones were analyzed for each group listed at top of each column. All biological processes identified are plotted as a row of individual boxes within each column and color coded based on significance. Red boxes indicate a biological pathway that was very significantly enriched in the dataset (p-value < 0.005), blue boxes were significantly enriched (p-value < 0.01), gray boxes were observed but not significant (p-value ≥ 0.01) and white boxes indicate a pathway that was not found in the given dataset. The top 30 pathways represented with red boxes in HIV in vivo dataset are broken down at left with each biological process named and associated p-value listed.

TABLES

Table 1. Integration site characteristics within individual mice

Mouse ID	Unique # IS	IS Within Genes	Intragenic %	Oncogenic %
240	428	354	82.71%	6.07%
244	346	287	82.95%	6.07%
246	144	112	77.78%	9.03%
247	3084	2432	78.86%	6.42%
249	1664	1391	83.59%	7.45%
127	252	203	80.56%	7.14%
128	81	62	76.54%	4.94%
130	260	198	76.15%	8.46%
TOTAL/AVG	**6259**	**5039**	**80.51%**	**6.95%**

Table 2. Number and orientation of IS in each dataset

Group	Total # of IS	% Forward orientation	Total IS within genes	% Forward orientation in genes	% Forward orientation of total vs. within gene
All HIV	240,765	50.08%	192,602	50.40	0.32
HIV in vivo	6,259	48.73%	5,039	52.31	3.58
HIV in vitro (all samples)	234,506	50.11%	187,563	50.35	0.24
HIV in vitro (primary cell)	93,758	50.13%	76,735	50.91	0.78
HIV in vitro (cell line)	140,748	50.10%	110,828	49.96	-0.14

Total number of IS found within genes and the orientation frequency for those sites in relation to gene transcript, and the difference of this orientation preference between total and within gene IS.

Table 3. Top 10 chromosomal bins enriched in vivo

Chromosome	Bin Position	Genes within Bin	Frequency Difference
Chr 11	65,418,950 – 65,443,949	*NEAT1*	1.3680E-03
Chr 17	4,844,439 – 4,869,438	*MINK1*	1.2586E-03
Chr 3	47,675,050 – 47,700,049	*SMARCC1*	9.2693E-04
Chr 17	81,244,439 – 81,269,438	*SLC38A10*	9.0333E-04
Chr 17	77,319,439 – 77,344,438	*SEPT9*	8.9697E-04
Chr 17	7,469,439 – 7,494,438	*POLR2A*, *ZBTB4*	8.0076E-04
Chr 1	39,300,001 – 39,325,000	*MACF1*	6.9999E-04
Chr 6	35,651,677 – 35,676,676	*FKBP5*	6.8936E-04
Chr 22	50,516,947 – 50,541,946	*NCAPH2*, *SCO2*, *TYMP*, *ODF3B*	6.8080E-04
Chr 19	17,388,713 – 17,413,712	*BST2*, *BSPR*	6.7432E-04

Chromosome number and specific nucleotide position, are included as well as gene transcripts falling within the window and the calculated frequency difference when compared to in vitro IS dataset for each one.

Table 4. Top 8 expanded clones within genes

In vivo, Total # IS: 6,259			
	Gene	**# Cells**	**Position**
1	*JAK2*	72	Chr9: 5,069,731
2	*CUL2*	61	Chr10: 35,030,954
3	*DENND5A*	52	Chr11: 9,179,134
4	*SEPT9*	50	Chr17: 77,339,546
5	*COG6*	49	Chr13: 39,696,821
6	*RRAS2*	43	Chr11: 14,290,096
7	*USP12*	39	Chr13: 27,077,135
8	*PPP6R3*	36	Chr11: 68,490,436

In vitro – primary cells, Total # IS: 93,758			
	Gene	**# Cells**	**Position**
1	*KDM2A*	5	Chr11: 67,170,694
2	*STK4*	5	Chr20: 44,972,952
3	*LIMA1*	5	Chr12: 50,243,017
4	*SET*	5	Chr9: 128,694,271
5	*PEX2*	5	Chr8: 76,987,392
6	*ARFGEF1*	5	Chr8: 67,254,071
7	*MSRA*	5	Chr8: 10,358,511
8	*IP6K1*	4	Chr3: 49,743,345

In vitro – cell line, Total # IS: 140,748			
	Gene	**# Cells**	**Position**
1	*NLK*	6	Chr17: 28,077,013
2	*SMG1P5*	6	Chr16: 30,311,106
3	*ANO6*	6	Chr12: 45,301,663
4	*PPARA*	6	Chr22: 46,207,255
5	*ESYT2*	6	Chr7: 158,746,347
6	*KMT2D*	6	Chr12: 49,022,494
7	*FBXO45*	6	Chr3: 196,575,569
8	*PHACTR4*	5	Chr1: 28,419,522

The total number of IS in each dataset is listed, which includes chromosomal position, total number of cells containing the IS, and the gene the IS falls within.

SUPPLEMENTAL INFORMATION for

HIV infection results in clonal expansions containing integrations within pathogenesis-related biological pathways

Kevin G. Haworth[1], Lauren E. Schefter[1], Zachary K. Norgaard[1], Christina Ironside[1], Jennifer E. Adair[1, 2], Hans-Peter Kiem[1,2,3].

List of Supplemental Items:

1. **Supplemental Tables 1–5.**
2. **Supplemental Figures 1–6.**

Supplemental Table 1. Breakdown of each individual sample used for IS analysis

HIV

#	File name	In vitro/ in vivo	HIV strain	HIV tropism	Unique # IS	# IS forward orientation	% Forward orientation	#IS within genes	% IS within genes
1	Jurkat MOI 0.1	In vitro	BaL	*CCR5*	37316	18727	50.18%	29425	78.85%
2	Jurkat MOI 0.001	In vitro	BaL	*CCR5*	23651	11821	49.98%	18842	79.67%
3	Jurkat MOI 0.1	In vitro	BaL	*CCR5*	20888	10528	50.40%	16604	79.49%
4	Jurkat NL4.3	In vitro	NL4.3	*CXCR4*	9166	4561	49.76%	7021	76.60%
5	Jurkat BAL	In vitro	BaL	*CCR5*	6859	3381	49.29%	5475	79.82%
6	Jurkat JRCSF	In vitro	JRCSF	*CCR5*	13326	6608	49.59%	9997	75.02%
7	Donor XHS5 CD4+	In vitro	BaL	*CCR5*	1078	534	49.54%	851	78.94%
8	Donor 1501522A CD4+	In vitro	BaL	*CCR5*	41798	20838	49.85%	34274	82.00%
9	Donor 1501522B CD4+	In vitro	BaL	*CCR5*	50882	25630	50.37%	41610	81.78%
10	Jurkat MOI 0.01	In vitro	BaL	*CCR5*	29623	14926	50.39%	23530	79.43%
11	Mouse 127 bone marrow	In vivo	BaL	*CCR5*	113	49	43.36%	85	75.22%
12	Mouse 127 spleen	In vivo	BaL	*CCR5*	141	63	44.68%	120	85.11%
13	Mouse 128 bone marrow	In vivo	BaL	*CCR5*	24	15	62.50%	18	75.00%
14	Mouse 128 spleen	In vivo	BaL	*CCR5*	57	28	49.12%	44	77.19%
15	Mouse 130 bone marrow	In vivo	BaL	*CCR5*	245	121	49.39%	187	76.33%
16	Mouse 130 spleen	In vivo	BaL	*CCR5*	15	10	66.67%	11	73.33%
17	Mouse 240 bone marrow	In vivo	BaL	*CCR5*	352	165	46.88%	292	82.95%
18	Mouse 240 spleen	In vivo	BaL	*CCR5*	76	35	46.05%	62	81.58%
19	Mouse 244 bone marrow	In vivo	BaL	*CCR5*	291	148	50.86%	245	84.19%
20	Mouse 244 spleen	In vivo	BaL	*CCR5*	55	29	52.73%	42	76.36%
21	Mouse 246 bone marrow	In vivo	NL4.3	*CXCR4*	74	32	43.24%	53	71.62%
22	Mouse 246 spleen	In vivo	NL4.3	*CXCR4*	71	34	47.89%	59	83.10%
23	Mouse 247 bone marrow	In vivo	JRCSF	*CCR5*	2095	1038	49.55%	1661	79.28%
24	Mouse 247 spleen	In vivo	JRCSF	*CCR5*	1039	512	49.28%	812	78.15%
25	Mouse 249 bone marrow	In vivo	BaL	*CCR5*	438	204	46.58%	359	81.96%
26	Mouse 249 spleen	In vivo	BaL	*CCR5*	1243	602	48.43%	1044	83.99%

Non-hematopoietic-lenti

#	File name	Primary/Cell line	Unique # IS	# IS Forward orientation	% Forward orientation	# IS within genes	% IS within genes
1	GHOST #1	Cell line	1827	928	50.79%	1285	70.33%
2	GHOST #2	Cell line	2118	1050	49.58%	1566	73.94%
3	GHOST #3	Cell line	5105	2615	51.22%	3662	71.73%
4	GHOST #4	Cell line	10966	5434	49.55%	7364	67.15%

5	GHOST #5	Cell line	894	458	51.23%	687	76.85%
6	HeLa #1	Cell line	5896	2889	49.00%	4192	71.10%
7	HeLa #3	Cell line	1326	665	50.15%	930	70.14%
8	HeLa #2	Cell line	2147	1083	50.44%	1449	67.49%
9	HeLa #4	Cell line	804	419	52.11%	602	74.88%
10	HeLa #5	Cell line	1671	830	49.67%	1223	73.19%

Hematopoietic-lenti

#	In vitro/ in vivo	Primary/cell Line	Vector	Unique # IS	# IS Forward orientation	% Forward orientation	# IS within genes	% IS within genes
1	In vivo	Primary cell	Lenti	975	452	46.36%	633	64.92%
2	In vivo	Primary cell	Lenti	1171	614	52.43%	695	59.35%
3	In vivo	Primary cell	Lenti	49	24	48.98%	39	79.59%
4	In vivo	Primary cell	Lenti	738	371	50.27%	415	56.23%
5	In vivo	Primary cell	Lenti	624	331	53.04%	331	53.04%
6	In vivo	Primary cell	Lenti	46	21	45.65%	27	58.70%
7	In vivo	Primary cell	Lenti	420	218	51.90%	286	68.10%
8	In vivo	Primary cell	Lenti	36	17	47.22%	31	86.11%
9	In vivo	Primary cell	Lenti	737	394	53.46%	394	53.46%
10	In vitro	Primary cell	Lenti	24144	12038	49.86%	18748	77.65%
11	In vitro	Primary cell	Lenti	8729	4416	50.59%	6483	74.27%
12	In vitro	Primary cell	Lenti	5604	2788	49.75%	4574	81.62%
13	In vitro	Primary cell	Lenti	10639	5324	50.04%	8489	79.79%
14	In vitro	Primary cell	Lenti	6985	3511	50.26%	5694	81.52%
15	In vitro	Primary cell	Lenti	11029	5529	50.13%	8956	81.20%
16	In vitro	Primary cell	Lenti	6449	3234	50.15%	5238	81.22%
17	In vitro	Primary cell	Lenti	7466	3730	49.96%	5990	80.23%
18	In vitro	Primary cell	Lenti	8390	4205	50.12%	6649	79.25%
19	In vitro	Primary cell	Lenti	668	326	48.80%	420	62.87%
20	In vitro	Primary cell	Lenti	789	392	49.68%	461	58.43%
21	In vitro	Primary cell	Lenti	264	146	55.30%	195	73.86%
22	In vitro	Primary cell	Lenti	742	370	49.87%	497	66.98%
23	In vitro	Primary cell	Lenti	886	396	44.70%	600	67.72%
24	In vitro	Primary cell	Lenti	278	127	45.68%	161	57.91%
25	In vitro	Primary cell	Lenti	303	153	50.50%	174	57.43%
26	In vitro	Primary cell	Lenti	469	237	50.53%	326	69.51%
27	In vitro	Primary cell	Lenti	356	174	48.88%	209	58.71%
28	In vitro	Primary cell	Lenti	422	222	52.61%	244	57.82%

29	In vitro	Primary cell	Lenti	613	319	52.04%	334	54.49%
30	In vitro	Primary cell	Lenti	209	92	44.02%	117	55.98%
31	In vitro	Primary cell	Lenti	760	373	49.08%	483	63.55%
32	In vitro	Primary cell	Lenti	422	225	53.32%	300	71.09%
33	In vitro	Primary cell	Lenti	110	57	51.82%	49	44.55%
34	In vitro	Primary cell	Lenti	173	79	45.66%	147	84.97%
35	In vitro	Primary cell	Lenti	230	116	50.43%	196	85.22%
36	In vitro	Primary cell	Lenti	263	142	53.99%	236	89.73%
37	In vitro	Primary cell	Lenti	214	92	42.99%	184	85.98%
38	In vitro	Primary cell	Lenti	165	93	56.36%	146	88.48%

Overview of all samples used for integration site analysis in this manuscript for (**a**) HIV infections, (**b**) Non-Hematopoietic-Lenti, and (**c**) Hematopoietic-Lenti. Table includes number of unique IS observed, chromosomal orientation, and within gene frequency.

Supplemental Table 2. Top 8 enriched bins between CCR5 and CXCR4.

Chromosome	Bin Position	Genes within Bin	Frequency Difference
***CCR5*-enriched**			
Chr 16	67,057,784 – 67,082,783	*CBFB*	5.8937E-04
Chr 8	144,274,725 – 144,299,724	*BOP1*	5.7593E-04
Chr 17	75,344,439 – 75,369,438	*GRB2*	5.6849E-04
Chr 17	75,369,439 – 75,394,438	*GRB2*	5.4880E-04
Chr 16	88,582,784 – 88,607,783	*ZC3H18*	5.4640E-04
Chr 17	4,219,429 – 4,244,438	*ANKFY1*	5.2167E-04
Chr 19	1,063,713 – 1,088,712	*ABCA7*	4.9214E-04
Chr 17	80,694,439 – 80,719,428	*RPTOR*	4.7245E-04
***CXCR4*-enriched**			
Chr 9	91,311,089 – 91,336,088	*AUH*	-7.0842E-04
Chr 17	78,044,439 – 78,069,438	*TNRC6C*	-6.7890E-04
Chr 9	121,186,089 – 121,211,088	*RAB14*	-6.2968E-04
Chr 20	32,821,097 – 32,846,096	*MAPRE1*	-5.2886E-04
Chr 11	93,718,950 – 93,743,949	*CEP295*	-5.2766E-04
Chr 4	157,054,491 – 157,079,490	*GLRB*	-5.2766E-04
Chr 9	81,036,089 – 81,061,088	none	-5.0797E-04
Chr 4	102,504,491 – 102,529,490	*NFKB1*	-5.0797E-04

Top 8 enriched bins within either the in vitro Jurkat *CCR5* or in vitro Jurkat *CXCR4* infection datasets. Each contains the chromosome number and specific nucleotide position, as well as gene transcripts falling within the window and the calculated frequency difference.

Supplemental Table 3. Compiled information for each IS dataset.

RIS Group Name	Cell Source	Source of Integrations	# of Samples	# of Unique IS	# of IS in Genes
HIV in vivo – humanized mice	Spleen and bone marrow	HIV infection	16	6,259	5,039
HIV in vitro – primary cells	Peripheral CD4+ cells	HIV infection	3	93,758	76,735
HIV in vitro – cell lines	Jurkat cells	HIV infection	7	140,748	110,828
Lenti – hematopoietic	Blood, bone marrow, or cultured cells	Lentiviral transduction	38	101,959	78,697
Lenti – non-hematopoietic	Ghost and HeLa cells	Lentiviral transduction	14	32,753	22,960

Overview for each of the 5 independent datasets analyzed in this manuscript including cell source, source of integration, total number of unique sequencing runs, and corresponding number of unique IS and number within genes for each. The all HIV dataset is a compilation of the first three rows.

Supplemental Table 4. Reference list of biological pathways enriched for IS

Biological Process	HIV in vivo	HIV in vitro - primary cells	HIV in vitro - cell lines	Lentivirus - hematopoietic	Lentivirus non-hematopoietic	Randomized IS library
GO:0016032~viral process	3.627E-08	1.626E-01	8.859E-01	Not found	9.528E-01	5.838E-01
GO:0016925~protein sumoylation	5.575E-07	6.441E-01	9.098E-01	9.051E-02	4.925E-01	6.060E-01
GO:0032508~DNA duplex unwinding	1.088E-05	1.488E-01	Not found	Not found	3.536E-01	Not found
GO:0000398~mRNA splicing, via spliceosome	3.394E-05	4.656E-01	8.820E-02	1.442E-02	4.993E-01	Not found
GO:0045944~positive regulation of transcription from RNA polymerase II promoter	3.523E-05	2.777E-01	9.562E-01	7.432E-02	7.539E-01	1.486E-01
GO:0061028~establishment of endothelial barrier	9.785E-05	Not found	Not found	Not found	Not found	Not found
GO:0008380~RNA splicing	1.407E-04	7.937E-01	1.432E-01	Not found	4.013E-01	7.588E-01
GO:0006897~endocytosis	3.077E-04	6.129E-02	1.903E-01	Not found	5.020E-01	2.228E-02
GO:0006397~mRNA processing	4.012E-04	7.372E-01	7.212E-01	Not found	3.559E-01	6.955E-01
GO:0006974~cellular response to DNA damage stimulus	5.108E-04	1.768E-01	6.215E-01	3.269E-05	5.422E-01	2.234E-01
GO:1901888~regulation of cell junction assembly	5.348E-04	Not found	Not found	Not found	Not found	Not found
GO:0018107~peptidyl-threonine phosphorylation	6.183E-04	Not found	8.331E-01	Not found	2.739E-01	Not found
GO:0006406~mRNA export from nucleus	6.355E-04	6.935E-01	8.443E-01	4.079E-02	7.124E-01	9.899E-01
GO:0045893~positive regulation of transcription, DNA-templated	8.121E-04	4.509E-01	4.185E-01	Not found	4.200E-01	1.723E-01
GO:0006468~protein phosphorylation	8.306E-04	8.632E-02	1.538E-01	Not found	3.220E-02	3.923E-01
GO:0042787~protein ubiquitination involved in ubiquitin-dependent protein catabolic process	9.800E-04	1.035E-01	2.717E-01	9.559E-02	6.759E-03	3.785E-01
GO:0045892~negative regulation of transcription, DNA-templated	1.410E-03	5.589E-01	8.842E-01	1.490E-02	1.203E-01	6.477E-01
GO:0006396~RNA processing	1.477E-03	8.377E-01	4.909E-03	Not found	9.368E-02	4.397E-01
GO:0006357~regulation of transcription from RNA polymerase II promoter	1.546E-03	5.013E-01	8.058E-01	8.210E-02	2.195E-01	4.065E-02
GO:0000122~negative regulation of transcription from RNA polymerase II promoter	1.621E-03	4.977E-01	9.460E-01	1.550E-02	2.896E-01	6.171E-01
GO:0007077~mitotic nuclear envelope disassembly	1.702E-03	3.391E-01	Not found	Not found	Not found	3.151E-01
GO:0031571~mitotic G1 DNA damage checkpoint	1.769E-03	4.095E-01	Not found	Not found	Not found	Not found
GO:0048745~smooth muscle tissue development	1.769E-03	4.095E-01	Not found	Not found	Not found	Not found
GO:0031065~positive regulation of histone deacetylation	1.769E-03	Not found	Not found	Not found	Not found	Not found
GO:0032147~activation of protein kinase activity	1.978E-03	8.844E-01	Not found	Not found	1.680E-01	3.275E-01
GO:0042059~negative regulation of epidermal growth factor receptor signaling pathway	2.197E-03	8.219E-01	Not found	Not found	8.297E-01	4.837E-01
GO:1902231~positive regulation of intrinsic apoptotic signaling pathway in response to DNA damage	2.472E-03	Not found	Not found	Not found	Not found	Not found
GO:0007067~mitotic nuclear division	2.518E-03	1.866E-01	3.519E-01	Not found	4.125E-03	9.316E-01
GO:0051301~cell division	2.652E-03	2.362E-02	3.446E-01	8.762E-02	3.079E-02	9.545E-01
GO:0006461~protein complex assembly	2.714E-03	7.971E-01	6.217E-01	Not found	3.215E-01	3.531E-02
GO:0006606~protein import into nucleus	2.816E-03	2.867E-01	9.351E-01	Not found	9.424E-01	7.397E-01

Biological Process	HIV in vivo	HIV in vitro - primary cells	HIV in vitro - cell lines	Lentivirus - hematopoietic	Lentivirus non-hematopoietic	Randomized IS library
GO:0017148~negative regulation of translation	2.816E-03	9.381E-01	9.351E-01	4.976E-02	5.308E-01	7.397E-01
GO:0006511~ubiquitin-dependent protein catabolic process	2.930E-03	3.461E-01	8.478E-01	4.189E-02	9.404E-01	5.723E-01
GO:0010467~gene expression	3.020E-03	3.906E-01	6.834E-02	Not found	7.867E-02	Not found
GO:0007265~Ras protein signal transduction	3.143E-03	6.417E-01	9.632E-01	Not found	1.846E-02	9.598E-01
GO:0048008~platelet-derived growth factor receptor signaling pathway	3.257E-03	Not found	Not found	Not found	7.597E-01	Not found
GO:0007283~spermatogenesis	3.347E-03	7.216E-01	Not found	Not found	3.018E-01	8.236E-01
GO:0046777~protein autophosphorylation	3.629E-03	9.884E-01	3.941E-01	Not found	1.240E-01	3.648E-01
GO:0000389~mRNA 3'-splice site recognition	4.159E-03	Not found	Not found	Not found	Not found	Not found
GO:0035735~intraciliary transport involved in cilium morphogenesis	4.159E-03	Not found	Not found	Not found	Not found	Not found
GO:0043161~proteasome-mediated ubiquitin-dependent protein catabolic process	4.487E-03	9.477E-02	3.316E-01	6.015E-03	5.109E-01	8.093E-01
GO:0060999~positive regulation of dendritic spine development	4.671E-03	Not found	Not found	Not found	8.591E-02	2.586E-01
GO:0043547~positive regulation of GTPase activity	4.740E-03	9.514E-01	8.704E-01	Not found	5.028E-02	3.364E-02
GO:0006928~movement of cell or subcellular component	4.790E-03	7.726E-01	Not found	Not found	5.959E-01	Not found
GO:1900034~regulation of cellular response to heat	5.208E-03	1.379E-01	6.765E-01	Not found	7.040E-01	9.681E-01
GO:0006409~tRNA export from nucleus	5.444E-03	4.449E-01	Not found	Not found	Not found	Not found
GO:1900026~positive regulation of substrate adhesion-dependent cell spreading	5.444E-03	Not found	Not found	Not found	1.717E-02	7.695E-01
GO:0008360~regulation of cell shape	6.082E-03	1.204E-01	3.156E-01	Not found	5.096E-01	1.016E-02
GO:0006198~cAMP catabolic process	6.198E-03	Not found	2.912E-02	Not found	5.216E-01	4.972E-01
GO:0001843~neural tube closure	6.282E-03	9.752E-01	8.749E-01	Not found	5.079E-01	8.652E-01
GO:0010827~regulation of glucose transport	6.368E-03	4.608E-01	Not found	Not found	Not found	Not found
GO:0030168~platelet activation	6.608E-03	6.289E-01	4.356E-01	Not found	6.511E-01	7.629E-01
GO:0006260~DNA replication	6.715E-03	1.107E-01	8.457E-01	Not found	1.702E-02	8.291E-01
GO:0007062~sister chromatid cohesion	7.015E-03	1.090E-01	9.537E-01	Not found	3.784E-03	9.485E-01
GO:0034644~cellular response to UV	7.045E-03	Not found	Not found	Not found	3.536E-01	Not found
GO:0016567~protein ubiquitination	7.074E-03	2.392E-01	4.840E-01	4.398E-02	2.755E-01	7.401E-01
GO:0000209~protein polyubiquitination	7.302E-03	2.601E-02	6.146E-01	Not found	5.238E-01	7.224E-01
GO:0071158~positive regulation of cell cycle arrest	8.286E-03	Not found	Not found	Not found	7.074E-01	6.822E-01
GO:0001701~in utero embryonic development	8.582E-03	5.128E-01	6.328E-01	Not found	3.563E-02	2.178E-01
GO:0006661~phosphatidylinositol biosynthetic process	9.745E-03	Not found	5.023E-01	Not found	6.173E-03	1.173E-01
GO:0030033~microvillus assembly	9.950E-03	Not found	Not found	4.080E-02	5.664E-01	Not found
GO:0031297~replication fork processing	4.121E-01	1.303E-03	Not found	Not found	Not found	Not found
GO:0006886~intracellular protein transport	3.854E-02	7.576E-03	7.845E-02	Not found	2.568E-05	6.604E-02

Biological Process	HIV in vivo	HIV in vitro - primary cells	HIV in vitro - cell lines	Lentivirus - hematopoietic	Lentivirus non-hematopoietic	Randomized IS library
GO:0007030~Golgi organization	7.456E-01	7.417E-03	6.676E-01	Not found	8.681E-03	8.500E-01
GO:0000387~spliceosomal snRNP assembly	7.752E-01	8.846E-03	Not found	Not found	Not found	Not found
GO:0006915~apoptotic process	8.968E-02	2.667E-01	1.610E-03	Not found	1.057E-01	7.864E-01
GO:0007399~nervous system development	7.315E-01	6.952E-01	7.009E-03	Not found	3.027E-01	2.191E-01
GO:0032088~negative regulation of NF-kappaB transcription factor activity	3.076E-01	2.376E-01	8.442E-01	1.462E-03	4.448E-01	3.939E-01
GO:0032922~circadian regulation of gene expression	3.427E-01	1.646E-02	7.443E-01	4.356E-03	5.448E-02	2.513E-01
GO:0006888~ER to Golgi vesicle-mediated transport	8.430E-01	2.162E-01	7.501E-01	4.176E-04	5.030E-01	8.489E-01
GO:0016310~phosphorylation	1.475E-01	1.871E-01	1.660E-02	6.237E-03	2.153E-02	4.661E-01
GO:1901215~negative regulation of neuron death	3.441E-01	5.639E-01	8.481E-01	9.403E-03	5.779E-01	5.403E-01
GO:0016569~covalent chromatin modification	3.707E-01	9.035E-01	7.689E-01	5.815E-03	9.125E-01	6.691E-02
GO:0007346~regulation of mitotic cell cycle	6.216E-01	8.530E-01	8.481E-01	9.403E-03	8.602E-01	Not found
GO:0098609~cell-cell adhesion	2.569E-02	2.972E-02	5.948E-01	4.597E-02	1.335E-03	2.364E-01
GO:0006890~retrograde vesicle-mediated transport, Golgi to ER	4.219E-01	5.378E-01	8.962E-01	1.292E-02	1.793E-03	8.875E-01
GO:0006351~transcription, DNA-templated	1.627E-02	1.530E-01	1.600E-01	1.434E-02	6.168E-03	3.591E-01
GO:0003007~heart morphogenesis	8.184E-02	7.842E-01	7.785E-01	Not found	3.704E-03	1.711E-01
GO:0060271~cilium morphogenesis	4.096E-01	7.678E-01	1.744E-01	Not found	2.223E-03	8.465E-02
GO:0007018~microtubule-based movement	4.115E-01	5.280E-01	7.261E-01	Not found	1.621E-03	2.833E-02
GO:0010628~positive regulation of gene expression	7.121E-01	8.686E-01	4.277E-01	Not found	3.879E-03	8.349E-01
GO:0007029~endoplasmic reticulum organization	7.629E-01	3.620E-01	7.197E-01	Not found	1.483E-03	Not found
GO:0008104~protein localization	3.905E-01	2.316E-02	9.437E-01	Not found	8.368E-03	Not found
GO:0009791~post-embryonic development	7.373E-01	8.616E-01	6.586E-01	Not found	7.954E-03	Not found
GO:0007010~cytoskeleton organization	4.055E-02	3.896E-02	8.679E-01	Not found	4.605E-02	3.245E-04
GO:0050731~positive regulation of peptidyl-tyrosine phosphorylation	1.327E-01	1.847E-01	9.791E-01	Not found	7.595E-01	3.548E-03
GO:0048015~phosphatidylinositol-mediated signaling	4.733E-01	5.562E-01	8.711E-01	Not found	1.362E-01	8.965E-04
GO:0022604~regulation of cell morphogenesis	3.749E-01	3.275E-01	Not found	Not found	3.386E-01	4.474E-03
GO:0032869~cellular response to insulin stimulus	3.699E-01	4.880E-01	Not found	Not found	8.896E-01	7.324E-03
GO:0043523~regulation of neuron apoptotic process	2.587E-01	Not found	9.888E-03	Not found	6.070E-01	Not found
GO:0031929~TOR signaling	5.257E-01	Not found	Not found	Not found	3.568E-03	Not found
GO:0070562~regulation of vitamin D receptor signaling pathway	Not found	6.345E-03	Not found	Not found	Not found	Not found
GO:0006364~rRNA processing	Not found	7.738E-02	1.385E-04	Not found	9.783E-01	8.491E-01
GO:0046685~response to arsenic-containing substance	Not found	4.635E-01	Not found	1.928E-04	Not found	Not found
GO:0033689~negative regulation of osteoblast proliferation	Not found	Not found	Not found	4.573E-03	Not found	Not found

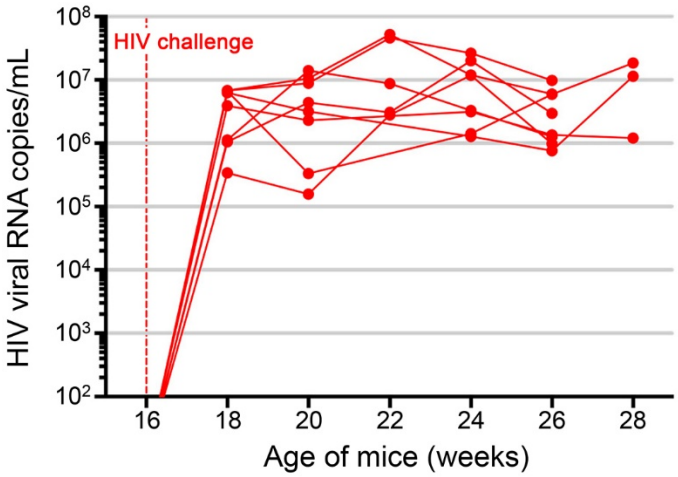
Biological Process	HIV in vivo	HIV in vitro - primary cells	HIV in vitro - cell lines	Lentivirus - hematopoietic	Lentivirus non-hematopoietic	Randomized IS library
GO:0052697~xenobiotic glucuronidation	Not found	Not found	Not found	Not found	2.326E-10	Not found
GO:1904224~negative regulation of glucuronosyltransferase activity	Not found	Not found	Not found	Not found	4.352E-09	Not found
GO:2001030~negative regulation of cellular glucuronidation	Not found	Not found	Not found	Not found	4.352E-09	Not found
GO:0045922~negative regulation of fatty acid metabolic process	Not found	Not found	Not found	Not found	1.877E-08	Not found
GO:0052695~cellular glucuronidation	Not found	Not found	Not found	Not found	4.434E-06	Not found
GO:0052696~flavonoid glucuronidation	Not found	Not found	Not found	Not found	4.728E-06	Not found
GO:0051552~flavone metabolic process	Not found	Not found	Not found	Not found	7.278E-05	Not found
GO:0009813~flavonoid biosynthetic process	Not found	Not found	Not found	Not found	2.588E-04	Not found
GO:0042573~retinoic acid metabolic process	Not found	Not found	Not found	Not found	3.484E-04	Not found
GO:0014898~cardiac muscle hypertrophy in response to stress	Not found	Not found	Not found	Not found	5.021E-04	Not found
GO:0030500~regulation of bone mineralization	Not found	Not found	Not found	Not found	1.271E-03	Not found
GO:0006207~'de novo' pyrimidine nucleobase biosynthetic process	Not found	Not found	Not found	Not found	1.969E-03	Not found
GO:1901018~positive regulation of potassium ion transmembrane transporter activity	Not found	Not found	Not found	Not found	1.969E-03	Not found
GO:0010882~regulation of cardiac muscle contraction by calcium ion signaling	Not found	Not found	Not found	Not found	3.322E-03	Not found
GO:0014808~release of sequestered calcium ion into cytosol by sarcoplasmic reticulum	Not found	Not found	Not found	Not found	Not found	2.734E-03
GO:0051775~response to redox state	Not found	Not found	Not found	Not found	Not found	8.475E-03

Overview of all biological processes identified in DAVID analysis. Individual biological processes listed in first column and the statistical significance (p-values) for each analyzed group are shown. Color code next to values indicates level of significance. Red boxes indicate a biological pathway that was very significantly enriched in the dataset (p-value < 0.005), blue boxes were significantly enriched (p-value < 0.01), gray boxes were observed but not significant (p-value ≥ 0.01) and white boxes indicate a pathway that was not found in the given dataset. Same representation as found in Figure 7.

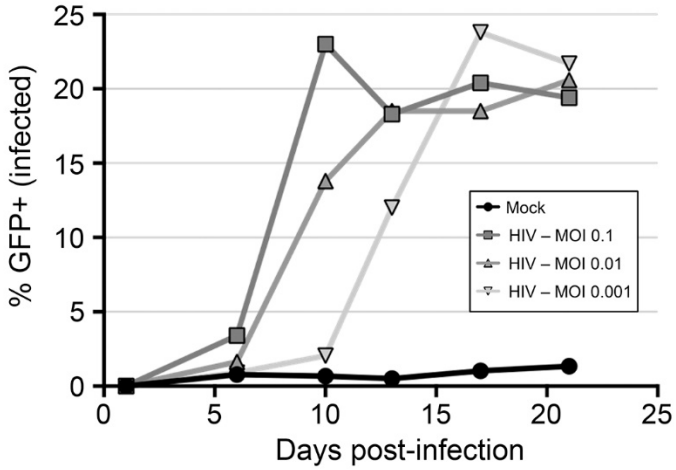
Supplemental Table 5. Primers used for IS analysis.

Primer Name	Primer Sequence	Region of Specificity
LCTLS (Linker 1)	CCTAACTGCTGTGCCACTGAATTCAGATC	
LCTU (Linker 2)	GACCCGGGAGATCTGAATTCAGTGGCACAGCAGTTAGG	
LC2 Primer (Forward 1st PCR)	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGATCTGAATTCAGTGGCACAG	GATCTGAATTCAGTGGCACAG
HIV3 Primer (Reverse 1st PCR)	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGTGACTCTGGTAACTAGAGATCCCTC	TGTGACTCTGGTAACTAGAGATCCCTC
Lenti LTR Primer (Reverse 1st PCR)	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGCTTGCCTTGAGTGCTTCAAGTAG	AGCTTGCCTTGAGTGCTTCAAGTAG
P5-5'-A/H (Forward 2nd PCR)	AATGATACGGCGACCACCGAGATCTACACXXXXXXXXTCGTCGGCAGCGTC	TCGTCGGCAGCGTC
P7-3'-1/12 (Reverse 2nd PCR)	CAAGCAGAAGACGGCATACGAGATXXXXXXXXGTCTCGTGGGCTCGG	GTCTCGTGGGCTCGG

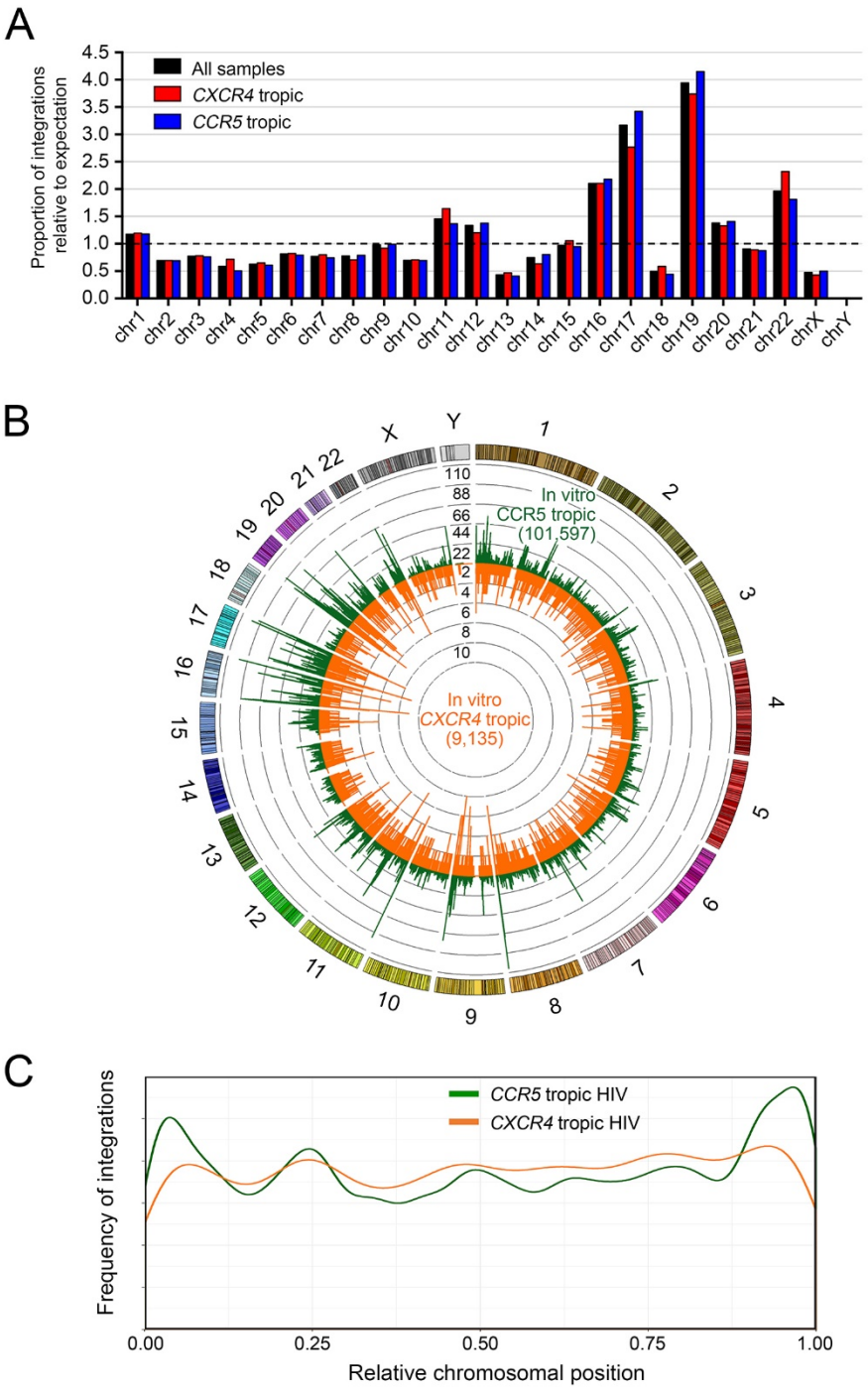
List of all primer sequences used for amplification of IS from either HIV infected or lentivirus transduced cell populations. LCTLS and LCTU primers are hybridized to form linker cassette for ligation. Region of specificity for each primer is also provided. Illumina barcode sequences used for multiplexing samples on sequencing runs are denoted as X's in 2nd PCR primers.



Supplemental Figure 1. Viremia of HIV infected mice. Viral load in plasma as measured in viral RNA copies per milliliter of blood. Lines represent individual mice tracked over time according to the age of mice on the x-axis. Vertical dashed red line indicates time of HIV challenge.

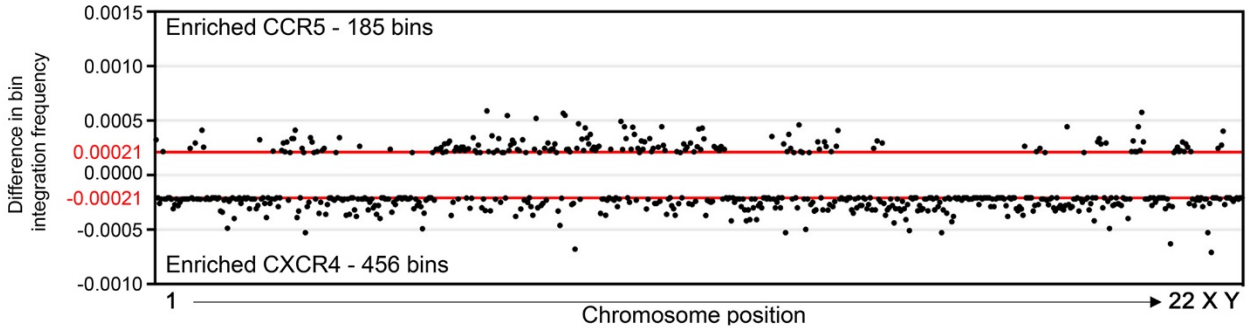


Supplemental Figure 2. HIV infection in Jurkat cell line. In vitro HIV infection of Jurkat cells challenged with various infectious units of HIV as compared to mock uninfected controls. Infection was tracked by flow cytometry for GFP expression under the control of an LTR promoter that is activated during active viral replication. Time post-infection is represented on x-axis in days and percent GFP expression in the cell population is plotted on the y-axis.

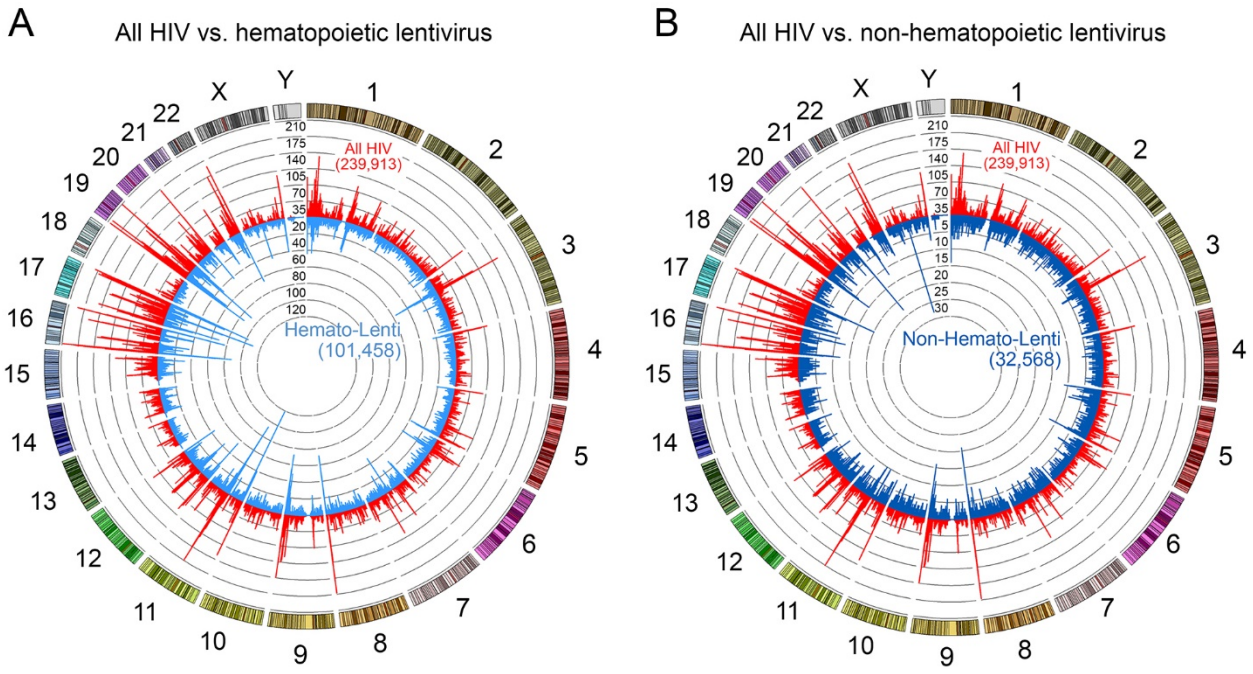


Supplemental Figure 3. Chromosomal integration sites in HIV strains with different tropisms. Either *CCR5* or *CXCR4* tropic HIV-1 viruses were analyzed for their chromosomal integration profiles. (**A**) All HIV integrations for each tropism were plotted for their frequency of integrations relative to unbiased expectation based on chromosome size, with larger chromosome expected to contain more unique integrations when compared to smaller ones. The y-axis is proportion of integrations observed relative to expectation based on chromosome size with horizontal dashed black line at a value of 1. (**B**) Circos plot comparing in vitro integrations of either a *CCR5* tropic strain (radiating outwards in green) or a *CXCR4* tropic strain (radiating inwards in orange) Each chromosome is represented on the exterior of each ring and is broken down into sequential bins, each 25kB in size. The total number of unique integrations occurring within each bin is

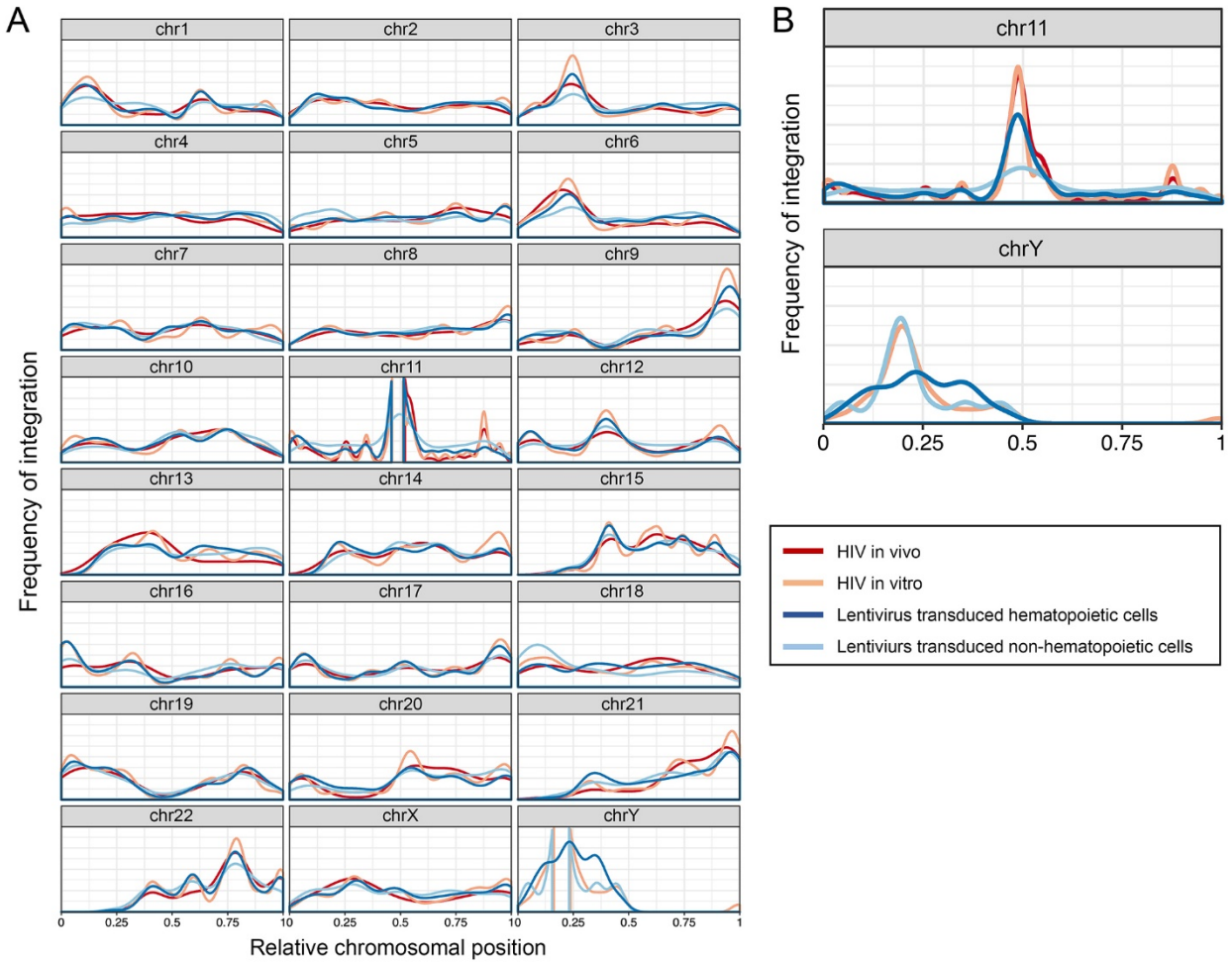
represented by the height of histogram bars. (**C**) Specific location of integrations relative to chromosomal length is plotted for in vitro Jurkat infections with either *CCR5* (green line) or *CXCR4* (orange line) viruses. The x-axis represents relative chromosomal position with 0.25, meaning 25% of the distance from beginning of chromosome and y-axis represents the frequency of observing an integration at each position throughout the chromosome.



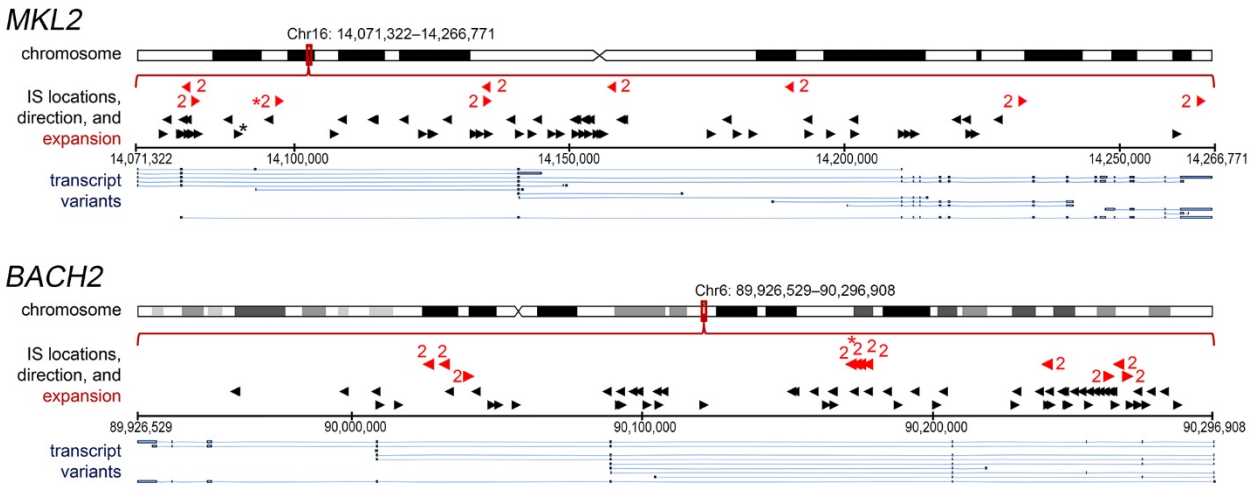
Supplemental Figure 4. Differential IS frequency between *CCR5* and *CXCR4* tropic viral strains. The in vitro *CCR5* dataset was directly compared to the *CXCR4* one in Jurkat cells to determine if there were specific locations within the genome that were significantly enriched for IS for these different tropisms. (**A)** Dot plot representing significantly enriched 25kB bin segments in one dataset compared to the other. Bins in the positive direction on y-axis were enriched for during in vitro Jurkat *CCR5* infection (185 bins, n=1), while bins in the negative direction were enriched for during in vitro Jurkat *CXCR4* infections (456 bins, n=6). X-axis represents relative chromosomal position, but is not normalized for chromosome length. Horizontal red lines indicate the 99th percentile of simulated data, similar to a p-value of 0.01.



Supplemental Figure 5. Circos plots comparing HIV profiles to lentiviral vectors. Each chromosome is represented on the exterior of each ring and is broken down into sequential bins, each 25kB in size. The total number of unique integrations occurring within each bin is represented by the height of histogram bars. All HIV ISs are represented as red histogram radiating outwards and compared to either (**A**) hematopoietic lentivirus transduced cells as light blue histogram or (**B**) non-hematopoietic lentivirus transduced cells as dark blue histogram; both of which radiate inward. Concentric rings function as a y-axis and have incremental values as outlined in each plot. Total number of IS within each dataset is represented as number in the parentheses.



Supplemental Figure 6. Positional integration frequency for each chromosome. (**A**) The location of integration relative to chromosomal length is plotted for each data set with HIV in vivo represented as a dark red line, HIV in vitro as a light red line, lentivirus transduced hematopoietic as a dark blue line, and lentivirus transduced non-hematopoietic as a light blue line. The x-axis represents relative chromosomal position, with 0.25 representing 25% of the distance from beginning of chromosome and y-axis represents the frequency of observing an integration at each position throughout the chromosome. (**B**) Expanded view for both chromosome 11 and Y.



Supplemental Figure 7. Identified IS in previously characterized genes. Representation of all unique HIV integration sites observed in two genes previously published as containing expanded clones or IS restricted to a very narrow window. Gene name is listed at top left for each group and specific location on chromosome is highlighted by red box. Gene transcript is expanded below chromosome, and each arrow indicates a unique IS, and arrow direction depicts orientation. Black arrows represent an IS found in one cell and red arrows represent IS found in 2 or more, with the number of clones detected noted next to each arrow. Blue lines at bottom of each graph represent all known transcript variants of each gene. Asterisks indicate integration sites that were found within in vivo mouse samples and are color coordinated as black or red to match arrow.