HHS Public Access

Author manuscript

Cell. Author manuscript; available in PMC 2016 January 29.

Published in final edited form as:

Cell. 2015 January 29; 160(3): 420–432. doi:10.1016/j.cell.2015.01.020.

HIV-1 integration landscape during latent and active infection

Lillian Cohn[1], **Israel T. Silva**[1,2], **Thiago Y. Oliveira**[1], **Rafael A. Rosales**[3], **Erica H. Parrish**[4], **Gerald H. Learn**[4], **Beatrice H. Hahn**[4], **Julie L. Czartoski**[5], **M. Juliana McElrath**[5], **Clara Lehmann**[6,7], **Florian Klein**[1], **Marina Caskey**[1], **Bruce D. Walker**[8,9], **Janet D. Siliciano**[10], **Robert F. Siliciano**[9,10], **Mila Jankovic**[1], and **Michel C. Nussenzweig**[1,9]

[1] Laboratory of Molecular Immunology, The Rockefeller University, New York, NY 10065, USA.

[2] National Institute of Science and Technology in Stem Cell and Cell Therapy and Center for Cell Based Therapy. Rua Catão Roxo, 2501, Ribeirão Preto, CEP 14051-140, SP, Brazil.

[3] Departamento de Computação e Matemática, Universidade de São Paulo. Av. Bandeirantes, 3900, Ribeirão Preto, CEP 14049-901.

[4] Departments of Medicine and Microbiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

[5] Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle WA, 98109, USA

[6] Department I of Internal Medicine, University Hospital of Cologne, Cologne, Germany.

[7] German Centre for Infection Research, partner site Bonn-Cologne, Cologne, Germany.

[8] Ragon Institute of MGH, Cambridge, MA 02139, USA

[9] Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA

[10] Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA.

SUMMARY

The barrier to curing HIV-1 is thought to reside primarily in $CD4^+$ T cells containing silent proviruses. To characterize these latently infected cells, we studied the integration profile of HIV-1 in viremic progressors, individuals receiving antiretroviral therapy, and viremic controllers. Clonally expanded T cells represented the majority of all integrations and increased during therapy. However, none of the 75 expanded T cell clones assayed contained intact virus. In contrast, the cells bearing single integration events decreased in frequency over time on therapy, and the surviving cells were enriched for HIV-1 integration in silent regions of the genome. Finally, there was a strong preference for integration into, or in close proximity to *Alu* repeats,

Correspondence: nussen@rockefeller.edu.

AUTHOR CONTRIBUTIONS

L.B.C planned and performed experiments, analyzed the data and wrote the manuscript. I.T.S. performed all bioinformatics analysis. T.Y.O. performed cancer gene analysis. R.A.R. performed bioinformatics and statistical analysis. E.H.P, G.H.L and B.H.H performed SGA and phylogenetic analysis. J.L.C., M.J.M, C.L., F.K., M.C., and B.D.W. selected and provided clinical samples. J.D.S and R.F.S performed QVOA and provided extensive discussion. M.J. planned and performed experiments, analyzed the data, and wrote the manuscript. M.C.N. planned experiments, analyzed the data and wrote the manuscript.

which were also enriched in local hotspots for integration. The data indicate that dividing clonally expanded T cells contain defective proviruses, and that the replication competent reservoir is primarily found in $CD4^+$ T cells that remain relatively quiescent.

INTRODUCTION

Despite effective therapy, HIV-1 can persist in a latent state as an integrated provirus in resting memory $CD4^+$ T cells (Chun et al., 1997; Finzi et al., 1997; Wong et al., 1997). The latent reservoir is established very early during infection, (Chun et al., 1998), and because of its long half-life of 44 months (Finzi et al., 1999) it is the major barrier to curing HIV-1 infection (Siliciano and Greene, 2011).

The HIV-1 latent reservoir has been difficult to define, in part because reactivation of latent viruses is difficult to induce and to measure. Viral outgrowth assays underestimate the size of the reservoir, while direct measurements of integrated HIV-1 DNA overestimate the reservoir because a large fraction of the integrated viruses are defective (Ho et al., 2013). Although the latent reservoir remains to be completely defined, establishing the reservoir requires intact retroviral integration into the genome and subsequent transcriptional silencing (Siliciano and Greene, 2011). Whether or not the genomic location of the integration impacts on latency is debated (Jordan et al., 2003; Jordan et al., 2001; Sherrill-Mix et al., 2013). However, HIV integration into the genome is known to favor the introns of expressed genes (Han et al., 2004), some of which, like *BACH2* and *MKL2* carry multiple independent HIV-1 integrations in different individuals and are considered hotspots for integration (Ikeda et al., 2007; Maldarelli et al., 2014; Wagner et al., 2014). However, there is currently no precise understanding of the nature of these hotspots or why they are targeted by HIV-1.

Viremia rebounds from the latent reservoir after interruption of long-term treatment with combination anti-retroviral therapy (cART). When it does, it appears to involve an increasing proportion of monotypic HIV-1 sequences, suggesting the proliferation of latently infected cells (Wagner et al., 2013). Based on this observation and the finding that a subset of cells bearing integrated HIV-1 undergoes clonal expansion in patients receiving suppressive anti-retroviral therapy, it has been proposed that the clonally expanded cells play a critical role in maintaining the reservoir (Maldarelli et al., 2014; Wagner et al., 2014).

To obtain additional insights into the regions of the genome that are favored by HIV-1 for integration and the role of clonal expansion in maintaining the reservoir, we developed a single cell method to identify a large number of HIV-1 integration sites from treated and untreated individuals, including “viremic controllers” who spontaneously maintain viral loads of <2000 RNA copies/ml and “typical progressors” who display viral loads >2000 RNA copies/ml.

RESULTS

Integration library construction

Twenty-four integration libraries were constructed from $CD4^+$ T cells from 13 individuals: 3 provided longitudinal samples before and after (0.1-7.2 years) initiation of therapy; 4 were untreated; 2 were treated; and 4 were viremic controllers (Table S1). Patients were grouped into three categories based on viral loads and therapy: 1. viremic progressors were untreated individuals with viral loads higher than 2000 viral RNA copies/mL of plasma; 2. progressors were treated individuals whose initial viral loads were higher than 2000 viral RNA copies/mL before therapy; 3. controllers were individuals who maintain low viral loads spontaneously in the absence of therapy (less than 2000 viral RNA copies/mL). The frequency of latently infected, resting $CD4^+$ T cells in our patients was similar to that reported by others as measured by quantitative viral outgrowth assay (Table S1 and (Laird et al., 2013)).

Libraries were produced from genomic DNA by a modification of the translocation-capture sequencing method that we refer to in this paper as integration sequencing (Figure 1A) (Janovitz et al., 2013; Klein et al., 2011). Virus integration sites were recovered by semi-nested ligation-mediated PCR from fragmented DNA using primers specific to the HIV-1 3' LTR (Table S2). PCR products were subjected to high-throughput paired-end sequencing, and reads were aligned to the human genome. Since sonication is random, it produces unique linker ligation points that identify the specific integration events in each infected $CD4^+$ T cell, which allows both single cell resolution and identification of expanded clones of cells with identical integrations ((Berry et al., 2012) and Figure 1A). Thus, integration sequencing can enumerate both the number of integration sites and the number of infected cells.

A total of 6719 unique virus integration sites were determined (Table S3): 873 unique integrations in viremic controllers; 987 integrations in untreated progressors; and 4859 integrations in treated progressors (Figure 1B).

Integrations are enriched in introns of highly expressed genes

We analyzed the genomic location of the integration sites obtained from viremic controllers, untreated and treated progressors and compared our results to published data obtained from HIV-1 infected individuals (Han et al., 2004; Ho et al., 2013; Ikeda et al., 2007; Schroder et al., 2002). In agreement with the work of others, the majority of integration sites in each group are genic (Figure 1C and Figure S1A). Moreover, integrations are found more frequently in the introns of highly expressed genes, and there is a slight bias for viral orientation that leads to convergent transcription (Figures 1D, E and F and Figure S1B-D) (Mitchell et al., 2004; Schroder et al., 2002). Thus, the general features of integrations defined by integration sequencing are similar to those obtained by others.

Although the differences between groups were small in magnitude, they were significant in that treated progressors had a smaller proportion of integrations in genic regions ($p<0.0001$ and $p<0.0001$, respectively) and in highly expressed genes ($p<0.0001$ and $p<0.0001$, respectively) when compared to viremic controllers and untreated progressors (Figure 1C, E

and Figure S1C). Conversely, the proportion of viral integrations in genes expressed at lower levels was increased in treated progressors compared to viremic controllers and untreated progressors ($p= 0.002$ and $p<0.0001$, respectively). Viremic controllers and treated progressors were not significantly different from each other in terms of the level of expression of the genes at the sites of integration (Figure 1E). Thus, therapy is associated with a relative decrease in the number of cells with viral integrations in highly expressed genes.

Identification of clonally expanded cells containing integrated HIV-1

Since we shear DNA ends randomly to produce our libraries, and by paired end sequencing can determine the precise site of both the integration and sheared end, we infer that identical integrations with unique sheared ends arise from clones of expanded cells (Figure 1A). Integrations can therefore be classified as clonally expanded (i.e. identical integrations with distinct sheared ends, deriving from the clonal expansion of an original unique, single integration event) or single integrations (i.e. unique integration site with a single sheared end).

Clonally expanded viral integrations were present in all individuals irrespective of therapy or viremia (Table S3). However, the proportion of clonally expanded viral integrations is significantly lower in viremic controllers (30%) and viremic progressors (27%) than in treated progressors (40%) ($p<0.0001$ and $p<0.0001$, Figure 2A and Figure S2A). Although the size of individual clones varied from 2-295 cells (Figure S2B), the relative increase in clonally expanded integrations during therapy consistently translated into an increase in the number of infected cells that derive from expanded clones (Figure 2B). The percentage of cells containing clonally expanded HIV-1 integrations was similar in untreated progressors (78%) and controllers (79%), but it was significantly increased in treated progressors (90%) ($p<0.0001$ and $p<0.0001$, Figure 2B and Figure S2C). Thus, therapy is associated with an increase in the frequency of clonal HIV-1 integrations and infected clonally expanded cells.

To determine whether the position of viral integration in the genome correlates with clonal expansion we compared the genomic clonally expanded to single integrations. Both types of integrations favored genes and their introns (Figure 2C, D and Figure S2D, E). However, the proportion of clonally expanded integrations in intergenic regions was greater than that of single integrations (Figure 2C, $p<0.0001$). Moreover, of the integrations in genes, single integrations were more likely to be found in highly expressed genes than clonal integrations (Figure 2E, $p<0.0001$ and Figure S2F). Thus, cells harboring viral integrations in intergenic regions and genes that are expressed at lower levels are more likely to be clonally expanded.

Large expanded clones are found in memory cells

Central memory cells are thought to be the major source of the HIV-1 reservoir (Chomont et al., 2009). To investigate the nature of the cells that comprise the expanded clones, we performed virus integration sequencing on genomic DNA from sorted central, transitional and effector memory $CD4^+$ T cells (Figure 2F). In both individuals studied, all three subsets of $CD4^+$ T cells contained expanded clones (Figure 2G, H and Figure S2G,H). Thus, central,

transitional and effector memory T cells, all of which have undergone antigen-stimulated cell division, harbor the expanded clones of HIV-1 integrants.

Clonally expanded integrations increase after therapy

The proportion of clonally expanded viral integrations is increased in treated progressors (Figure 2A and (Wagner et al., 2014)). To further examine the effect of therapy on clonal expansion we analyzed longitudinal samples from three typical progressors before and during therapy (Table S1). We found an increase in the number of clonally expanded integrations throughout the treatment period of up to 7.2 years in two of the three patients (Figure 3A, p=0.017 and Figure S3A) as well as an increase in the number of cells that contained clonally expanded viral integrations (Figure S3B). Correspondingly, there was also an overall decrease over time in single integrations (p=0.017), with a half-life of 127 months assuming a non-linear regression model for one-phase decay (Figure 3B). Thus, our data suggests that the numbers of single integrations decay very slowly over time, while clonally expanded integrations increase with time on cART.

The increase in the number of clonal integrations during cART did not favor genic or intergenic regions (p=0.65), indicating that this effect is independent of the location of the integration in the genome (Figure 3C and E, Figure S3C). In contrast, single integrations decrease significantly in genic regions (Figure 3D, p=0.036 and Figure S3D) and increase proportionally in intergenic regions (Figure 3F, p=0.036). Thus the fate of cells harboring single viral integrations in cART treated progressors differs from clonal integration. Moreover, the fate of single integrations is dependent on their location in the genome whereas the clonal integrations are not. These results suggest that cells bearing genic single integrations are selected against during therapy and that clonal expansion is not.

Clonally expanded integrations in the same genes in multiple patients

In the 3 progressors who provided longitudinal samples, approximately 5% of the clonal integrations persisted through successive time points without selection for genic or intergenic regions compared to all clonal integrations (Figure 4A and B). Furthermore, of the genic integrations that persisted, there was also no selection for or against those in highly expressed genes (Figure 4C). Thus, the persistent clonal integrations are indistinguishable from the larger pool of clonally expanded viral integrations in terms of their position in the genome.

To determine whether specific genes or groups of genes are permissive for clonal expansion we looked for overlap in genic integration sites between samples (Figure 4D-F). Despite a higher number of single integrations, there was much greater overlap of the genes that harbor clonally expanded integrations between individuals irrespective of treatment or level of viremic control (p<0.0001) (Figure 4D-F). On average, there is 13% and 3% overlap between genes harboring clonally expanded and single viral integrations, respectively. The genes containing clonally expanded viral integrations in multiple patients are expressed at lower levels than genes containing overlapping single viral integrations (Figure 4G). Taken together, these results suggest that cells that carry integrations in highly expressed genes

tolerate clonal expansion less well than cells with integrations in genes with lower levels of expression.

Since clonal integrations have been associated with genes involved in malignant transformation (Wagner et al., 2014), we examined our entire data set for enrichment of integrations in cancer-associated genes (n = 743 cancer associated genes (Vogelstein et al., 2013; Zhao et al., 2013)). Although there was an overall enrichment among for integrations in cancer genes (329/4410 = 7.5%) compared to all genes in the human genome (743/25,660 =2.8%) ($p<0.0001$), this preference does not seem to be significant because it is similar to the overall preference for integration into highly expressed genes (Figure S4). Furthermore, we observed no overrepresentation of single, clonal or persistent integrations in cancer genes (Figure 4H). Importantly, a significant decrease in integrations in cancer related genes was observed in longitudinal samples (Figure 4I) suggesting that these are selected against with therapy.

Expanded clones contain defective viruses

Our method of integration sequencing captures the end of the 3' LTR and identifies the genomic site of viral integration. To determine whether the viruses found in expanded clones are intact, we used nested integration site-specific PCR primers that were anchored in the host genome to amplify the 5'LTRs of 75 expanded clones from 8 individuals (Table S2). The clones selected for PCR verification varied in size from 5-200 out of $0.3\text{-}2\times10^6$ CD4 T cells. Of the 75 sequences obtained, 24 showed fragmented 5'LTRs flanked by the correct genomic site, and an additional 44 of the proviruses did not have a recoverable 5' end (Figure 5B). The remaining 8 viruses with intact 5' LTRs were amplified in limiting dilution conditions using integration site-specific primers and HIV-1 primers (Figure 5C). Three of the 8 viruses could not be amplified; 4 had large deletions in Env, 1 had a frameshift mutation in *pol* and 1 had undergone APOBEC3G mediated hypermutation to produce a premature stop codon in *env* (Figure 5D and Data S1). Thus, we were unable to find a single intact integrated provirus among 75 expanded clones.

Hotspots for virus integration

Overlap between integrations in the genes of different patients suggests the existence of hotspots for HIV-1 integration. A number of individual genes have been identified as preferential sites for HIV-1 integration including *BACH2, MKL2, DMNT1, MDC1 and STAT5B* (Ikeda et al., 2007; Maldarelli et al., 2014; Wagner et al., 2014). To identify hotspots for HIV-1 integration genome-wide, we subjected our data set to hot_scan analysis (Silva et al., 2014), which defines hotspots by identifying regions of local enrichment using scan statistics. This analysis identified 55, 85, and 247 hotspots for controllers, viremic and treated progressors, respectively (Figure 6A). For example, the intron between exons 5 and 6 in *MKL2* is a hotspot for integration in patient 11, contains an expanded clonal family in patient 10 and was also identified as a site of enrichment for integration by others (Maldarelli et al., 2014) (Figure 6B).

To validate our *in silico* analysis and to further characterize the *MKL2* hotspot, we sequenced the *gag* gene from proviruses integrated into *MKL2* by amplification with nested

genomic primers specific for *MKL2* and HIV-1 *gag*. Sequences obtained from patient 10, who showed only one expanded clone are very closely related to each other, which is consistent with a single clonally expanded integration (Figure 6C). In contrast, sequences obtained from patient 11 are far more diverse suggesting that there were several different viral integrations in the *MKL2* hotspot (Figure 6C). We conclude that the hotspots defined by hot_scan represent multiple distinct integration events in close proximity.

Viremic progressors had the highest proportion of integration events in hotspots, indicating that in the case of high-level viremia there are specific genomic locations that favor integration (Figure 6D). Although the majority of all integrations fall outside of hotspots (Figure 6D), hotspots resemble other integrations in that they are preferentially found within genes with a preponderance of these in introns (Figure 6E and F). In all cases hotspots are enriched in highly expressed genes, and consistent with the overall decrease in viral integrations in highly expressed genes during therapy, the proportion of hotspots in these genes also decreases (Figure 6G and 1E). Thus, the general characteristics of hotspots are similar to features of all integrations.

To determine whether there is a relationship between hotspots and clonally expanded viral integrations we enumerated single and clonally expanded integrations in hotspots (Figure 6H). Only a small fraction (11-18%) of all single integrations were found in hotspots with untreated viremic progressors showing the highest level (Figure 6H). In contrast, there was a much higher proportion of clonal integrations in hotspots (30-46%) with the lowest proportion in treated progressors (Figure 6H). This observation is consistent with the finding that there is a greater degree of overlap in genes that harbor clonally expanded than single integrations (Figure 4D-F) and that clonally expanded integrations are more likely to be hotspots than single integrations (Figure 6H, $p<0.0001$).

Viral integration enriched in sites containing a DNA sequence motif

To determine whether there are specific genomic features associated with sites of viral integration and hotspots, we examined 100 base pairs (bp) centered on all integration sites for the presence of a consensus sequence (Bailey and Elkan, 1994). We found 7% of all integrations within 100bp of a highly conserved 30bp motif (INT-motif) (Figure 7A). The majority of the integrations identified in this analysis were single integration events with the ratio of single to clonal integrations being significantly different from the expected (Figure 7B, $p<0.0001$). When HIV-1 integrates directly into the INT-motif, the 5’ end of the motif is recurrently found 20bps from the site of viral integration (Figure 7C). The INT-motif is asymmetrically distributed in *Alu* repeats and its position coincides with a peak of viral integration (Figure 7D). Furthermore, there is a significant overall enrichment of integrations inside *Alu* repeats (Figure 7E), and in close proximity to *Alu* repeats, irrespective of whether the integration is inside genes or in intergenic regions (Figure 7F). Thus a preference for *Alu* is independent of a preference for integration in genes.

Previous studies have shown a preference for integration into *Alu* repeats, potentially because *Alu* repeats are enriched in gene-rich regions (Schroder et al., 2002). To further examine the relationship between *Alu* repeats and transcription of genes, we determined the distance between *Alu* repeats and the center of all genes. There was no positive correlation

between the position of *Alu* and the level of transcription (Figure 7G). To determine whether the distance between integration and *Alu* repeats correlates with transcription, we measured the distance between the sites of integration and *Alu* repeats in all genes (Figure 7H). There was no significant difference between integration distance to *Alu* repeats in highly expressed, silent or trace level expressed genes. Therefore the rate of transcription does not impact integration distance to *Alu* repeats and integration at these sites must be independent of transcription.

Finally, the number of *Alu* repeats in a hotspot is directly correlated with the number of integration events in that hotspot (Fig. 7I, χ=0.86). We conclude that HIV-1 has a preference for integration in close proximity to sites in the genome that are enriched in *Alu* repeats and that this preference is independent of the level of transcription.

DISCUSSION

$CD4^+$ T cells that are actively infected with HIV-1 are rapidly eliminated during anti-retroviral therapy, but this form of treatment is relatively ineffective in selecting against latently infected $CD4^+$ T cells, which have an estimated half-life of 44 months. Abolishing the latent reservoir is the current hurdle to finding a cure for HIV-1 infection. Although we have learned a great deal about the location of the latent compartment and its persistence during therapy, it has been difficult to uncover whether there are specific genomic features associated with latency (Siliciano and Greene, 2011). One of the major impediments to understanding latency is our inability to purify cells harboring latent HIV-1 as opposed to cells containing defective viruses. To further investigate the latent compartment, we used a high throughput method that uncovers sites of HIV-1 integration while enumerating clones of expanded T cells that bear identical integrations.

By comparing HIV-1 integration in controllers, untreated and treated progressors, including longitudinal samples obtained before and after therapy, we found that proliferating clones of infected cells accumulate over time. However, we were unable to detect intact, full-length viral sequences in these clones. Instead, our evidence suggests that the reservoir resides primarily in cells bearing unique integrations that are selected against by cART in an integration specific manner, favoring the persistence of integrations in intergenic regions and silent genes, with decay kinetics that argue against homeostatic proliferation.

A number of different investigators have shown that HIV-1 prefers to integrate into the introns of highly expressed genes (Craigie and Bushman, 2012). This is true for all of the individuals in our study irrespective of their status as controllers or treatment with cART. Although the level of intrinsic viremic control has no detectable effect on integration site selection, therapy selects against genic integrations and more specifically, against integrations in highly expressed genes, when compared to untreated progressors or controllers. Given that cART selects for cells that bear silent proviruses, the results suggest that viruses integrated into genes are less likely to become latent than those found in intergenic regions. Moreover, the data indicate that among the proviruses integrated into genes, those that are found in genes expressed at low levels are also more likely to become latent. These findings are entirely consistent with *in vitro* experiments in cell lines showing

that level of HIV-1 transcription is dependent in part on the status of surrounding chromatin (Jordan et al., 2003; Jordan et al., 2001; Sherrill-Mix et al., 2013).

HIV-1 integration has been studied in multiple cell types, but large libraries of integrations sites in primary infected T cell have only recently become available (Maldarelli et al., 2014; Wagner et al., 2014). Integration sites obtained from in vitro infected cell lines and primary T cells are distinct (Brady et al., 2009; Sherrill-Mix et al., 2013). Nevertheless common features of HIV-1 integration have been defined including the observation that integration favors *Alu* repeats (Schroder et al., 2002). This association was thought to be dependent on the presence of these repeats in the introns of gene-rich regions and not on a particular sequence feature (Schroder et al., 2002). However, we observed that integration preference into highly transcribed genes and into *Alu* repeats seem to be independently important and furthermore integrations are enriched near *Alu* repeats both in genic and intergenic regions. One possible explanation for preference for *Alu* seems to be the presence of an INT-motif. TG-$(N)_{5-7}$-CA sequence has been associated with sites of HIV-1 integration, but an integration consensus has not been defined (Brady et al., 2009; Holman and Coffin, 2005; Lewinski et al., 2006; Serrao et al., 2014; Wang et al., 2007; Wu et al., 2005). We found a 30bp INT-motif within 100bp of 7% of all integrations, the large majority of which are single events. As expected, the HIV-1 INT-motif contains a signature TG-$(N)_{5-7}$-CA and can form a hairpin structure, anchored on 5'-NTG-3', 5'-CAN-3'. This motif is frequently found at the 3' end of *Alu* where it coincides with a peak of viral integration events and viruses integrated directly in this motif show a dramatic specificity for insertion site. The asymmetric peak and specificity of the integration site are remarkable. Nevertheless, we are likely underestimating the frequency of integrations within *Alu* due because we can only map unique reads.

The observation that HIV-1 prefers to integrate in the neighborhood of *Alu* repeats is consistent with the finding that different individuals have been reported to have multiple integrations in selected genes (Ikeda et al., 2007; Maldarelli et al., 2014; Schroder et al., 2002; Wagner et al., 2014). Our experiments define a group of overlapping hotspots for integration that share many of the features of all HIV-1 integrations including preference for introns of highly expressed genes and high density of *Alu* repeats. Viremic progressors showed the highest levels of hotspot integration, possibly because persistent integration leads to over-representation of these favored sites. Alternatively, these integrations might be positively selected by mechanisms that remain to be determined.

Individuals receiving cART show increasing numbers of cells with identical viral genomes by SGA suggesting clonal expansion of a subset of cells bearing integrated proviruses (Buzon et al., 2014; Chomont et al., 2009; Wagner et al., 2013). Two independent groups have recently documented the long-term persistence of expanded clones of cells during therapy with cART (Maldarelli et al., 2014; Wagner et al., 2014). Our analysis confirms and extends these observations by showing that when considered as a group, expanded clones are less likely to occur when the provirus is in a genic region, and clones that are associated with genes tend to be in genes that are expressed at lower levels than single integrations. Thus, proviruses inserted into active regions of the genome, which would be more likely to

support viral re-activation during T cell proliferation, are generally selected against during clonal expansion.

Why certain integration sites are permissive for clonal expansion is not known, but finding that expanded clones with integrations occur in cancer related genes led to the suggestion that integration into genes that regulate cell division promotes proliferation (Wagner et al., 2014). While we also found a higher proportion of integrations in cancer-related genes as compared to random, this bias was not different from that observed for other highly expressed genes favored by HIV-1. Further, we do not see any differential bias for integration in cancer related genes in clonally expanded cells compared to single integrations and an overall decrease in the number of integrations in cancer related genes during the course of therapy. Since the number and size of clones increases with time on therapy, the data indicate that integration into cancer genes is unlikely to be a general contributor to the proliferation of infected cells.

Our data show that cART selects for expanded clones and that viremic controllers resemble treated progressors in showing a higher proportion of expanded clones than untreated viremics. cART selects for clonal integrations irrespective of the location in the genome. This is in contrast to single integrations, which are selected against by therapy. cART specifically favors the survival of single integrations in intergenic regions and is biased against genic regions with an overall half-life for single integrations of 127 months. The half-life of single integrations is not too dissimilar from the current estimate for the latent reservoir, which is believed to decay with a half-life of 44 months on cART (Finzi et al., 1999).

The major outstanding question after the discovery of clonally expanded cells with integrated HIV-1 is whether the virus from these cells contribute to the latent reservoir (Maldarelli et al., 2014; Wagner et al., 2014). Several different independent lines of evidence argue against this idea. First, although the latent reservoir is thought to be contained primarily in resting central memory $CD4^+$ T cells (Siliciano and Greene, 2011), we find that clonally expanded viral integrations are found in all three memory T cell compartments. Second, whereas the reservoir appears to decay with time on cART, we find that clonally expanded integrations increase with time and do so irrespective of whether they are found in genes or intergenic regions. In contrast, single integrations in more active parts of the genome, which are more likely to support HIV-1 reactivation, are selected against with time on ART. Finally, all 75 of the clonally expanded proviruses tested were defective, which is in agreement with 2 examples in the literature (Imamichi et al., 2014; Josefsson et al., 2013). Thus, we conclude intact virus is not enriched in infected expanded cells. However, we cannot rule out the possibility that a rare clone of cells contains an active virus. Nevertheless, the 90% of all cells bearing integrated proviruses that account for expanded clones of infected cells in cART treated progressors appear to be unlikely to be the major source of the rebounding latent reservoir. Instead, the replication competent reservoir is likely to be contained in the remaining 10% of cells that harbor single integrations that decline with a long half-life on cART (Figure S5).

In conclusion, the data indicate that HIV-1 infected T cells that undergo clonal expansion are able to do so because their proviruses are defective and that the replication competent reservoir is found in the subset of $CD4^+$ T cells that remain quiescent.

EXPERIMENTAL PROCEDURES

$CD4^+$ T cell isolation for Integration Library construction

Human samples were collected after signed informed consent in accordance with Institutional Review Board (IRB)-reviewed protocols by all participating institutions. Patients 1, 2, and 3 were selected from the Seattle HIV longitudinal cohort studies at Fred Hutchinson Cancer Research Center. Patients 4, 8 and 9 were recruited from the University of Cologne and samples were obtained at Rockefeller University (MNU_0628). Patients 5, 6 and 7 were selected from the Rockefeller University HIV-1 antibody therapy clinical trial. Patients 10, 11, 12, and 13 were selected from a group of elite controllers that were followed at the Ragon Institute in Boston.

$CD4^+$ T cells were isolated from whole PBMC using anti-CD4 microbeads (Miltenyi Biotec). The percentage of live cells was determined by flow cytometry based on forward and side scatter. Purity of $CD4^+$ T cells was determined by labeling isolated cells with anti-human CD3, CD4, CD8, CD19 and HLA-DR and gating on CD3, CD4 double positive cells. Isolated cells were used for library construction only if purity was >75%. $CD4^+$ T cell subsets were isolated by FACSorting on a BD Aria II by labeling cells with anti-human CD3, CD4, CD8, CD66b, CD335, HLA-DR, CCR7, CD27 and CD45RA. Analysis of $CD4^+$ T cell subsets was done by pooling cellular DNA isolated from multiple sorts of the same sample.

Quantitative viral outgrowth assay

Viral outgrowth was performed as previously described. (Laird et al., 2013)

Integration Library

The method for integration library construction was adapted from TC-Seq (Klein et al., 2011).

DNA preparation—DNA from 0.2-2 million $CD4^+$ cells from HIV-1 infected patients was isolated and prepared as previously described (Klein et al., 2011). Fragments were ligated to 200pmol of annealed linkers (Table S2). Virus sequences were eliminated by digestion with BglII (NEB) and fragments were purified.

Integration site amplification—Semi-nested ligation-mediated PCR was performed on DNA. All PCRs were performed using Phusion Polymerase (Thermo). DNA was divided into 700ng aliquots and subjected to single-primer PCR with biotinylated LTR1 [1x(98C-1min) 12x(98C-15s, 62C-30s, 72C-30s) 1x(72C-5min)] (Table S2). Each reaction was spiked with pLinker and subjected to additional cycles of PCR [1x(98C-1min) 25x(98C-15s, 62C-30s, 72C-30s) 1x(72C-5min)]. Products of 300-1000bp were isolated by agarose gel electrophoresis and magnetic streptavidin bead purification. Semi-nested PCR

was performed on the magnetic beads first with a single primer LTR2 (same cycling conditions as above) followed by spiking in pLinker and additional cycles (Table S2). Products of 300-1000bp were isolated by gel electrophoresis.

Paired-end library preparation—Linkers were digested by AscI such that a 6-nucleotide barcode (CGCGCC) was left on the DNA fragments, indicating linker-dependent amplification. Fragments were blunted by End-It DNA Repair Kit (Epicenter), purified with AmPure beads (Agencourt) and ligated to NextFlex paired-end adapters. Adaptor-ligated fragments were enriched by 35 cycles of PCR with NextFlex primers [1x(98C-1min) 35x(98C-15s, 66C-30s, 72C-30s) 1x(72C-5min)] and fragments between 300-1000bp were isolated by gel electrophoresis. Two or three libraries were mixed in equimolar ratios and sequenced by either 150bp paired-end sequencing on Illumina MiSeq or 150bp or 100bp paired-end sequencing on an Illumina 2500 HiSeq. Data is accessible via NCBI SRA using the accession number: SRP045822.

Computational Analysis

Read Alignment—Paired-end reads were mapped to the HIV-1 sequence (designated as a bait) using BLAT (Kent, 2002) with default settings. Reads that were mapped to the bait without mismatches were checked for the linker barcode in the paired-end read, and was mapped to the human genome reference GRCh37/hg19) with *Bowtie* (Langmead et al., 2009). Only uniquely mapped reads (allowing for to 2 mismatches) were used as defined in the best alignment stratum (command line options: **-v2 -all -best -strata -m1**). Identical reads generated by PCR amplification were merged.

Integration determination—Once the paired-end reads were properly mapped in the bait and human genome (see above), we determined the integration breakpoint by aligning the remaining nucleotide sequence containing the 3' terminus of the HIV-1 LTR to the human genome using BLAT (default settings). Only uniquely mapped reads up to 1Kb away from its partner were kept. Adjacent (within 50 nucleotides) putative integrations sites were merged. Finally, the 5' end of the paired end reads were used to deduce the integration and shear position sites in the human genome.

Hotspot detection—To detect preferred sites of HIV-1 integration genome-wide, we subjected our dataset to *hot_scan* software analysis (Silva et al., 2014), which defines hotspots by scan statistics. Hotspots obtained by *hot_scan* were defined using different window widths (100, 200, 500, 1000, 2000, 5000, 10000, 20000, 50000 and 100000 bp).

Motif analysis—To determine a consensus motif, 100bp flanking each integration site was analyzed for the presence of 30bp consensus sequence using MEME software (Bailey and Elkan, 1994).

Monte Carlo Simulation for virus integration and hotspots—Monte Carlo simulation was conducted by shuffling the genomic locations of all virus integration sites 10000 times using *bedtools shuffle* utility (Quinlan and Hall, 2010). Then, we compared the observed number with the median number of in the randomized list. We assessed

enrichments by *P-value* by counting the frequency of observed events being equal or higher than the number of randomized events divided by N=10000.

Statistical analysis—Proportion test is the standard test for the difference between proportions, also known as a two-proportion z-test. We used R's implementation of this via the prop.test() function.

Integration library verification

To verify our integration sequencing strategy, we constructed two libraries from DNA isolated from un-infected individuals. We recovered 13 sequences that mapped to integration sites. We subtracted these "integration sites" from all libraries before proceeding.

To test the saturation of our method, two separate integration libraries were constructed from identical samples for three patients. We found that both libraries contained the same expanded clonal families, but the majority of single virus integrations were unique to each sample of cells used for library construction. Single viral integrations found in both libraries were less than 1% of observed viral integrations.

PCR verification—Genomic DNA isolated as described was serially diluted and subjected to nested-PCR using genomic specific primers and primers LTR1 and LTR2 (Table S2) using HotStart Taq Polymerase (Qiagen) [1x(98C-14min) 40x(98C-30s, 55C-30s, 72C-30s) 1x(72C-5min)]. Products were isolated by gel electrophoresis and sequenced directly. Analysis of clones in this manner identified that we underestimate the size of clones by 4-5 times (data not shown).

$CD4^+$ T cell subset sorting

To isolate $CD4^+$ subsets, we labeled PBMCs with antibodies to CD45RA, CD4, CD8 CD66b, CCR7, CD335, HLA-DR, CD3, and CD27. We separated T cell subsets by FACS Aria (BD Biosciences) to very high purity (>98%).

Virus sequencing

5'LTR—5' LTRs from large clones were amplified with nested genomic primers and LTR2Rev (Table S2) using Platinum High Fidelity Taq (Invitrogen) [1x(98C-14min) 40x(98C-30s, 55C-30s, 68C-1min) 1x(68C-5min)]. Products were isolated by gel electrophoresis and sequenced directly.

Full Length Virus—Full length genomic DNA from infected patients was isolated as described and serially diluted. Each well was filled to a final volume of 50μL with PCR reaction mixture (Platinum Taq MasterMix, Invitrogen) and primers to amplify virus from a specific integration site in the genome (Table S2 and (Ho et al., 2013)) using touchdown cycling to increase specificity. Then, 2μL aliquots from the first PCR were subjected to nested genomic PCR and 1% gel electrophoresis. The positive wells were gel-purified and fragments were sequenced directly.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

We would like to acknowledge all patients who contributed to this study. We thank Dr. Qiao Wang for invaluable discussions; Dr. Johannes Scheid for the coordination and preparation of viremic controller samples; Joshua Horwitz for samples during the initial development of integration sequencing; Klara Velinzon and Gaelle Breton for FACSorting; Zoran Jankovic for laboratory support; New York Genome Center for sequencing; David Chambliss for assistance in obtaining human samples; Arlene Hurley and Gisela Kremer for assistance in patient coordination; all members of Nussenzweig and Mucida labs for valuable discussion and advice. This work was supported in part by the Bill and Melinda Gates Foundation Collaboration for AIDS Vaccine Discovery Grants 1040753 and 38619 (to M.C.N.). This work was also supported in part by grant #UL1 TR000043 from the National Center for Advancing Translational Sciences (NCATS), AI 100663-01 and CHAVI-ID Award UM1AI100663, Center for HIV/AIDS Vaccine Immunology and Immunogen Discovery, AI 100148-01. E.H.P, G.H.L and B.H.H. are supported by UM1 AI100645 and R37 AI 066998. M.C.N., R.F.S. and B.D.W. are Howard Hughes Medical Institute Investigators.

REFERENCES

Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings / International Conference on Intelligent Systems for Molecular Biology ; ISMB International Conference on Intelligent Systems for Molecular Biology. 1994; 2:28–36.

Berry CC, Gillet NA, Melamed A, Gormley N, Bangham CR, Bushman FD. Estimating abundances of retroviral insertion sites from DNA fragment length data. Bioinformatics. 2012; 28:755–762. [PubMed: 22238265]

Brady T, Agosto LM, Malani N, Berry CC, O'Doherty U, Bushman F. HIV integration site distributions in resting and activated CD4+ T cells infected in culture. Aids. 2009; 23:1461–1471. [PubMed: 19550285]

Buzon MJ, Sun H, Li C, Shaw A, Seiss K, Ouyang Z, Martin-Gayo E, Leng J, Henrich TJ, Li JZ, et al. HIV-1 persistence in CD4+ T cells with stem cell-like properties. Nature medicine. 2014; 20:139–142.

Chomont N, El-Far M, Ancuta P, Trautmann L, Procopio FA, Yassine-Diab B, Boucher G, Boulassel MR, Ghattas G, Brenchley JM, et al. HIV reservoir size and persistence are driven by T cell survival and homeostatic proliferation. Nature medicine. 2009; 15:893–900.

Chun TW, Carruth L, Finzi D, Shen X, DiGiuseppe JA, Taylor H, Hermankova M, Chadwick K, Margolick J, Quinn TC, et al. Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. Nature. 1997; 387:183–188. [PubMed: 9144289]

Chun TW, Engel D, Berrey MM, Shea T, Corey L, Fauci AS. Early establishment of a pool of latently infected, resting CD4(+) T cells during primary HIV-1 infection. Proceedings of the National Academy of Sciences of the United States of America. 1998; 95:8869–8873. [PubMed: 9671771]

Craigie R, Bushman FD. HIV DNA integration. Cold Spring Harbor perspectives in medicine. 2012; 2:a006890. [PubMed: 22762018]

Finzi D, Blankson J, Siliciano JD, Margolick JB, Chadwick K, Pierson T, Smith K, Lisziewicz J, Lori F, Flexner C, et al. Latent infection of CD4+ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. Nature medicine. 1999; 5:512–517.

Finzi D, Hermankova M, Pierson T, Carruth LM, Buck C, Chaisson RE, Quinn TC, Chadwick K, Margolick J, Brookmeyer R, et al. Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. Science. 1997; 278:1295–1300. [PubMed: 9360927]

Han Y, Lassen K, Monie D, Sedaghat AR, Shimoji S, Liu X, Pierson TC, Margolick JB, Siliciano RF, Siliciano JD. Resting CD4+ T cells from human immunodeficiency virus type 1 (HIV-1)-infected individuals carry integrated HIV-1 genomes within actively transcribed host genes. Journal of virology. 2004; 78:6122–6133. [PubMed: 15163705]

Ho YC, Shan L, Hosmane NN, Wang J, Laskey SB, Rosenbloom DI, Lai J, Blankson JN, Siliciano JD, Siliciano RF. Replication- competent noninduced proviruses in the latent reservoir increase barrier to HIV-1 cure. Cell. 2013; 155:540–551. [PubMed: 24243014]

Holman AG, Coffin JM. Symmetrical base preferences surrounding HIV-1, avian sarcoma/leukosis virus, and murine leukemia virus integration sites. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102:6103–6107. [PubMed: 15802467]

Ikeda T, Shibata J, Yoshimura K, Koito A, Matsushita S. Recurrent HIV-1 integration at the BACH2 locus in resting CD4+ T cell populations during effective highly active antiretroviral therapy. The Journal of infectious diseases. 2007; 195:716–725. [PubMed: 17262715]

Imamichi H, Natarajan V, Adelsberger JW, Rehm CA, Lempicki RA, Das B, Hazen A, Imamichi T, Lane HC. Lifespan of effector memory CD4+ T cells determined by replication-incompetent integrated HIV-1 provirus. Aids. 2014

Janovitz T, Klein IA, Oliveira T, Mukherjee P, Nussenzweig MC, Sadelain M, Falck-Pedersen E. High-throughput sequencing reveals principles of adeno-associated virus serotype 2 integration. Journal of virology. 2013; 87:8559–8568. [PubMed: 23720718]

Jordan A, Bisgrove D, Verdin E. HIV reproducibly establishes a latent infection after acute infection of T cells in vitro. The EMBO journal. 2003; 22:1868–1877. [PubMed: 12682019]

Jordan A, Defechereux P, Verdin E. The site of HIV-1 integration in the human genome determines basal transcriptional activity and response to Tat transactivation. The EMBO journal. 2001; 20:1726–1738. [PubMed: 11285236]

Josefsson L, von Stockenstrom S, Faria NR, Sinclair E, Bacchetti P, Killian M, Epling L, Tan A, Ho T, Lemey P, et al. The HIV-1 reservoir in eight patients on long-term suppressive antiretroviral therapy is stable with few genetic changes over time. Proceedings of the National Academy of Sciences of the United States of America. 2013; 110:E4987–4996. [PubMed: 24277811]

Klein IA, Resch W, Jankovic M, Oliveira T, Yamane A, Nakahashi H, Di Virgilio M, Bothmer A, Nussenzweig A, Robbiani DF, et al. Translocation-capture sequencing reveals the extent and nature of chromosomal rearrangements in B lymphocytes. Cell. 2011; 147:95–106. [PubMed: 21962510]

Laird GM, Eisele EE, Rabi SA, Lai J, Chioma S, Blankson JN, Siliciano JD, Siliciano RF. Rapid quantification of the latent reservoir for HIV-1 using a viral outgrowth assay. PLoS pathogens. 2013; 9:e1003398. [PubMed: 23737751]

Lewinski MK, Yamashita M, Emerman M, Ciuffi A, Marshall H, Crawford G, Collins F, Shinn P, Leipzig J, Hannenhalli S, et al. Retroviral DNA integration: viral and cellular determinants of target-site selection. PLoS pathogens. 2006; 2:e60. [PubMed: 16789841]

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; 345:179–183. [PubMed: 24968937]

Mitchell RS, Beitzel BF, Schroder AR, Shinn P, Chen H, Berry CC, Ecker JR, Bushman FD. Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. PLoS biology. 2004; 2:E234. [PubMed: 15314653]

Rose PP, Korber BT. Detecting hypermutations in viral sequences with an emphasis on G --> A hypermutation. Bioinformatics. 2000; 16:400–401. [PubMed: 10869039]

Schroder AR, Shinn P, Chen H, Berry C, Ecker JR, Bushman F. HIV-1 integration in the human genome favors active genes and local hotspots. Cell. 2002; 110:521–529. [PubMed: 12202041]

Serrao E, Krishnan L, Shun MC, Li X, Cherepanov P, Engelman A, Maertens GN. Integrase residues that determine nucleotide preferences at sites of HIV-1 integration: implications for the mechanism of target DNA binding. Nucleic acids research. 2014; 42:5164–5176. [PubMed: 24520116]

Sherrill-Mix S, Lewinski MK, Famiglietti M, Bosque A, Malani N, Ocwieja KE, Berry CC, Looney D, Shan L, Agosto LM, et al. HIV latency and integration site placement in five cell-based models. Retrovirology. 2013; 10:90. [PubMed: 23953889]

Siliciano RF, Greene WC. HIV latency. Cold Spring Harbor perspectives in medicine. 2011; 1:a007096. [PubMed: 22229121]

Silva IT, Rosales RA, Holanda AJ, Nussenzweig MC, Jankovic M. Identification of cromosomal translocation hotspots via scan statistics. Bioinformatics. 2014; 30(18):2551–8. [PubMed: 24860160]

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr. Kinzler KW. Cancer genome landscapes. Science. 2013; 339:1546–1558. [PubMed: 23539594]

Wagner TA, McKernan JL, Tobin NH, Tapia KA, Mullins JI, Frenkel LM. An increasing proportion of monotypic HIV-1 DNA sequences during antiretroviral treatment suggests proliferation of HIV-infected cells. Journal of virology. 2013; 87:1770–1778. [PubMed: 23175380]

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

Wang GP, Ciuffi A, Leipzig J, Berry CC, Bushman FD. HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. Genome research. 2007; 17:1186–1194. [PubMed: 17545577]

Wong JK, Hezareh M, Gunthard HF, Havlir DV, Ignacio CC, Spina CA, Richman DD. Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. Science. 1997; 278:1291–1295. [PubMed: 9360926]

Wu X, Li Y, Crise B, Burgess SM, Munroe DJ. Weak palindromic consensus sequences are a common feature found at the integration target sites of many retroviruses. Journal of virology. 2005; 79:5211–5214. [PubMed: 15795304]

Zhao M, Sun J, Zhao Z. TSGene: a web resource for tumor suppressor genes. Nucleic acids research. 2013; 41:D970–976. [PubMed: 23066107]

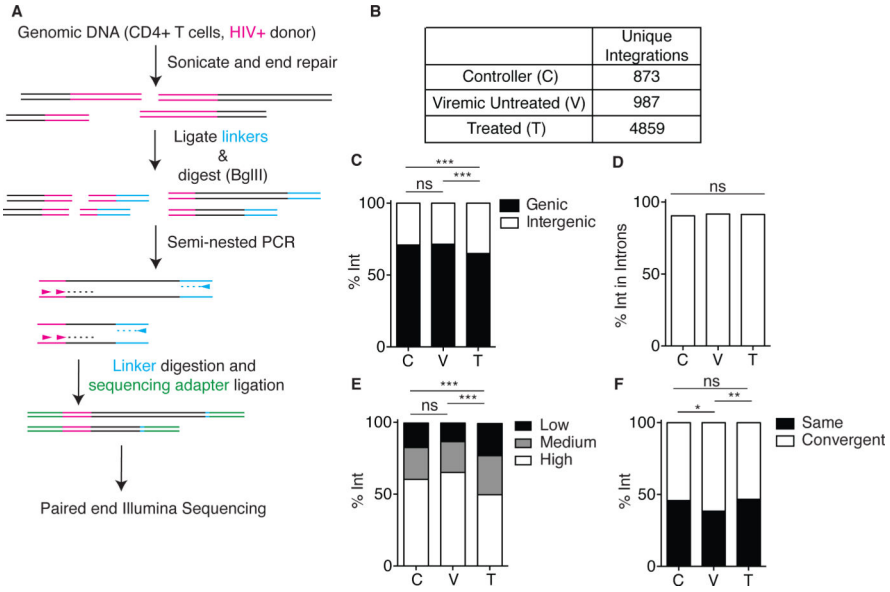

Figure 1. HIV-1 integration libraries, see also Figure S1

A) Diagram of integration library construction. B) Table of unique integrations identified in viremic controllers (C), viremic untreated progressors (V), and treated progressors (T). C) Proportion of integrations (Int) that are in genic or intergenic regions in C, V or T. D) Proportion of genic integrations located in introns in C, V or T. E). Proportion of integrations in genes with high, medium or low expression. P-values refer to proportion of integrations in highly expressed genes. F) Transcriptional orientation of integrated HIV-1 relative to host gene in controllers, viremic or treated progressors. ns: not significant *P<0.05**P< 0.01 ***P<0.0001 using two-proportion z-test.

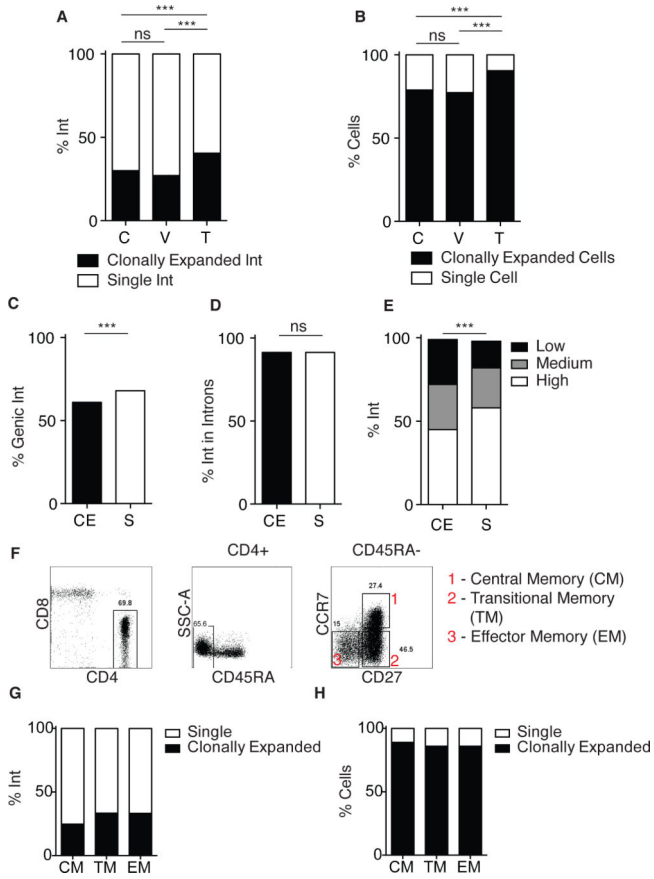

Figure 2. Identification of clonally expanded cells bearing integrated HIV-1, See also Figure S2

A) Proportion of viral integrations (Int) that are clonally expanded, as identified by the same integration site with multiple shears in controllers (C), viremic (V) or treated progressors (T). B) Proportion of infected cells deriving from clonal expansion in C, V or T.. C) Proportion of clonally expanded (CE) and single (S) viral integrations in genic or intergenic regions. D) Proportion of clonally expanded and single viral integrations in introns. E) Proportion of clonally expanded or single viral integrations in genes with high, medium or low expression. P values refer to proportion of integrations in highly expressed genes. F) Seven-parameter flow cytometry sorting strategy to identify $CD4^+$ T cell subsets. CM, TM, and EM cell subsets were identified based on their CD45RA, CCR7, and CD27 expression. Shown is one representative sort. G) Proportion of viral integrations (Int) that are clonally expanded, as identified by the same integration site with multiple shears in sorted $CD4^+$ T cell subsets from patient 9. H) Proportion of infected cells deriving from clonal expansion in sorted $CD4^+$ T cell subsets from patient 9. ns: not significant ***P<0.0001 using two-proportion z-test.

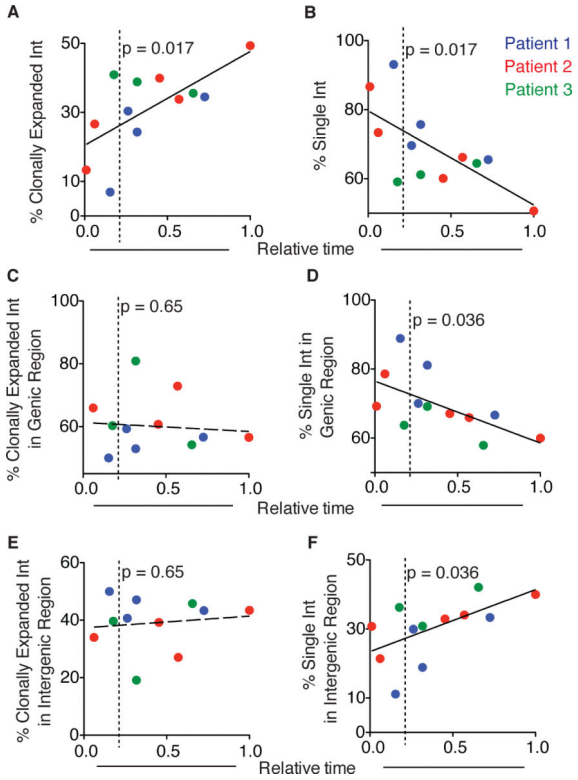

Figure 3. Clonally expanded viral integrations increase and single integrations decrease during therapy, See also Figure S3

Graphs show data from patients 1 (blue), 2 (red) and 3 (green) from longitudinal time points (Table S1). Time was normalized from 0 to 1 (727 days pre therapy to 2617 days post therapy). Dotted line at t = 0.21 marks therapy initiation. Trendline was determined by linear regression model. Solid lines indicate significant change in proportion of events; dashed lines indicate insignificant change in proportion of events. A) Proportion of clonally expanded viral integrations (Int). B) Proportion of single viral integrations. C) Proportion of genic clonally expanded viral integrations. D) Proportion of genic single viral integrations. E) Proportion of intergenic clonally expanded viral integrations. F) Proportion of intergenic single viral integrations.

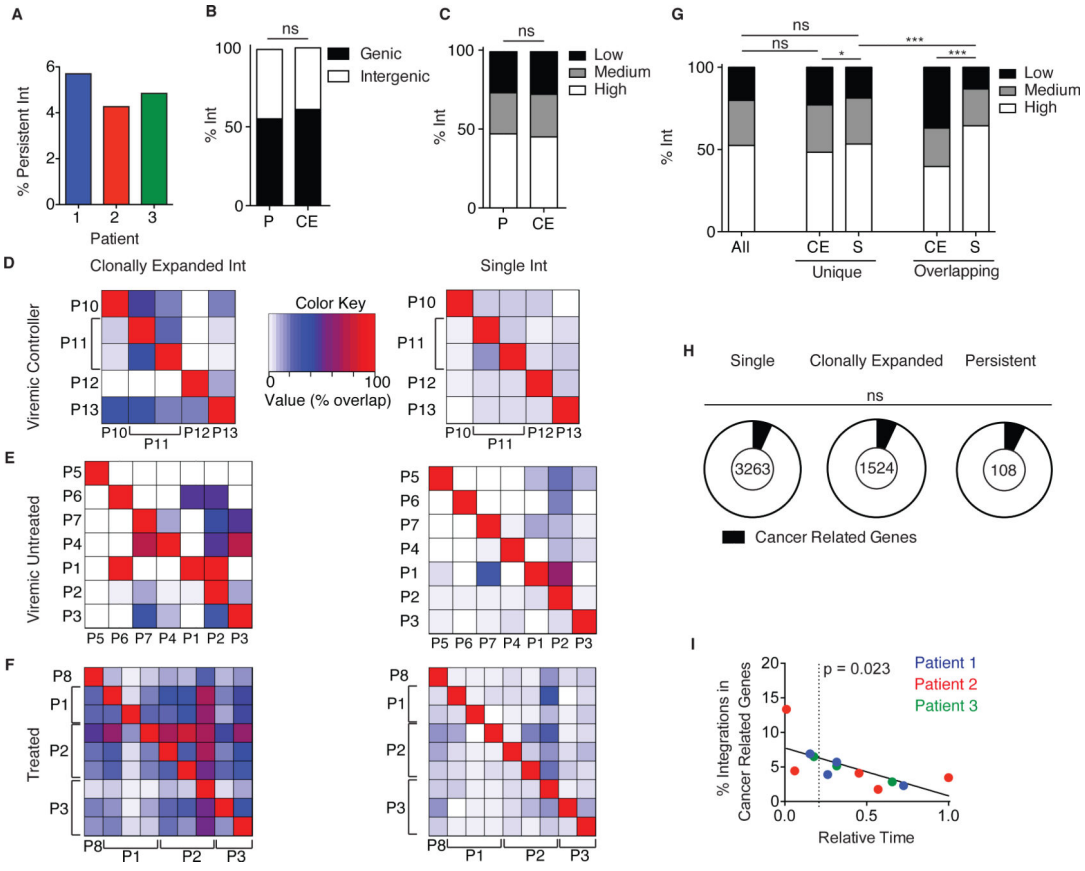

Figure 4. Integrations in genes permissive for clonal expansion occur in multiple patients. See also Figure S4

A) Percent viral integrations present in more than one time point (persistent integrations) in patients 1, 2 and 3 (Table S1). B) Comparison of persistent (P) and clonally expanded (CE) viral integrations in genic or intergenic region. C) Proportion of persistent and clonally expanded viral integrations in genes with high, medium or low expression. P values refer to proportion of integrations in highly expressed genes. D-F) Heatmap showing overlap between samples of genes containing clonally expanded or single viral integrations between samples. Patients are indicated by P1-13. Multiple samples from one individual are marked by a bracket. The amount of overlap is denoted by color (see legend); Red = 100% overlap. G) Genes containing single or clonally expanded viral integrations were analyzed for their presence in multiple patients. Genes with integrations in more than one individual were classified as “overlapping”; genes with integrations in only one individual were classified as “unique.” Shown is the proportion of single and clonally expanded unique and overlapping viral integrations in genes with high, medium or low expression. P values refer to proportion of integrations in highly expressed genes. H) Genes with integrations were analyzed for their association with cancer. Proportions of cancer-associated genes are shown for single, clonally expanded and persistent viral integrations. The number indicates the total number of genes from each category. I) Graph shows proportion of integrations in cancer-related genes from patients 1 (blue), 2 (red) and 3 (green) from longitudinal time points (Table S1). Time was normalized from 0 to 1 (727 days pre therapy to 2617 days post therapy). Dotted line at t = 0.21 marks therapy initiation. Trendline was determined by linear regression model and

indicates significant change in proportion of events, p=0.023. ns: not significant *P<0.05**P< 0.01 ***P<0.0001 using two-proportion z-test.

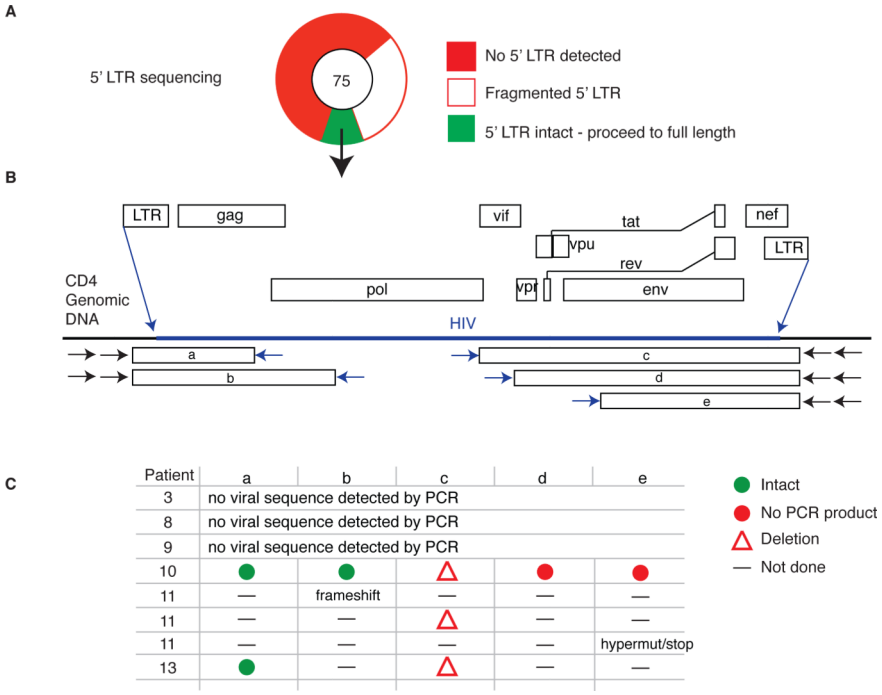

Figure 5. Large expanded clones are defective, See also Data S1

A) Sequence analysis of 5'LTRs in clonally expanded integrations. Of 75 different clonally expanded integrations from 8 individuals, 24 showed fragmented 5' LTRs, 44 didn't have a recoverable 5' LTR, and 8 contained intact 5'LTRs. B) Strategy for HIV-1 sequencing. 8 proviruses were analyzed for intact viral sequence. Nested genomic primers and internal HIV primers were used in a PCR walking strategy to amplify fragments *a-e* from specific clonally expanded integrations. PCR products were sequenced directly. C) Summary of HIV-1 sequencing from large expanded clones. Sequences were aligned to HXB2 and examined for presence of large internal deletions. Intact sequences were analyzed for G → A hypermutation by Los Alamos Hypermut algorithm (Rose and Korber, 2000). Non hypermutated products were analyzed for intact reading frames and frameshift mutations by Los Alamos HIVQC. Green dot: intact, non hypermutated sequence. Red dot: no PCR product recovered. Red triangle: sequence with internal deletion. – : not done.

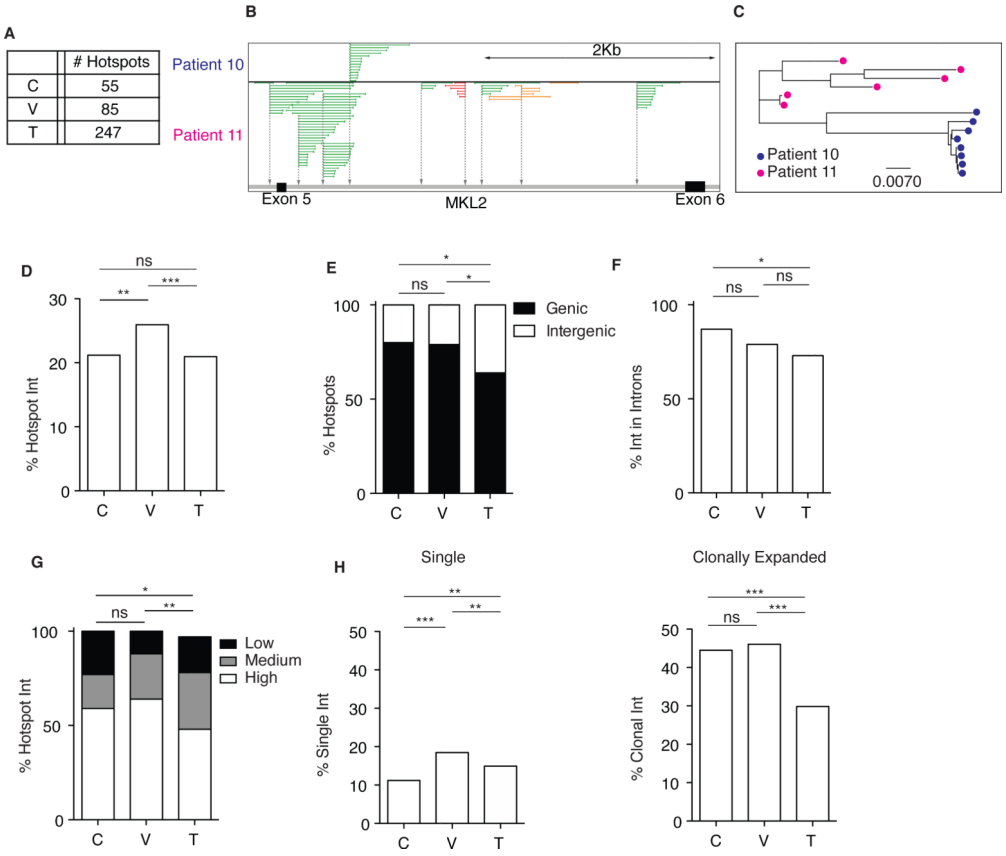

Figure 6. Identification of hotspots for HIV-1 integration

A) Number of hotspots identified by hot-scan in viremic controllers (C), viremic untreated (V) and treated progressors (T). B) Integrations in *MKL2* from patients 10 and 11. Gray vertical arrows indicate site of integrations. Colored horizontal lines show fragments of DNA spanning the point of integration through sheared end. Green: viruses integrated in the same orientation as gene. Red: convergent orientation. Orange: viruses integrated with both orientations. C) HIV-1 *gag* was amplified from integrated proviruses in *MKL2* from patients 10 and 11. PCR was performed using nested integration site-specific primers and HIV-1 *gag* primers. Sequences were clustered to assess DNA sequence similarity. The scale bar represents 0.007 substitutions per site. D) Proportion of virus integrations inside hotspots. E) Proportion of hotspots in genic and intergenic regions. F) Proportion of hotspots in introns. G) Proportion of hotspots in genes with high, medium or low expression. P values refer to proportion of integrations in highly expressed genes. H) Percentage of total single and clonally expanded viral integrations inside hotspots. Enrichment of clonally expanded viral integrations compared to single integrations is significant, p <0.0001. ns: not significant *P<0.05 **P< 0.01 ***P<0.0001 using proportion test

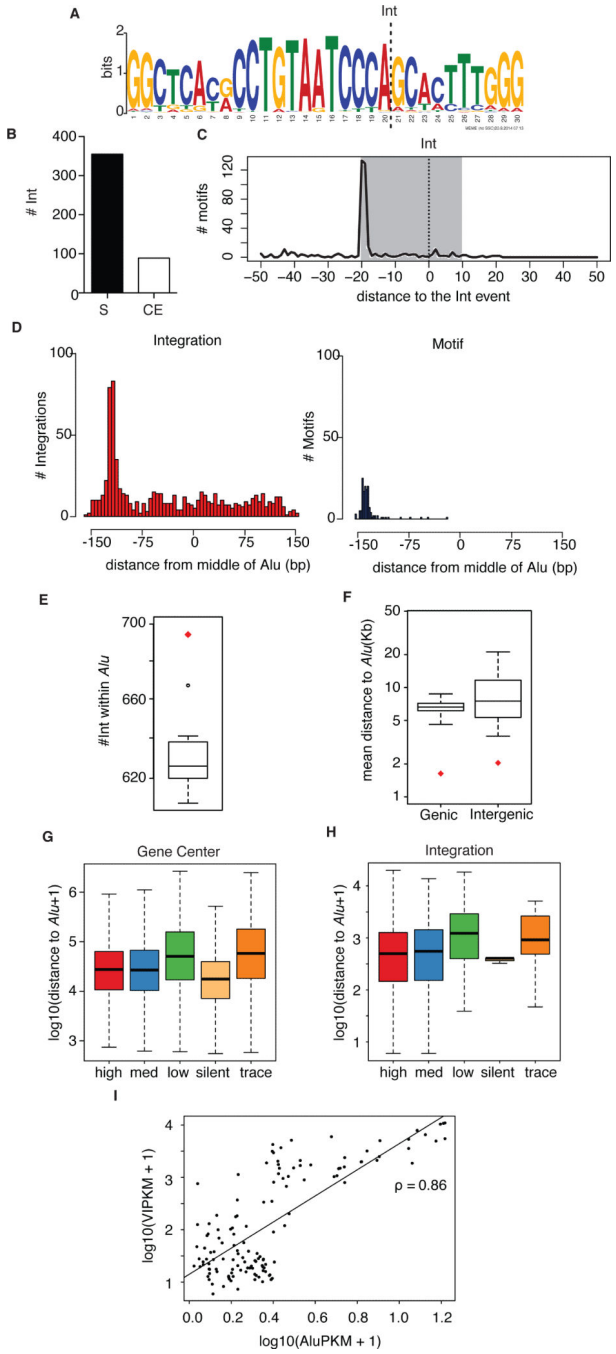

Figure 7. Consensus motif for viral integration

A) 30bp sequence consensus motif (INT-motif). 100bp around all viral integration sites were analyzed for a consensus sequence by MEME (Bailey and Elkan, 1994). 444 integration sites were identified with the INT-motif. E-value: 6.4×10^{-4071}. The dotted line shows the preferred site of integration (see also (C)). B) Number of single (S) and clonally expanded (CE) which were identified to contain INT-motif within 100bp of the integration site. P<0.0001, using two-proportion z-test. C) Conserved integration site within INT-motif. Histogram maps the start site (5' end) of INT-motif with respect to the integration site

(dotted line). Peak shows the majority of integration sites occur 20bp from the 5' end of the motif start site. Shaded region represents the location of the INT-motif relative to the majority of the integration sites. D) Location of integration preference and INT-motif inside *Alu* repeats is overlapping. Left, location of integration sites *Alu* repeats were plotted relative to the midpoint of the repeat. Right, the location of the start site of INT-motifs within *Alu* repeats. E) Integrations are enriched inside *Alu* repeats. Total integrations identified inside *Alu* repeats were enumerated (red diamond) and compared to the expected value as defined by Monte Carlo simulation. The boxplot displays the variation of the number of random integrations identified inside *Alu* repeats by each iteration of the simulation. F) Integrations are near *Alu* repeats in genes and intergenic regions. Average distance to the nearest *Alu* repeat for all integrations inside genes or intergenic regions was calculated (red diamond) and compared to the expected distance as defined by Monte Carlo simulation. The boxplot displays the variation of the distance of random integrations from *Alu* repeats in genes or intergenic regions by each iteration of the simulation. G) Distance to *Alu* repeats from the center of highly, medium, low, trace or silently expressed genes. H) Distance to *Alu* repeats in highly, medium, low, trace or silently expressed genes. I) Positive correlation between *Alu* repeats and integrations inside hotspots. Graph shows number of *Alu* repeats (X axis) vs. integrations in hotspots (Y axis). Hotspots not containing *Alu* repeats were removed from this analysis. The scatter plot shows the linear relationship between the number of INT-motifs and integrations inside hotspots (Pearson's correlation, $\rho = 0.86$).