Assessment of HIV-1 integration in tissues and subsets across infection stages

Vincent H. Wu,[1,2] Christopher L. Nobles,[1] Leticia Kuri-Cervantes,[1,2] Kevin McCormick,[1] John K. Everett,[1] Son Nguyen,[1,2] Perla M. del Rio Estrada,[3] Mauricio González-Navarro,[3] Fernanda Torres-Ruiz,[3] Santiago Ávila-Ríos,[3] Gustavo Reyes-Terán,[3] Frederic D. Bushman,[1] and Michael R. Betts[1,2]

[1]Department of Microbiology and [2]Institute for Immunology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [3]Centro de Investigación en Enfermedades Infecciosas, Instituto Nacional de Enfermedades Respiratorias, Mexico City, Mexico.

The integration of HIV DNA into the host genome contributes to lifelong infection in most individuals. Few studies have examined integration in lymphoid tissue, where HIV predominantly persists before and after antiretroviral treatment (ART). Of particular interest is whether integration site distributions differ between infection stages with paired blood and tissue comparisons. Here, we profiled HIV integration site distributions in sorted memory, tissue-resident, and/or follicular helper CD4$^+$ T cell subsets from paired blood and lymphoid tissue samples from acute, chronic, and ART-treated individuals. We observed minor differences in the frequency of nonintronic and nondistal intergenic sites, varying with tissue and residency phenotypes during ART. Genomic and epigenetic annotations were generally similar. Clonal expansion of cells marked by identical integration sites was detected, with increased detection in chronic and ART-treated individuals. However, overlap between or within CD4$^+$ T cell subsets or tissue compartments was only observed in 8 unique sites of the 3540 sites studied. Together, these findings suggest that shared integration sites between blood and tissue may, depending on the tissue site, be the exception rather than the rule and indicate that additional studies are necessary to fully understand the heterogeneity of tissue-sequestered HIV reservoirs.

Introduction

HIV infection remains an important global health issue, despite the advent of antiretroviral therapy (ART). An HIV cure remains elusive, mainly due to integration of viral DNA into the genome of potentially long-lived memory CD4$^+$ T cells. This integration of intact HIV-1 proviruses can establish a latent reservoir that is capable of lifelong persistence based on estimated rates of viral decay (1, 2). ART interruption typically leads to viral rebound, indicating that intact proviruses lie dormant during ART treatment but can later be reactivated (3). Increasing our understanding of the characteristics of the HIV-1 reservoir is necessary for the development of a functional cure.

HIV-1 integrates DNA copies of the RNA genome into host cells, predominantly through a coordinated interaction between HIV integrase and host-encoded LEDGF/p75 (4–6). HIV-1 integration sites are found more commonly in active transcription units, a finding which is largely explained by the ability of LEDGF/p75 to tether to these sites (4, 7–10). As HIV-1 integration is not site specific, the host-viral junction site in the genome provides a unique marker of infected cells that allows for the tracking of these cells through integration site analyses. Longitudinal tracking studies have shown an increase in clonally expanded cells during ART, with a selection for cells containing integration sites within cancer-related genes over time (11, 12). These clones are established early in HIV-1 infection and persist after ART treatment (11, 13, 14). However, most integration site studies have been carried out on bulk PBMCs or peripheral blood CD4$^+$ T cells, which may not fully recapitulate the complexity of the integrated reservoir in the body for two reasons: (a) the majority of HIV-1–infected CD4$^+$ T cells reside in lymphoid tissues (15–17) and (b) CD4$^+$ T cells are composed of multiple heterogeneous tissue-sequestered and circulating subsets in vivo (18–22).

CD4$^+$ T cells display diverse memory and effector phenotypes, each with different epigenetic and transcriptomic signatures. While many studies have compared various internal proviral sequences across cell subsets (12, 23–27), few studies have compared the integration site landscape between different cell subsets (12).

Conflict of interest: The authors have declared that no conflict of interest exists.

Submitted: May 11, 2020
Accepted: September 17, 2020
Published: September 24, 2020

Reference information: *JCI Insight*. 2020;5(20):e139783. https://doi.org/10.1172/jci.insight.139783.

Viral sequence studies generally observed overlapping sequences between cell subsets, but one study has suggested potential compartmentalization based on *nef* sequences between memory cell subsets in 3 of 5 individuals (23). In vivo compartmentalization of CD4$^+$ T cells can be seen with tissue-resident memory cells in different contexts, including HIV-1 infection (28, 29). This raises the hypothesis that tissue residency may lead to a different landscape of integration sites between tissue and blood, especially for CD4$^+$ T cells that are either tissue resident or have limited egress capabilities. Overall, these studies highlight a gap in our understanding of HIV-1 reservoirs, demonstrating the need to better understand the integration site landscape in the context of tissue/blood compartments and CD4$^+$ T cell heterogeneity at different stages of HIV-1 infection.

To address this, we obtained PBMCs and lymph node mononuclear cells (LNMCs) from cervical lymph nodes (CLN) and inguinal lymph nodes (ILN) from HIV-infected untreated acutely infected/chronically infected (acute/chronic) individuals or tonsils from ART-treated individuals. From these, we profiled integration sites within memory, follicular, and resident CD4$^+$ T cell subsets to assess the effect of compartmentalization and cellular residency on integration sites. While we observed that HIV-1 integration site targeting characteristics were largely shared in the different compartments, cell subsets, and stages of HIV-1 infection, overlap of integration sites between subsets and compartments was rare, despite the presence of clonally expanded sites. Together, these findings suggest that tissue compartmentalization of the HIV-1 reservoir can occur between different anatomical sites, supporting the need to further understand the dynamics and fate of HIV-infected cells from different cell subsets and tissues.

Results

Study design and identification of HIV-1 integration sites. To assess the effect of compartmentalization and cellular residency on integration site distributions, we sorted LNMCs from ART-treated individuals into 4 CD4$^+$ T cell subsets (Supplemental Figure 1; supplemental material available online with this article; https://doi.org/10.1172/jci.insight.139783DS1): (a) CD45RA$^+$CCR7$^+$ (naive; for individuals T1 and T2 only), (b) HLA-DR$^-$CXCR5^{hi}PD-1^{hi} (germinal center T follicular helper [GC-Tfh]; HLA-DR$^-$ GC-Tfh), (c) non–GC-Tfh memory HLA-DR$^-$CD69$^+$ (labeled as HLA-DR$^-$CD69$^+$), and (d) non–GC-Tfh memory HLA-DR$^-$CD69$^-$ (labeled as HLA-DR$^-$CD69$^-$). As HLA-DR has been shown to be a marker of T cell activation (30), we sorted HLA-DR$^-$ cells to enrich for the latent reservoir in ART-treated individuals as seen in previous studies (31–33). Because the residency marker CD69 is also upregulated by T cells upon activation (34–36), we selectively sorted HLA-DR$^-$CD69$^+$ cells to focus on resting resident CD4$^+$ T cells.

From acutely and chronically infected lymph nodes, we sorted 4 CD4$^+$ T cell subsets: (a) CD45RA$^+$CCR7$^+$ (naive), (b) nonnaive PD-1^{hi}CXCR5^{hi} (GC-Tfh), (c) non–GC-Tfh memory CD69$^+$ (labeled as CD69$^+$), and (d) non–GC-Tfh memory CD69$^-$ (labeled as CD69$^-$) (Supplemental Figure 2). Given the high viral load and ongoing inflammation in these individuals, we did not exclude HLA-DR$^+$CD4$^+$ T cells. We did include the CD69$^+$ sorting strategy from ART-treated samples to allow analysis of integration sites in resident memory CD4$^+$ T cells; however, some of these cells may have been recently activated. Regardless, CD69 expression would still limit tissue egress potential of any recently activated CD69$^+$ cells, resulting in potentially unequal distribution of integration sites between tissues and blood.

PBMCs were sorted into memory phenotypes as described in Supplemental Figures 1 and 2. These subsets consisted of (a) CD45RA$^+$CCR7$^+$ (naive), (b) CD45RA$^-$CCR7$^+$ (central memory T [Tcm]), (c) CD45RA$^-$CCR7$^-$ (effector memory T [Tem]), and (d) CD45RA$^+$CCR7$^-$ (Tem cells reexpressing CD45RA [Temra]) for acute and chronic samples only. For ART-treated individuals, the Tcm and Tem subsets further excluded CD69$^+$ cells to select for resting cells. All sort counts are detailed in Supplemental Table 1, while information on individuals is provided in Table 1.

Sorted cells were assayed for integration sites using an established pipeline (37, 38), wherein sonicated adapter-ligated genomic DNA was amplified with a nested PCR using HIV-1 U3 and U5 long terminal repeat primers. Since random fragment lengths are generated during sonication, fragments with different lengths that are associated with integration sites mapping to the same genomic position can be counted as a proxy for the number of infected cells (termed sonic abundance) (38). Even with the limited number of cells (ranging from ~3000 to >400,000 cells per subset), integration site analyses are sufficient to detect clonal CD4$^+$ T cell expansions carrying integrated virus (39). We recovered a total of 3540 unique sites and an inferred count of 3973 cells with integration sites across 9 individuals. The breakdown of sites by infection stage, location, and cell subset are shown in Tables 2, 3, and 4, while the entire data set of sites is provided in Supplemental File 1.

Table 1. Individual information

Individual	HIV-1 infection stage	Time on ART (mo)	Age (yr)	Sex	Compartments	Viral load (cps/ml)	$CD4^+$ T cell counts (cells/μl)
A1	Acute (Fiebig IV)	–	22	M	P, I, C	179,832	696
A2	Acute (Fiebig IV)	–	27	M	P, I, C	5,112,511	589
A3	Acute (Fiebig IV)	–	24	M	P, I, C	58,355	501
C1	Chronic	–	32	M	P, I, C	1,771,593	462
C2	Chronic	–	23	M	P, I, C	143,916	509
C3	Chronic	–	21	M	P, I, C	1,019,989	251
T1	ART treated	16.9	24	M	P, T	<40	584
T2	ART treated	93	32	M	P, T	<40	497
T3	ART treated	38.64	32	M	P, T	<40	346

M, male; P, PBMC; I, inguinal lymph node; C, cervical lymph node; T, tonsil.

Comparable characteristics in HIV integration sites across infection stage, compartment, and cell subset. For our initial analysis, we conducted a deep characterization of the genomic characteristics of the integration sites obtained from each donor. Comparable to previous studies (4, 7, 8), the majority of sites were enriched in transcription units, with about 15% being in distal intergenic regions. The majority of the integration sites (71%) were found within introns. When separated by experimental factors (infection stage, compartment, and cell subset), the pattern of favored integration events in transcription units was also observed at similar relative values (Figure 1 and Supplemental Figure 3A). In order to assess any positional biases in chromosomal position, we generated bins after normalizing for chromosomal size (Supplemental Figure 3B). Most integration sites were detected near the end of the chromosome as previously observed (40). While there was no discernible pattern between infection stages, we observed a higher distribution of integration sites at the ends of chromosomes from acute individuals compared with those from chronic or ART individuals (Supplemental Figure 3B). This parallels findings from Haworth et al., where their sites recovered from in vitro cultures had increased frequency at either end of the chromosome compared with sites from 12- to 14-week infections in humanized mice (40).

When grouping our recovered integration sites by infection stage and compartment, we observed similar results with the majority of recovered sites being in intronic regions of transcription units followed by distal intergenic regions. However, there was a trend for an increase in nonintron and nondistal intergenic sites from tissue compartments during ART (Supplemental Figure 3, C and D). When combining these specific annotations (labeled "Other") into one category and depicting each subset, we observed a significant increase from tissue samples during ART (Supplemental Figure 4A; $P < 0.05$; Wilcoxon rank sum test). This increase was also observed in sites obtained from cellular subsets with a residency phenotype during ART (Supplemental Figure 4B; $P < 0.01$; Wilcoxon rank sum test), suggesting integration sites in resident cells have different integration site profiles or that there are different selective pressures compared with infected nonresident cells during ART. Furthermore, integration sites in distal intergenic regions were enriched in the blood compared with tissue during ART (Supplemental Figure 4A; $P < 0.05$; Wilcoxon rank sum test).

We next pooled the unique transcription unit sites across individuals by infection stage and anatomical compartment and assessed the enrichment of gene ontology annotations using Metascape (41) (Supplemental Figures 5–7). The significant gene ontology terms were similar across the stages, with potentially more enrichment of chromatin modification/organization-related genes in sites identified from lymphoid tissue than those found in blood. Due to the sparsity of our samples, we used a previously published bulk RNA-Seq data set on GC-Tfh and non-Tfh cells from lymph nodes of healthy donors (42) to generate bins of increasing gene expression. Genic integration sites from GC-Tfh cells (independent of HLA-DR) were in genes found in the higher expression bins for the GC-Tfh RNA-Seq data set, but other genic integration sites also followed a similar pattern (Supplemental Figure 8A). This was similarly observed with the non-Tfh cell RNA-Seq data set (Supplemental Figure 8B), suggesting that there was no bias observed for subset-specific integration sites based on gene expression from published RNA-Seq data sets. The overall enrichment of normal cellular pathway and lymphocyte function terms along with the RNA-Seq comparison supports the preferential integration of HIV DNA into genes with higher levels of transcription (43–45).

Table 2. Total integration sites and clones profiled for acute individuals

State	Individual	Compartment	Cell subset	Subset sonic abundance	Subset no. clones[A]	Location sonic abundance sum	Location no. clones[A]	Individual sonic abundance	Individual no. clones[A]
Acute	A1	CLN	CD69^{-}	138	2	285	2	540	9
			GC-Tfh	123	0				
			CD69^{+}	22	0				
			Naive	2	0				
		PBMC	Tcm	122	4	188	6		
			Tem	66	2				
		ILN	CD69^{-}	30	0	67	1		
			GC-Tfh	18	1				
			CD69^{+}	15	0				
			Naive	4	0				
	A2	ILN	CD69^{+}	97	0	213	2	496	3
			CD69^{-}	84	2				
			GC-Tfh	30	0				
			Naive	2	0				
		CLN	CD69^{+}	51	0	143	0		
			GC-Tfh	48	0				
			CD69^{-}	43	0				
			Naive	1	0				
		PBMC	Tem	100	1	140	1		
			Tcm	36	0				
			Temra	4	0				
	A3	PBMC	Tcm	13	0	13	0	25	0
		CLN	CD69^{-}	5	0	7	0		
			CD69^{+}	1	0				
			Naive	1	0				
		ILN	CD69^{-}	4	0	5	0		
			CD69^{+}	1	0				

[A] Clones determined by ≥3 sonic abundance.

We then compared genomic annotations of the recovered integration sites, including features such as DNaseI accessibility, CpG counts/density, and GC content (46, 47). Briefly, 3 random control sites were chosen in the human genome for each integration site in the data set to calculate a receiver operating characteristic (ROC) curve. A ROC value of 1 indicates that all actual integration sites in the data have a given feature (or have higher values) compared with all random controls; the opposite is true for a ROC value of 0. A ROC value of 0.5 indicates that this feature is unable to discern between actual and random control sites. For annotations with varied genomic windows, the overall integration site characteristics were comparable across infection stages, compartments, and cell subsets (Supplemental Figure 9A). While there were subtle differences (i.e., GC content within a 1000 bp window being more significantly enriched in the acute and ART-treated samples compared with chronic samples), our findings are consistent with the tendencies of preferential HIV integration into genic regions (4, 44, 46, 48).

Finally, we assessed epigenetic annotations within a 10 kb window of the integration site (±5000 bp). Due to the lack of available annotations specific for each sorted subset, we used epigenetic annotations from bulk CD4^{+} T cells for all comparisons (47). While in general we observed comparable epigenetic annotations, we did identify differences across the different combinations of factors (Supplemental Figure 9B). For instance, H3K79me3 histone methylation was lower in integration sites during ART and chronic states compared with random controls but not during acute infection (Supplemental Figure 9B). H3K79me3 is an important factor for transcriptional regulation that correlates with active gene transcription (49, 50). Further studies would be needed to determine if these sites are depleted in H3K79me3 during ART and chronic states in the specific cell subsets.

Identification of integration hotspots during multiple stages of HIV infection. We next identified genes commonly targeted for integration during different stages of HIV infection. We filtered the integration sites

Table 3. Total integration sites and clones profiled for chronic individuals

State	Individual	Compartment	Cell Subset	Subset sonic abundance	Subset no. clones	Location sonic abundance sum	Location no. clones	Individual sonic abundance	Individual no. clones
Chronic	C3	CLN	CD69^{-}	335	44	486	50	721	64
			GC-Tfh	128	5				
			CD69^{+}	12	1				
			Naive	11	0				
		ILN	CD69^{-}	101	6	159	9		
			GC-Tfh	49	3				
			CD69^{+}	7	0				
			Naive	2	0				
		PBMC	Tcm	36	2	76	5		
			Tem	34	2				
			Naive	4	1				
			Temra	2	0				
	C2	CLN	GC-Tfh	138	9	301	25	598.5	45
			CD69^{-}	117	12				
			CD69^{+}	45	4				
			Naive	1	0				
		ILN	CD69^{-}	109	6	156.5	11		
			CD69^{+}	27.5	1				
			GC-Tfh	12	2				
			Naive	8	2				
		PBMC	Tem	75	4	141	9		
			Tcm	62	5				
			Naive	4	0				
	C1	CLN	CD69^{-}	157	4	238	9	568	22
			CD69^{+}	63	2				
			GC-Tfh	10	1				
			Naive	8	2				
		ILN	CD69^{-}	85	4	231	6		
			GC-Tfh	76	2				
			CD69^{+}	64	0				
			Naive	6	0				
		PBMC	Tem	65	4	99	7		
			Tcm	23	1				
			Temra	8	1				
			Naive	3	1				

for genes (or nearest genes within 50 kb) present in 3 or more combinations of individuals/cell subsets/compartments for acute and ART-treated individuals (Figure 2, A and C). We raised this threshold to 4 or more combinations for chronic individuals due to the larger number of genes (Figure 2B).

For acute and chronic infections (Figure 2, A and B, respectively), we identified numerous hotspots. Integrations in or near *PACS1* and *ANKRD17* were the most common hotspot in acute and chronic data sets, respectively. Furthermore, we detected integrations within or near *STAT5B* and *FOXK2* genes during acute and/or chronic states, lending support to the notion that clonally expanded sites during ART treatment can be established early during infection (13, 14, 51). We also assessed pairwise enrichment of genes between infection stages and identified genes trending toward enrichment in different infection states (Supplemental Tables 2–4). We did not assess statistical significance due to the limited power of our sample size.

In ART-treated individuals, we observed a number of genes classically associated with integration, including *STAT5B*, *FOXK2*, *BACH2*, and *TET2*/*TET2-AS1* (13, 51, 52), across different subjects, tissues, and CD4^{+} T cell subsets. However, none of these integration-associated genes were found in every individual or within all of the assessed cell subsets within an individual. In the HLA-DR^{-}CD69^{-} population from tonsils of individual T1, we detected a *TET2* (and *TET2-AS1*) integration in intron 2, which is of clinical interest given

Table 4. Total integration sites and clones profiled for ART-treated individuals

State	Individual	Compartment	Cell Subset	Subset sonic abundance	Subset no. clones	Location sonic abundance sum	Location no. clones	Individual sonic abundance	Individual no. clones
ART	T1	Tonsil	HLA-DR$^-$ CD69$^+$	272.5	17	595	37	625.5	38
			HLA-DR$^-$ CD69$^-$	237.5	14				
			HLA-DR$^-$ GC-Tfh	83	6				
			Naive	2	0				
		PBMC	CD69$^-$ Tem	27.5	1	30.5	1		
			Naive	3	0				
	T2	PBMC	CD69$^-$ Tem	188	2	207	6	245	7
			CD69$^-$ Tcm	19	4				
		Tonsil	HLA-DR$^-$ CD69$^-$	33	1	38	1		
			HLA-DR$^-$ CD69$^+$	5	0				
	T3	PBMC	CD69$^-$ Tem	95	13	131	14	154	16
			CD69$^-$ Tcm	35	1				
			Naive	1	0				
		Tonsil	HLA-DR$^-$ GC-Tfh	11	0	23	2		
			HLA-DR$^-$ CD69$^-$	6	2				
			HLA-DR$^-$ CD69$^+$	6	0				

a recent report on the in vivo expansion of a CAR T cell with an integrated lentiviral vector in *TET2* (47, 52). Together, our data support previous studies documenting the enrichment of certain sites during ART treatment, but highlight the differential complexity of the integrated reservoir within each individual. We did not find any single specific gene preferentially associated with integration across all individuals or tissues.

Detection of clonal expansion across multiple stages of infection. Clonal expansion in peripheral blood has been documented previously during ART treatment (13, 51) as well as acute and chronic infection (14) but has not been compared between blood and lymphoid tissue at multiple stages of infection. Using sonic abundance, we could identify separate fragments with the same integration site originating from clonally expanded cells with the same integration site in all 3 stages of infection (Supplemental File 1 and Figures 3, 4, and 5). To identify cell clones during acute infection (Figure 3), we used a sonic abundance threshold of 3 to ensure the detection of at least 2 infected cells, as cells undergoing mitosis during acute infection may die (14). We found limited examples of clonal expansion during acute infection, with the most dominant clone (by absolute sonic abundance) integrated in *TUBGCP6* in the PBMC Tcm subset from individual A1 (Figure 3). For both the chronic and ART-treated samples, the sonic abundance threshold was set to 2. In chronic samples, we detected higher levels of clonal expansion, with some clones representing 10% or more of the total relative abundance in a cell subset (Figure 4). The most dominant clone detected in a chronic individual was integrated in *ZNF34* within the CD69$^+$ subset in CLN from individual C2. During ART treatment, some clones also represented 10% or more of the total infected cell population detected in a given cell subset, similar to our observations for chronic individuals (Figure 5). The most dominant clone, by absolute sonic abundance, contained an integrated provirus in *PTPN9,* within the HLA-DR$^-$ CD69$^-$ subset in the tonsil from individual T2. Overall, these results indicate that clonal expansion can be detected in tissues early in infection as well as during chronic infection and ART treatment.

Within the subset and infection stage resolution across our cohort, we asked if any particular subsets (pooled across tissue sites) were enriched for clones. As detailed in Supplemental Table 5, most subsets contained clones. In acute individuals, Tcm cells had the highest percentage, with 2.78% of detected unique sites

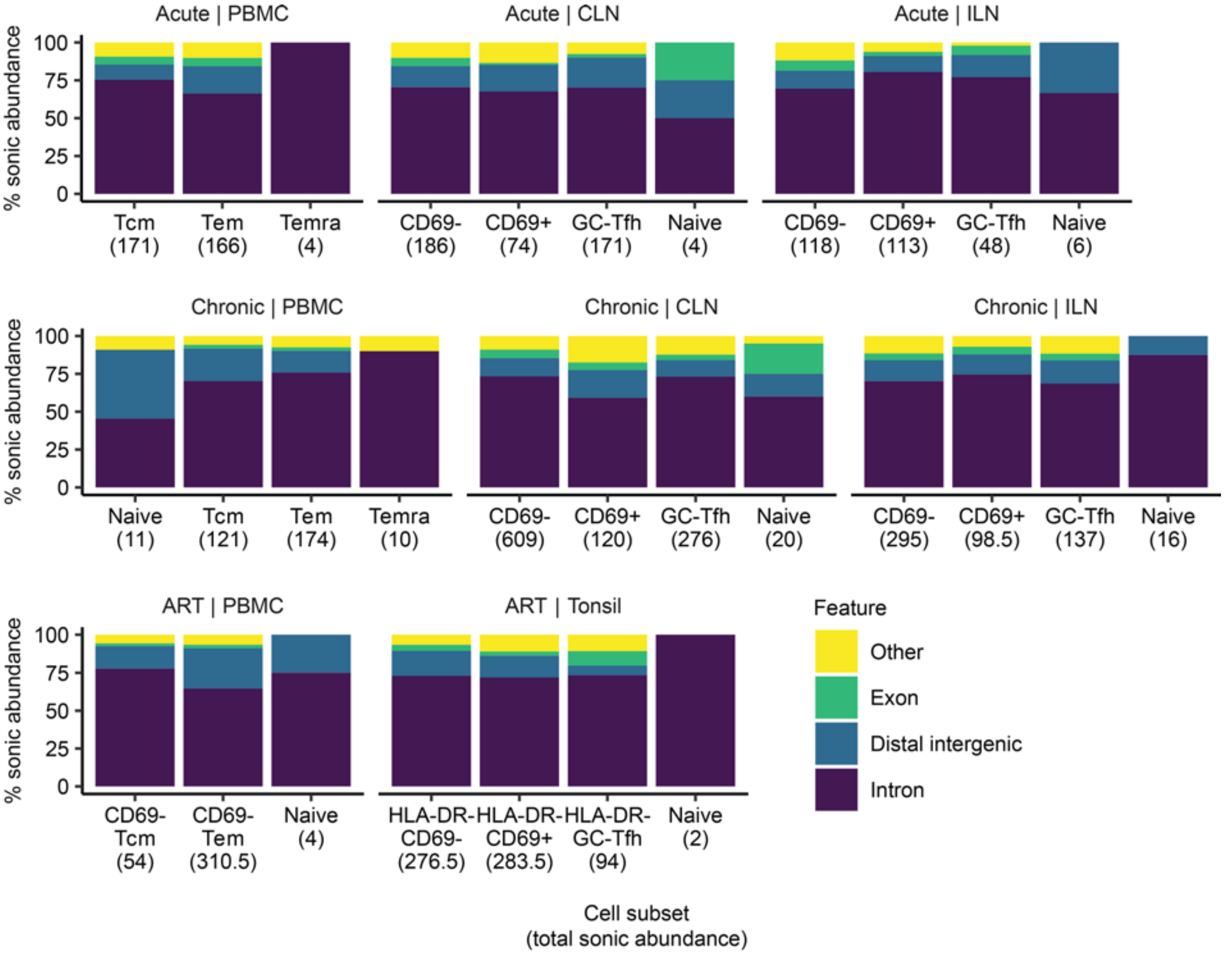

Figure 1. HIV-1 integration sites are primarily found in intronic regions across infection stages, compartments, and cell subsets. For each cell subset, the percentage of total integration sites combined across individuals is displayed for each feature annotation. Integration sites are binned based on cell subset, where inferred cell numbers by sonic abundance are shown in parentheses.

being clonal. The maximum values were 9.54% ($CD69^-$ from tissues) from chronic individuals and 7.56% ($HLA-DR^-CD69^-$ from PBMCs) from ART-treated individuals when considering only subsets in which we recovered more than 100 unique sites. Furthermore, clonal proportions significantly differed between infection stages (Supplemental Figure 10). There was also a trend toward higher clonality in nonresident phenotypes compared with resident phenotypes in chronic and ART-treated individuals, but this did not reach statistical significance, likely due to limited sample size. Together, these data suggest that the proportion of clonal sites increases over time during the course of infection and with the installment of ART and highlights the need to further explore tissue-resident phenotypes in the context of HIV-1 reservoirs.

Limited overlap detected between cell subsets and compartments. Recent reports have documented overlap of proviral sequences across compartments and across subsets. We thus examined whether there was overlap of integration sites within individuals between different T cell subsets. Since integration sites can serve as a unique barcode of an infected cell, we used integration site data to examine if infected cells after cell division have trafficked and/or differentiated.

For acute infection, we detected no overlap in any of the individuals (Figure 6A), likely due to the rampant viral replication and death/turnover of infected cells. Despite the depth of our analysis, we only detected a few overlaps of integration sites between data from T cell subsets or anatomic sites in 2 chronically infected individuals (Figure 6B). These included a clone in *LINC02068* shared between ILN (naive) and CLN ($CD69^-$), a clone near *LINC00466* in PBMCs shared between naive and Tem cells, a clone in *UBE2W* shared between CLN ($CD69^-$) and PBMCs (Tcm), and a clone in *INPP4B* shared between CLN ($CD69^-$) and PBMCs (Temra).

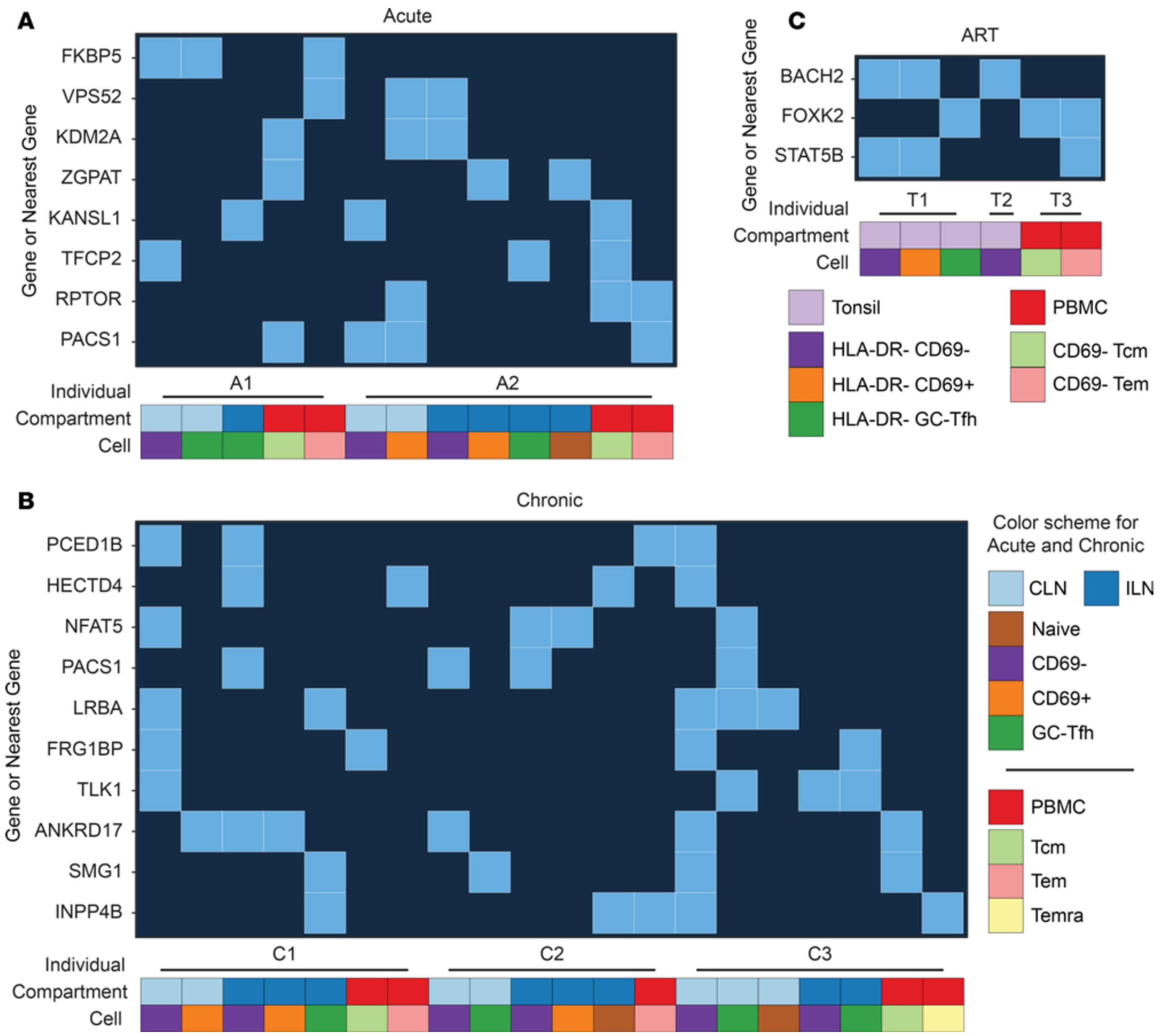

Figure 2. Hotspots of HIV-1 integration across individuals, compartments, and subsets during different infection stages. Integration sites in a gene or within 50 kb of a gene were identified across combinations of individual, compartment, and cell subsets. (**A**) Hotspots identified in 3 or more combinations in acute individuals. (**B**) Hotspots identified in 4 or more combinations in individuals with chronic HIV-1 infection. (**C**) Hotspots identified in 3 or more combinations in ART-treated individuals.

We only detected overlap of integration sites between subsets or tissue in 1 of the 3 ART-treated individuals (Figure 6C). These included a *BACH2* clone shared between tonsil HLA-DR$^-$CD69$^+$ and HLA-DR$^-$CD69$^-$CD4$^+$ T cells, 2 separate *TRABD* clones between tonsil HLA-DR$^-$ GC-Tfh and HLA-DR$^-$CD69$^+$CD4$^+$ T cells, and an expanded clone near *P4HB* between tonsil HLA-DR$^-$CD69$^-$ and HLA-DR$^-$CD69$^+$CD4$^+$ T cells. No overlaps were detected between peripheral blood CD4$^+$ T cells and tonsil cells in any of the 3 ART-treated individuals, despite the detection of clonal expansion (Figure 6C).

We assessed partitioning of clones in different anatomical sites by comparing our observations to results of computational simulations. We assumed that clones could be shared randomly between compartments and/or subsets. For each individual, we pooled all the sites according to their sonic abundance and then randomly sampled the pool based on the original sample size recovered in a given cell subset. The number of overlaps were counted across 10,000 simulations. Our simulations indicate that considerable overlap is expected in all pairs of samples tested (Supplemental Figure 11), in contrast to our observations of limited overlap. Thus, our integration site data set was not dominated by expanded clones that were common to multiple compartments, though it will be valuable to challenge this conclusion in future studies with deeper sampling over more subjects.

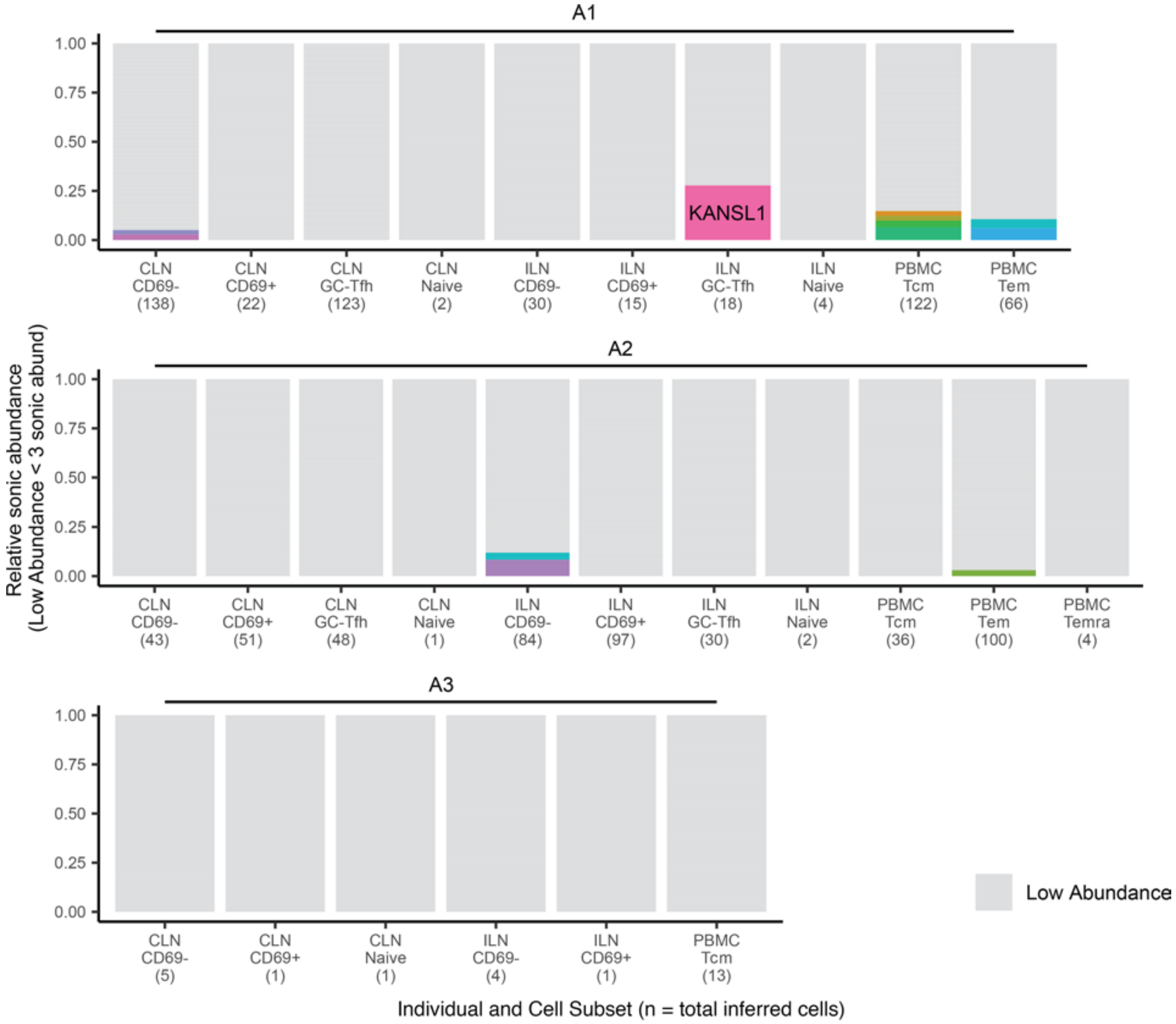

Figure 3. Identification of clonal expansion from individuals during acute infection. Relative sonic abundance for each integration site is shown for each acutely infected individual. Sites with color indicate that the site is clonal (defined as a sonic abundance ≥3 for acute infection). Different colors indicate different integration sites across an individual (i.e., two different integration sites in the same gene with clonal expansion would have different colors). Clones consisting of 10% or more of the total relative sonic abundance are labeled with the integrated gene or nearest gene. Genes within 50 kb of the integration site are denoted with "+," while genes that are more then 50 kb from the integration site are demoted with "#." The number of inferred total cells is indicated parenthetically on the *x* axis.

Discussion

Our current understanding of the integrated HIV-1 reservoir is largely derived from studies of the peripheral blood on bulk CD4$^+$ T cells, with the assumption that these represent infected CD4$^+$ T cells in lymphoid tissues. While some infected CD4$^+$ T cells may traffic between tissues and blood, it is unclear to what degree this occurs. Moreover, some memory CD4$^+$ T cells can establish residence in tissues and thus may not be well represented in the circulation. For example, Tfh cells, a population known to be a key HIV-1 reservoir, are mostly anatomically restricted to lymphoid tissues. These anatomical and subset characteristics of CD4$^+$ T cells raise the question of whether bulk peripheral blood CD4$^+$ T cell populations provide an accurate representation of the HIV-1 integration landscape throughout the body at multiple stages of infection. To address this gap, we examined integration site distributions in acute, chronic, and ART-treated HIV-1 individuals at compartment and subset resolutions. We found comparable characteristics of integration site distributions, where most sites were found in transcription units. The CD4$^+$ T cell subset, tissue of origin, or the infection stage had at most slight effects on the initial integration site selection based on comparison to genomic and epigenetic annotations. Previously identified genes that

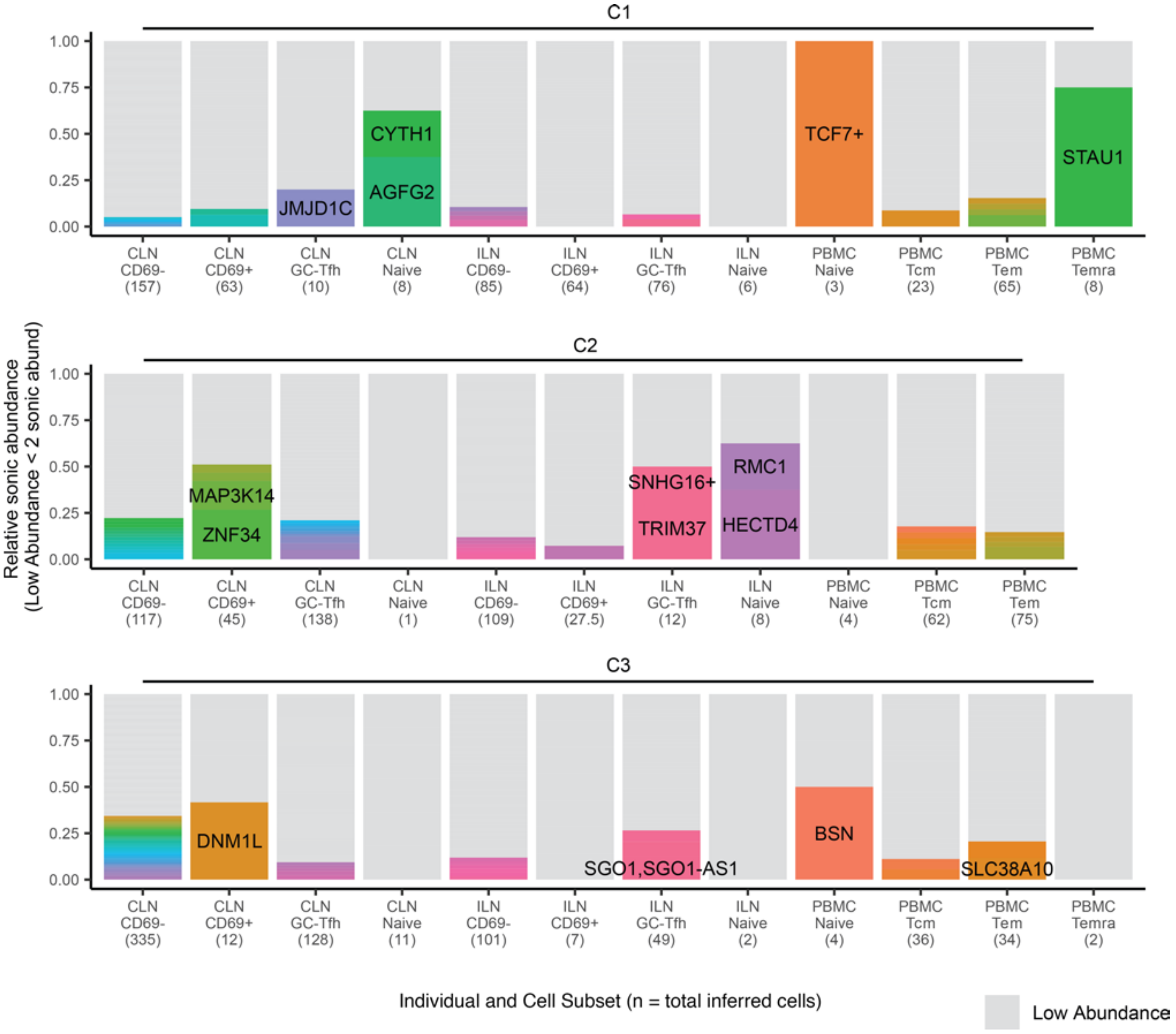

Figure 4. Identification of clonal expansion from individuals during chronic infection. Relative sonic abundance for each integration site is shown for each individual. Sites with color indicate that the site is clonal (defined as a sonic abundance ≥2 for chronic infection). Different colors indicate different integration sites across an individual (i.e., two different integration sites in the same gene with clonal expansion would have different colors). Clones consisting of 10% or more of the total relative sonic abundance are labeled with the integrated gene or nearest gene. Symbols appear after gene names as detailed in the legend for Figure 3.

are commonly detected in infected cells were also found among our data set. As expected, clonal expansion was found in most compartments and was more prominent in chronic and ART-treated individuals. Despite these expanded clones, we detected little overlap between subsets and compartments.

The general observation of similar characteristics of HIV-1 integration across subset and compartment supports previous studies documenting the tendencies of HIV-1 integration. While it is known for HIV-1 to prefer sites of active transcription, our observed lack of noticeable transcriptional activity bias by subset could potentially be explained by similar global transcriptomic profiles of CD4$^+$ T cell subsets, irrespective of known differentially expressed subset-defining genes. Despite largely similar characteristics, we were still able to detect some differences in our comparisons. We observed a corresponding increase in frequencies of nonintronic and nondistal-intergenic sites during ART in tissue and resident subsets, suggesting that infected resident cells might have subtle differences compared with nonresident infected cells. Furthermore, the trend in increased detection of clonality in nonresident cell subsets requires more sampling to better understand how residency phenotype may impact HIV-1 reservoirs.

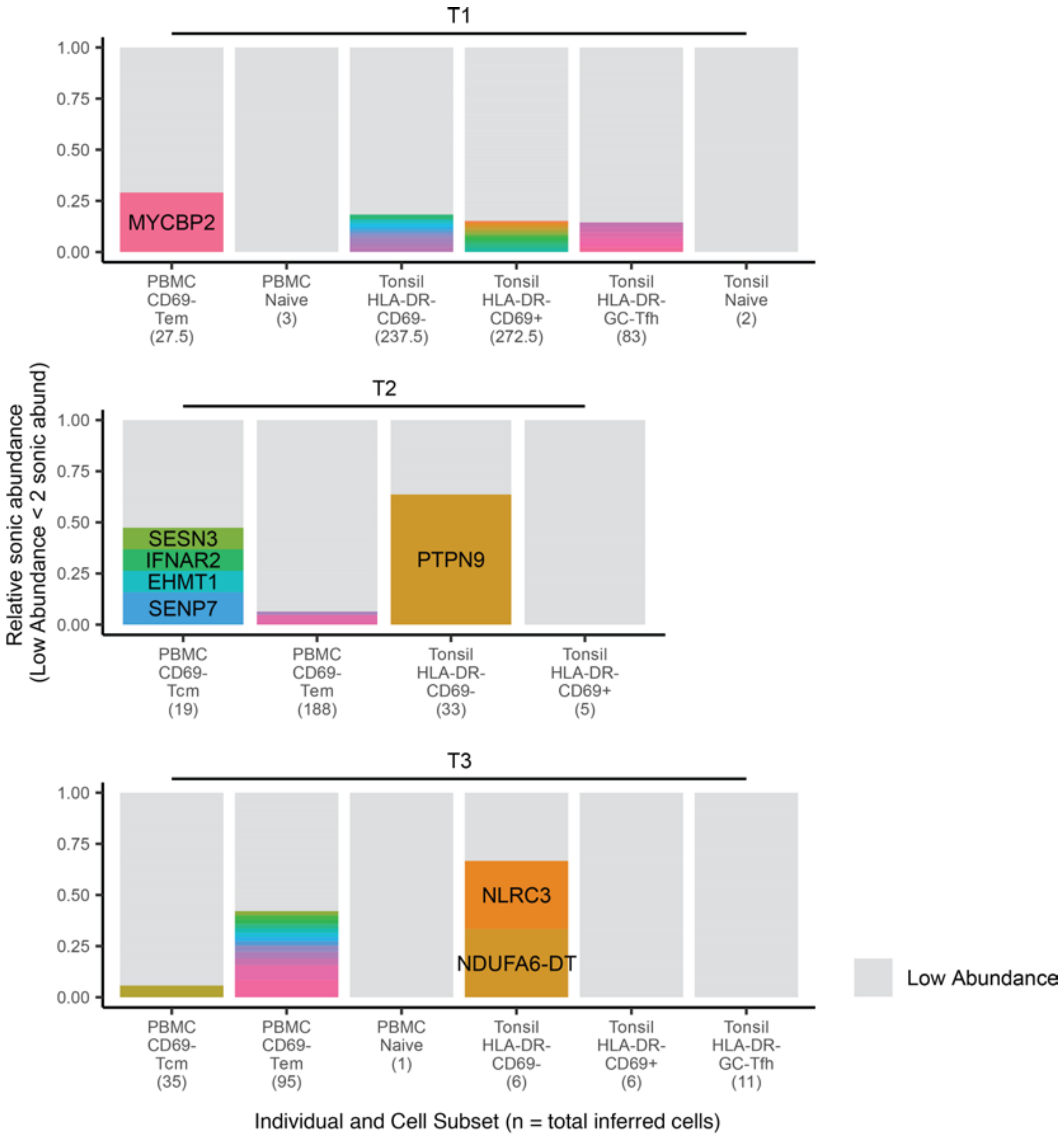

Figure 5. Identification of clonal expansion from ART-treated individuals. Relative sonic abundance for each integration site is shown for each individual. Sites with color indicate that the site is clonal (defined as a sonic abundance ≥2 for ART-treated individuals). Different colors indicate different integration sites across an individual (i.e., two different integration sites in the same gene with clonal expansion would have different colors). Clones consisting of 10% or more of the total relative sonic abundance are labeled with the integrated gene or nearest gene. Symbols appear after gene names as detailed in the legend for Figure 3.

Recent studies have investigated the HIV-1 integration site landscape in bulk CD4$^+$ T cells or LNMCs from lymphoid tissue, finding integration site overlaps between tissue compartments in ART-treated individuals (39, 53). This suggests that trafficking of infected cells may occur between compartments over long-term ART treatment. Other groups have observed identical internal proviral sequences between lymphoid tissue and peripheral blood (24, 27, 39, 53–55). However, the existence of identical proviral sequences between compartments does not indicate clonality, as different cells can harbor the same proviral sequence integrated at different sites in the host genome (39).

The limited overlap between integration sites in different subset and compartments in our data set has potential ramifications. With limited ongoing viral replication in lymphoid tissue during ART (53) and the low likelihood of the same integration site occurring from two separate cellular infection events, the emergence of identical sites in different compartments would require proliferation of the infected cell followed by migration and/or differentiation. This lack of integration site overlap suggests that migration and/or differentiation may impart a fitness cost (whether by immune clearance from viral reactivation, disruption of required cellular/differentiation

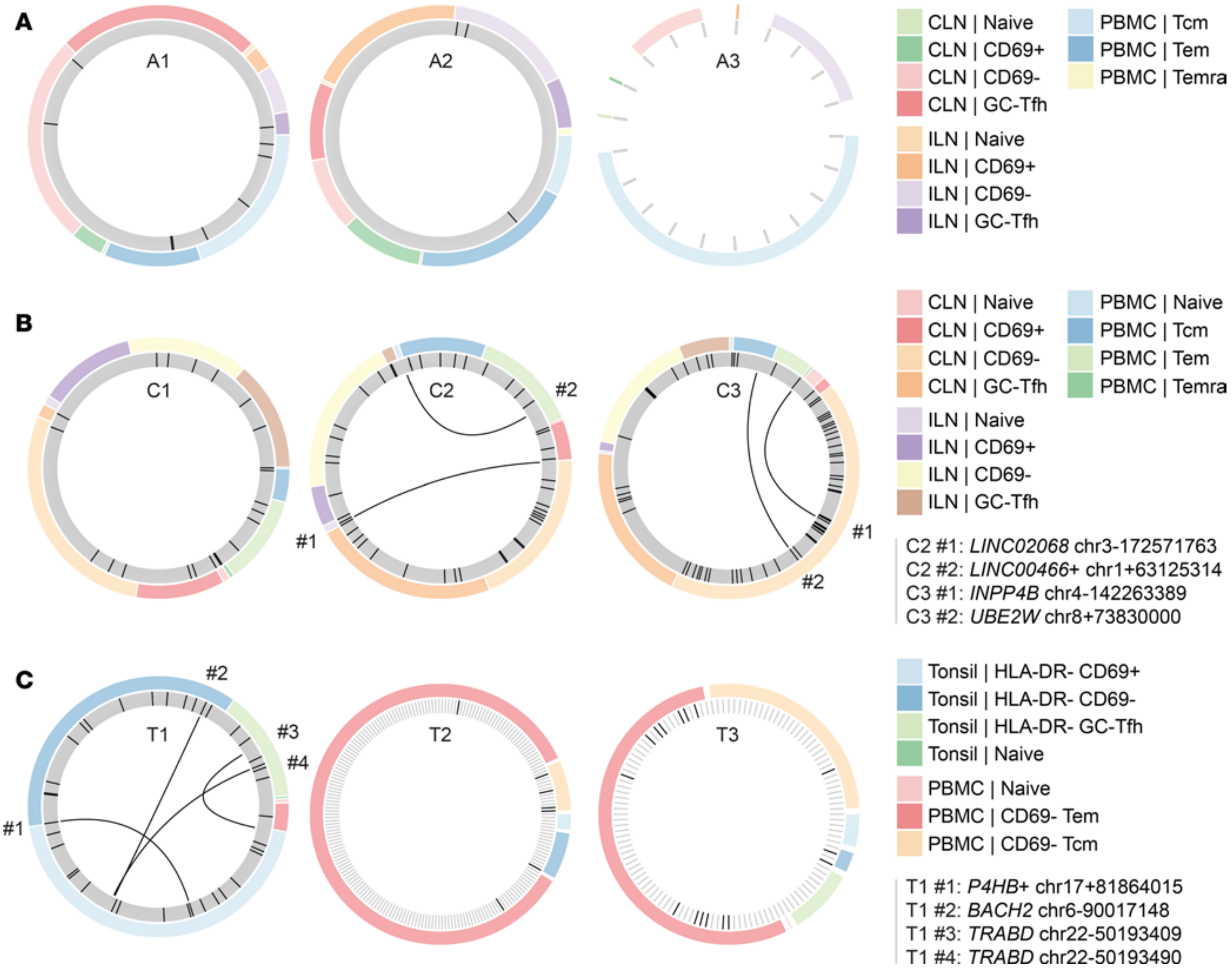

Figure 6. Limited overlap of integration sites detected within cell subsets and compartments. For (**A**) acute, (**B**) chronic, and (**C**) ART-treated states, each plot represents an individual. Each spoke on the inner wheel represents a unique integration site. Spokes with a darker color represent clonally expanded sites. The outer wheel represents the compartment and subset where the integration site was detected. Connecting lines indicate that the same integration site was detected in a different cell subset and is labeled with the gene (or nearest gene) along with genomic location. Symbols appear after gene names as detailed in the legend for Figure 3.

programs due to integration, or other means). This scenario is supported by the observation that the infection of resting memory CD4+ T cells likely stems from an infection event during an activated state prior to the transition to a quiescent state, thus escaping elimination (56). Any exit from quiescence could lead to the presentation of viral antigens and subsequent clearance, resulting in the general maintenance of segregation. The observation of shared proviral sequences from before ART treatment (27) and our limited observation of integration site overlap supports the notion of a genetic bottleneck (likely from immune pressure prior to ART, ref. 39).

Insertional mutagenesis affecting specific cellular genes may be linked to cellular proliferation. Some of the overlapping genes that we detected have been documented for their roles in oncogenesis. *LINC00466* has been implicated as a potential factor in promoting lung adenocarcinoma (57). *BACH2* has known tumor immunosuppressive capacity (58), and *P4HB* upregulation may be involved with glioblastoma multiforme (59) as noticed for integration sites recovered from the ART-treated individuals, thus suggesting that oncogenic genes may play a role in the overlap. Strikingly, these limited overlaps were all within the same compartment or involved the nonresident subsets (i.e., all subsets except tissue CD69+ or GC-Tfh phenotypes), suggesting that migration of infected clones through lymphatic tissues does occur. The dilution of a given infected clone among the vast network of lymphatics and vasculature could explain the small degree of sharing among the nonresident subsets.

Recent studies reported overlap in bulk CD4$^+$ T cells between tonsils and peripheral blood in ART-treated individuals (39, 53), but our data highlighted the lack of overlap in our subjects. The statistical concern of the "unseen species problem" (where the overlapping clones might have been undetected given the size of the pool of infected cells) is always present for these types of studies, especially with our samples from peripheral blood during ART. However, clonal expansion was still detectable from both compartments in our data set. The discrepancy between our findings and these previous studies can potentially be explained by two hypotheses. First, the duration of ART treatment may play a role in the emergence of sharing between blood and tissues. Our study cohort was sampled at a generally shorter duration of ART (1.4, 7.8, and 3.2 years) compared with previous studies (5.7, 13, and <15.8 years) (53). Time may be needed under suppressive therapy to allow cells to differentiate after antigenic stimulation or equilibrate between the subsets and/or compartments. Second, the overlaps between blood and tonsil may exist in the HLA-DR$^+$-activated subset, as we specifically examined resting HLA-DR$^–$ cells for tonsil samples. A recent study showed that there was an increase in infected HLA-DR$^+$ cells over time during ART (26), suggesting that these two hypotheses are not mutually exclusive. Both HLA-DR$^+$ and HLA-DR$^–$ cells are capable of harboring intact proviruses and in some cases have identical p6-RT sequences (26). However, integration site distributions were not assessed, thus leaving the question of true overlap between these subsets unresolved. The presence of HLA-DR and the likelihood of an activation state may suggest the ability for cells to traffic between compartments.

The limited overlap of integration sites reported here in the HIV-1 reservoir has consequences for cure strategies. Latency reversal agents to date have generally been unable to "shock" the entire reservoir, thus limiting the effectiveness of viral elimination. A recent study has demonstrated differential effects in activating HIV-1 RNA transcription via latency reversal agents in diverse CD4$^+$ T cell memory subsets from blood (60). These previous findings, combined with the differences we observed based on residency phenotypes, highlight the need to better understand the HIV-1 reservoir at an increased subset resolution.

Our study has several limitations, with the primary limitation being our sample size. Due to the large size of the HIV-1 reservoir and the possibility of single sites being actual clones, more studies with a variety of lymphoid tissues, including the gut-associated lymphoid tissue (GALT), are needed from individuals in different cohorts to validate our findings. While we did profile from 2 lymph nodes that are distal to each other, this does not preclude the possibility of clonally expanded cells being in the remaining hundreds of other structures distributed throughout the body. However, these samples, especially the multiple paired lymph nodes are difficult to acquire and costly to sequence at the cell subset resolution. Another limitation is the lack of dual integration site and viral DNA sequencing at the time of our study, which was difficult with the number of cells procured from each subset. As studies have documented the existence of replication competent viruses in clonally expanded cells (25, 39, 61), the understanding of which cells in our study have replication competent viruses would add additional resolution. New techniques for paired integration site and proviral profiling could be used for further studies (25, 39). The identification of integration sites, however, is sufficient for determining clonal expansion in infected cells (39).

Overall, our data set provides a resource to the HIV community, as it is the first to our knowledge to profile integration sites from acute, chronic, and ART-treated individuals with paired tissue and blood comparisons at a subset resolution. This resource will be valuable for both the HIV-1 reservoir field as well as for clinical methods using retroviral vectors. Our data highlight the need to better understand the reservoir — in terms of specific subreservoirs that may need to be differentially targeted with a focus on the residency status of infected cells.

Methods

Samples. For acute and chronic infection studies, peripheral blood and tissue samples (CLN and ILN) were simultaneously obtained from HIV-1$^+$ individuals prior to the introduction of ART (n = 3 for acute and n = 3 for chronic). For ART-treated studies, peripheral blood and tonsil samples were simultaneously obtained at the same time point from HIV-1$^+$ individuals undergoing ART treatment (n = 3). Donor information is summarized in Table 1.

PBMCs and LNMCs were prepared as previously described (62). All cells were cryopreserved before being shipped on dry ice for experimental studies.

CD4$^+$ T cell subset sorting. PBMCs and LNMCs were thawed and suspended at 2 million cells/mL in complete RPMI medium supplemented with 10% FBS, 1% L-glutamine, and 1% penicillin/streptomycin (R10). Cells were rested overnight in a humidified incubator with 5% CO_2 at 37°C, and RNase-free DNase I (10 units for every mL media used; MilliporeSigma) was added to reduce cell clumping for sorting. After overnight rests, cells were washed with PBS prior to extracellular staining of cellular markers for FACS on the same day.

Cells were first stained with CCR7 APC-Cy7 (G043H7, BioLegend, 353211) for 10 minutes in a humidified incubator with 5% CO_2 at 37°C. LIVE/DEAD Fixable Aqua (Thermo Fisher Scientific) was then added to the cells for a 10-minute incubation at room temperature. Cells were stained with a master mix of monoclonal antibodies with fluorescent conjugates (antibodies and clones are listed below) for 20 minutes at room temperature. Samples with greater than or equal to 15×10^6 cells were subjected to 2× of the titration volumes for all staining steps. Cells were then washed with PBS before being resuspended in 350 μl phenol red–free RPMI if samples had greater than or equal to 15 million cells. Otherwise, cells were resuspended in 250 μl phenol red–free RPMI. Prior to FACS, the cells were subjected through a 35 μm nylon mesh strainer-top tube (Thermo Fisher Scientific) to reduce cell clumping. All samples were sorted with a FACS Aria II SORP (BD) situated in a biosafety cabinet for biohazardous samples using standard methodology. The gating strategy is described in Supplemental Figures 1 and 2. Individual T3 had a further $CD32^-$ gate on cells from tonsil, as this sample's sort was done from a previous study (63). Cells were sorted into the 1.5 mL DNA LoBind tubes (Eppendorf) prefilled with 200–300 μl of a solution containing 1 part FBS and 1 part phenol red–free RPMI or PBS.

Sorted cells were pelleted in a centrifuge at 3000*g* for 5 minutes, and supernatants were discarded. Cell pellets were resuspended in 200 μl PBS and snap frozen on dry ice before storing at –80°C.

Antibodies. The following monoclonal antibodies were used for FACS: CCR7 APC-Cy7 (G043H7, Biolegend, 353211), CD69 PE (FN50, BD, 555531), CD3 APC R700 (UCHT1, BD, 565119), PD-1 BV421 (EH12.2H7, Biolegend, 329920), HLA-DR BV605 (G46-6, BD, 562845), CXCR5 BB515 (RF8B2, BD, 564624), CD14 PE-Cy5 (61D3, abcam, ab25395), CD16 PE-Cy5 (3G8, Biolegend, 3012010), CD19 PE-Cy5 (HIB19, Biolegend, 302210), CD45RA PE CF594 (HI100, BD, 562298), CD8 PE-Cy5.5 (RPA-T8, eBioscience, 35-0088-42), and CD4 PE Cy7 (RPA-T4, Biolegend, 300512).

Integration site profiling. Sorted cells were thawed for DNA extraction using the DNeasy Blood & Tissue Kit (Qiagen) using standard protocols. The final eluate was reeluted through the column to improve DNA yields, as recommended in the manufacturer's protocol. 100 μl of the DNA eluate was supplemented with 30 μl molecular-grade water (mH_2O) to proceed with library preparation as previously published (37, 38). Briefly, genomic DNA was sonicated using a Covaris M220 unit (peak power: 50 W, duty factor: 5%, cycles/burst: 200, treatment time: 60 seconds, water temperature: 25°C) to generate approximately 800–900 bp fragments. A genomic uninfected DNA control (purified from 293T cells; Bushman Lab) was added during the sonication stage. Fragmented DNA was cleaned using AMPure XP beads (Beckman Coulter) (ratio of 0.7 beads/1 DNA volume) and resuspended in 40 μl mH_2O.

DNA ends were repaired using NEBNext Ultra End Repair/dA-Tailing Module (New England Biolabs) and incubated at 20°C for 30 minutes, 65°C for 30 minutes, and rested at 4°C indefinitely. A no-template control of mH_2O was added during the end-repair stage. End-repaired fragmented DNA was then ligated with unique linkers using the NEBNext Ultra Ligation Module (New England Biolabs) for 16 hours at 16°C and rested at 4°C indefinitely (38). The resulting product was then bead purified using AMPure XP beads (0.7 beads/1 DNA volume) and eluted in 60 μl of mH_2O.

The first round of nested PCR (PCR1) was performed in replicates of 4 per sample using the Advantage 2 PCR kit (Takara Bio). 15 μl DNA from the previous step was used for each replicate. U3 and U5 primers targeting the HIV-1 LTR region primers and unique linker primers were added at a final concentration of 300 nM each. Final volume was adjusted to 25 μl per replicate with mH_2O. Thermocycler settings were set for 1 cycle of 95°C for 1 minute; a linear amplification of 5 cycles of 95°C for 30 seconds plus 70°C for 1 minute and 30 seconds; a log amplification of 20 cycles of 95°C for 30 seconds plus 67°C for 1 minute and 30 seconds; a final extension cycle of 70°C for 4 minutes; and an indefinite hold at 4°C. 5 μl PCR1 product was run on a 1% agarose-TAE gel to check for smears.

The second round of nested PCR (PCR2) was performed on 2 μl PCR1 product for each replicate. Samples at this stage were separated into 2 different reactions, one containing U3 primers and the other containing U5 primers. All LTR primers contained sequencing adaptors as well as a unique barcode with the goal of all samples (including replicates) having a unique LTR-associated barcode and linker combination. PCR2 linker primers and U3 or U5 primers were added at a final concentration of 300 nM each. No internal fragment blocking oligo was added. Final volume was adjusted to 25 μl with mH_2O. Thermocycler settings were set for 1 cycle of 95°C for 1 minute; a linear amplification of 5 cycles of 95°C for 30 seconds plus 70°C for 1 minute and 30 seconds; a log amplification of 15 cycles of 95°C for 30 seconds plus 67°C for 1 minute and 30 seconds; a final extension cycle of 70°C for 4 minutes; and an indefinite hold at 4°C. 5 μl PCR2 product was run on a 1% agarose-TAE gel to check for smears.

All PCR2 replicates were pooled together (U3 and U5 products are kept separate). PCR2 replicate pools were bead purified with AMPure XP beads (0.7 beads/1 DNA volume) and eluted in 40 μl mH_2O. Libraries were then assessed for molarity of readable amplicons using quantitative PCR with the 2X MasterMix, ROX Low kit (Roche). Average fragment size was determined with an Agilent 2200 TapeStation using a High Sensitivity D1000 ScreenTape Assay.

Samples were pooled using the calculated readable molarity to attempt to equalize molarities across samples by using the Microsoft Excel Solver add-in. 1 μl each of uninfected and no-template controls were added to this pool. Bead purification with AMPure XP beads (0.7 beads/1 DNA volume) was performed as needed to get a final molarity of approximately 2 nM or greater. Final readable molarity concentrations were determined for the pooled library in addition to an average fragment size determination. Samples were then prepared for paired-end sequencing using a 300 cycle MiSeq v2 Reagent Kit (Illumina) with 30% PhiX spike-in and custom primers on the MiSeq platform (Illumina).

Primers. All primers are displayed in a 5′–3′ direction and were manufactured by Integrated DNA Technologies. Linker-associated primers and barcodes are from a previous study (38). The following primers were used: HIV-1 PCR1 (U3): CCCTGGCCCTGGTGTGTAGTTCTG; HIV-1 PCR1 (U5), GAACCCACTGCTTAAGCCTCAATAAAG; HIV-1 PCR2 (U3), CAAGCAGAAGACGGCATACGAGAT**BARCODE**AGTCAGTCAGCCCAGGGAAGTAGCCTTGTGTGTGGT; HIV-1 PCR2 (U5), CAAGCAGAAGACGGCATACGAGAT**BARCODE**AGTCAGTCAGCCCAAGTAGTGTGTGCCCGTCTGTTG; sequencing R1, ATCTACACCAGGACTGACGCTATGGTAATTGT; sequencing index (U3), CAAGGCTACTTCCCTGGGCTGACTGACT; sequencing index (U5), AGTCAGTCAGCCCAGGGAAGTAGCCTTG; sequencing R2 (U3), GGGCACACACTACTTGGGCTGACTGACT; and sequencing R2 (U5), AGTCAGTCAGCCCAAGTAGTGTGTGCCC.

Read to site mapping and abundance scoring. The processing of raw reads to final integration site counts was done using the cHIVa pipeline (https://github.com/cnobles/chiva; branch, master; commit ID, d43d01ebc577bb626dbdcda2eddefc9dc7387469), which is a streamlined and updated version of the original INSPIIRED pipeline (https://github.com/BushmanLab/INSPIIRED; branch, master; commit ID, 35e4b0b06182e3dcdb15f2abb6dbaab45b0ac225) (37). Briefly, raw bcl files from sequencing were converted into fastq files using bcl2fastq2 (Illumina). The fastq files are demultiplexed, trimmed, filtered, consolidated, aligned, and further processed with crossover filtering to get human (hg38) and viral integration site junctions by sample. Sonic abundance was then calculated as previously described by counting the unique fragment lengths for a given integration site to infer the number of infected cells containing the site of interest (37, 38). Since U3 and U5 were multiplexed together, the average of U3 and U5 sonic abundances for each original PCR replicate was used if there were both U3 and U5 reads detected with the same integration site. Additional R code (https://github.com/wuv21/intsite_analysis; branch, master; commit ID, 2d770df836b556cb841caef0d1124c1aeed5c058) was used to process and combine data from different runs for graphs and summary information. The ChIPSeeker package in R was used to annotate sites based on genomic location (64). Demultiplexed sequencing files are deposited in the Sequence Read Archive database (accession PRJNA655671).

RNA-Seq analysis. Bins of genes with increasing transcriptional activity were derived from published RNA-Seq data sets from GEO series GSE130793 (42). GC-Tfh samples (GSM3753986, GSM3753990, and GSM3753994) and non-Tfh (GSM3753987, GSM3753991, and GSM3753995) from lymph nodes were used. The raw counts were read using the R statistical language with the DESeq2 package (65). After normalization and scaling with default parameters, the transcriptional levels for each gene were then placed into 10 equal bins. The levels were determined by the samples within each subset (GC-Tfh or non-Tfh). Integration sites were then filtered from our data set to select for those residing in a gene. Sites were pooled together based on subset (as indicated in Supplemental Figure 8). Since genes may overlap, we included all of the overlapping genes in the comparison. The proportion of our recovered genic integration sites within each transcriptional bin was counted and plotted.

Genomic and epigenetic annotation heatmaps. Integration sites were grouped based on criteria of interest. Groups of sites were only included if there were 30 or more sites in preparation for site annotations as previously detailed (46–48). Briefly, 3 matched random control (MRC) sites were chosen for each integration site in the filtered data set. The actual and MRC sites were then marked for different genomic and $CD4^+$

T cell epigenetic annotations in order to compute a ROC curve. Significance for the proportions between actual integration sites and MRC sites was found using a χ^2 test.

Pooled overlap sampling. For each individual, all integration sites were pooled based on their sonic abundance. The sites were then randomly sampled by the original sonic abundance of a given cell subset to preserve our original sequencing depth and breadth. If a site was found in 2 or more of these simulated cell subsets, the site would be recorded as being a simulated overlap. This simulation was repeated 10,000 times, and the number of simulated overlaps detected was plotted as a cumulative distribution frequency plot.

Statistics. Statistical analyses are detailed in the Methods and figures legends and are consolidated here. Differences between genic annotations were assessed using a Wilcoxon rank sum test. For the genomic/epigenetic heatmaps, significance between actual integration sites and MRCs was assessed using a χ^2 test. Clonal proportions by residency phenotype and by stage were assessed using a Wilcoxon rank sum test. P values below 0.05 were deemed significant.

Study approval. All donors were recruited by the Department of Infectious Diseases at the National Institute for Respiratory Diseases (INER-CIENI) in Mexico City, Mexico. All donors provided written informed consent in compliance with protocols set forth by the INER-CIENI Ethics Committee and the Institutional Review Board at the University of Pennsylvania. Protocols were approved by the Institutional Review Boards at the University of Pennsylvania (no. 809316) and Comité de Ciencia y Bioética en Investigación (Committee for Science and Bioethics in Research) from INER (Mexico City, Mexico; study no. B03-16).

Author contributions

VHW, FDB, and MRB conceptualized experiments, analyzed data, and wrote the manuscript. PMDRE, MGN, SAR, and GRT provided the cells for the study. VHW and SN conducted the cell culture before and after sorting. LKC performed the flow cytometry sorting. FTR performed the surgeries to acquire inguinal lymph nodes. VHW and SN performed the integration site assay. VHW, CLN, KM, and JKE analyzed the sequencing data and performed downstream computational analyses.

Acknowledgments

We thank the individuals who donated samples to our study. We would also like to thank members of the MRB and FDB labs for their assistance, especially Scott Sherrill-Mix for very helpful discussions regarding statistical modeling. MRB was supported by the NIH (R01-AI118694) and BEAT-HIV Delaney Collaboratory (UM1-AI126620) grants. FDB was supported by the NIH (U19-AI117950 and R01-AI129661). Both MRB and FDB are further supported by the Penn Center for AIDS Research (P30-AI045008). VHW is supported by NIH grant (T32-AI007632).

Address correspondence to: Michael R. Betts, Department of Microbiology, 402c Johnson Pavilion, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. Phone: 215.573.2773; Email: betts@pennmedicine.upenn.edu.

1. Siliciano JD, et al. Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4+ T cells. *Nat Med*. 2003;9(6):727–728.
2. Crooks AM, et al. Precise quantitation of the latent HIV-1 reservoir: implications for eradication strategies. *J Infect Dis*. 2015;212(9):1361–1365.
3. Wong JK, et al. Reduction of HIV-1 in blood and lymph nodes following potent antiretroviral therapy and the virologic correlates of treatment failure. *Proc Natl Acad Sci USA*. 1997;94(23):12574–12579.
4. Ciuffi A, et al. A role for LEDGF/p75 in targeting HIV DNA integration. *Nat Med*. 2005;11(12):1287–1289.
5. Van Maele B, Busschots K, Vandekerckhove L, Christ F, Debyser Z. Cellular co-factors of HIV-1 integration. *Trends Biochem Sci*. 2006;31(2):98–105.
6. Engelman AN, Singh PK. Cellular and molecular mechanisms of HIV-1 integration targeting. *Cell Mol Life Sci*. 2018;75(14):2491–2507.
7. Cherepanov P, et al. HIV-1 integrase forms stable tetramers and associates with LEDGF/p75 protein in human cells. *J Biol Chem*. 2003;278(1):372–381.
8. Maertens G, et al. LEDGF/p75 is essential for nuclear and chromosomal targeting of HIV-1 integrase in human cells. *J Biol Chem*. 2003;278(35):33528–33539.
9. Han Y, et al. Resting CD4+ T cells from human immunodeficiency virus type 1 (HIV-1)-infected individuals carry integrated HIV-1 genomes within actively transcribed host genes. *J Virol*. 2004;78(12):6122–6133.
10. Ikeda T, Shibata J, Yoshimura K, Koito A, Matsushita S. Recurrent HIV-1 integration at the BACH2 locus in resting CD4+ T cell populations during effective highly active antiretroviral therapy. *J Infect Dis*. 2007;195(5):716–725.
11. Wagner TA, et al. HIV latency. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection.

Science. 2014;345(6196):570–573.

12. Cohn LB, et al. HIV-1 integration landscape during latent and active infection. *Cell*. 2015;160(3):420–432.
13. Maldarelli F, et al. HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science*. 2014;345(6193):179–183.
14. Coffin JM, et al. Clones of infected cells arise early in HIV-infected individuals. *JCI Insight*. 2019;4(12):128432.
15. Pantaleo G, et al. HIV infection is active and progressive in lymphoid tissue during the clinically latent stage of disease. *Nature*. 1993;362(6418):355–358.
16. Deleage C, et al. Defining HIV and SIV reservoirs in lymphoid tissues. *Pathog Immun*. 2016;1(1):68–106.
17. Estes JD, et al. Defining total-body AIDS-virus burden with implications for curative strategies. *Nat Med*. 2017;23(11):1271–1276.
18. Bucy RP, et al. Initial increase in blood CD4(+) lymphocytes after HIV antiretroviral therapy reflects redistribution from lymphoid tissues. *J Clin Invest*. 1999;103(10):1391–1398.
19. Chun TW, et al. Persistence of HIV in gut-associated lymphoid tissue despite long-term antiretroviral therapy. *J Infect Dis*. 2008;197(5):714–720.
20. Miles B, et al. Follicular regulatory T cells impair follicular T helper cells in HIV and SIV infection. *Nat Commun*. 2015;6:8608.
21. Banga R, et al. PD-1(+) and follicular helper T cells are responsible for persistent HIV-1 transcription in treated aviremic individuals. *Nat Med*. 2016;22(7):754–761.
22. Buggert M, et al. Limited immune surveillance in lymphoid tissue by cytolytic CD4+ T cells during health and HIV disease. *PLoS Pathog*. 2018;14(4):e1006973.
23. Jones BR, et al. Genetic diversity, compartmentalization, and age of HIV proviruses persisting in CD4$^+$ T cell subsets during long-term combination antiretroviral therapy. *J Virol*. 2020;94(5):e01786-19.
24. De Scheerder MA, et al. HIV rebound is predominantly fueled by genetically identical viral expansions from diverse reservoirs. *Cell Host Microbe*. 2019;26(3):347–358.e7.
25. Einkauf KB, et al. Intact HIV-1 proviruses accumulate at distinct chromosomal positions during prolonged antiretroviral therapy. *J Clin Invest*. 2019;129(3):988–998.
26. Lee E, et al. Memory CD4 + T-cells expressing HLA-DR contribute to HIV persistence during prolonged antiretroviral therapy. *Front Microbiol*. 2019;10:2214.
27. Lee E, et al. Impact of antiretroviral therapy duration on HIV-1 Infection of T cells within anatomic sites. *J Virol*. 2020;94(3):e01270-19.
28. Beura LK, et al. CD4$^+$ resident memory T cells dominate immunosurveillance and orchestrate local recall responses. *J Exp Med*. 2019;216(5):1214–1229.
29. Cantero-Pérez J, et al. Resident memory T cells are a cellular reservoir for HIV in the cervical mucosa. *Nat Commun*. 2019;10(1):4739.
30. Moriya N, Sanjoh K, Yokoyama S, Hayashi T. Mechanisms of HLA-DR antigen expression in phytohemagglutinin-activated T cells in man. Requirement of T cell recognition of self HLA-DR antigen expressed on the surface of monocytes. *J Immunol*. 1987;139(10):3281–3286.
31. Chun TW, Finzi D, Margolick J, Chadwick K, Schwartz D, Siliciano RF. In vivo fate of HIV-1-infected T cells: quantitative analysis of the transition to stable latency. *Nat Med*. 1995;1(12):1284–1290.
32. Finzi D, et al. Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science*. 1997;278(5341):1295–1300.
33. Chun TW, et al. Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. *Nature*. 1997;387(6629):183–188.
34. Testi R, Phillips JH, Lanier LL. Leu 23 induction as an early marker of functional CD3/T cell antigen receptor triggering. Requirement for receptor cross-linking, prolonged elevation of intracellular [Ca++] and stimulation of protein kinase C. *J Immunol*. 1989;142(6):1854–1860.
35. Yokoyama WM, et al. Characterization of a cell surface-expressed disulfide-linked dimer involved in murine T cell activation. *J Immunol*. 1988;141(2):369–376.
36. Shiow LR, et al. CD69 acts downstream of interferon-alpha/beta to inhibit S1P1 and lymphocyte egress from lymphoid organs. *Nature*. 2006;440(7083):540–544.
37. Berry CC, et al. INSPIIRED: Quantification and visualization tools for analyzing integration site distributions. *Mol Ther Methods Clin Dev*. 2017;4:17–26.
38. Sherman E, et al. INSPIIRED: A pipeline for quantitative analysis of sites of new DNA integration in cellular genomes. *Mol Ther Methods Clin Dev*. 2017;4:39–49.
39. Patro SC, et al. Combined HIV-1 sequence and integration site analysis informs viral dynamics and allows reconstruction of replicating viral ancestors. *Proc Natl Acad Sci USA*. 2019;116(51):25891–25899.
40. Haworth KG, Schefter LE, Norgaard ZK, Ironside C, Adair JE, Kiem HP. HIV infection results in clonal expansions containing integrations within pathogenesis-related biological pathways. *JCI Insight*. 2018;3(13):99127.
41. Zhou Y, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun*. 2019;10(1):1523.
42. Vella LA, et al. T follicular helper cells in human efferent lymph retain lymphoid characteristics. *J Clin Invest*. 2019;129(8):3185–3200.
43. 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. *EMBO J*. 2001;20(7):1726–1738.
44. Schröder AR, Shinn P, Chen H, Berry C, Ecker JR, Bushman F. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell*. 2002;110(4):521–529.
45. Mitchell RS, et al. Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol*. 2004;2(8):E234.
46. Berry C, Hannenhalli S, Leipzig J, Bushman FD. Selection of target sites for mobile DNA integration in the human genome. *PLoS Comput Biol*. 2006;2(11):e157.
47. Nobles CL, et al. CD19-targeting CAR T cell immunotherapy outcomes correlate with genomic modification by vector integration.

J Clin Invest. 2020;130(2):673–685.
48. 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 Res*. 2007;17(8):1186–1194.
49. Wang Z, et al. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet*. 2008;40(7):897–903.
50. Farooq Z, Banday S, Pandita TK, Altaf M. The many faces of histone H3K79 methylation. *Mutat Res Rev Mutat Res*. 2016;768:46–52.
51. Wagner TA, et al. HIV latency. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. *Science*. 2014;345(6196):570–573.
52. Fraietta JA, et al. Disruption of TET2 promotes the therapeutic efficacy of CD19-targeted T cells. *Nature*. 2018;558(7709):307–312.
53. McManus WR, et al. HIV-1 in lymph nodes is maintained by cellular proliferation during antiretroviral therapy. *J Clin Invest*. 2019;129(11):4629–4642.
54. Vibholm LK, et al. Characterization of intact proviruses in blood and lymph node from HIV-infected individuals undergoing analytical treatment interruption. *J Virol*. 2019;93(8):e01920-18.
55. Kuo HH, et al. Blood and lymph node dissemination of clonal genome-intact human immunodeficiency virus 1 DNA sequences during suppressive antiretroviral therapy. *J Infect Dis*. 2020;222(4):655–660.
56. Shan L, et al. Transcriptional reprogramming during effector-to-memory transition renders CD4$^+$ T cells permissive for latent HIV-1 infection. *Immunity*. 2017;47(4):766–775.e3.
57. Ma T, Hu Y, Guo Y, Yan B. Tumor-promoting activity of long noncoding RNA LINC00466 in lung adenocarcinoma via miR-144-regulated HOXA10 axis. *Am J Pathol*. 2019;189(11):2154–2170.
58. Roychoudhuri R, et al. The transcription factor BACH2 promotes tumor immunosuppression. *J Clin Invest*. 2016;126(2):599–604.
59. Sun S, et al. Endoplasmic reticulum chaperone prolyl 4-hydroxylase, beta polypeptide (P4HB) promotes malignant phenotypes in glioma via MAPK signaling. *Oncotarget*. 2017;8(42):71911–71923.
60. Grau-Expósito J, et al. Latency reversal agents affect differently the latent reservoir present in distinct CD4+ T subpopulations. *PLoS Pathog*. 2019;15(8):e1007991.
61. Simonetti FR, et al. Clonally expanded CD4+ T cells can produce infectious HIV-1 in vivo. *Proc Natl Acad Sci USA*. 2016;113(7):1883–1888.
62. Nguyen S, et al. Elite control of HIV is associated with distinct functional and transcriptional signatures in lymphoid tissue CD8$^+$ T cells. *Sci Transl Med*. 2019;11(523):eaax4077.
63. Abdel-Mohsen M, et al. CD32 is expressed on cells with transcriptionally active HIV but does not enrich for HIV DNA in resting T cells. *Sci Transl Med*. 2018;10(437):eaar6759.
64. Yu G, Wang LG, He QY. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*. 2015;31(14):2382–2383.
65. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550.