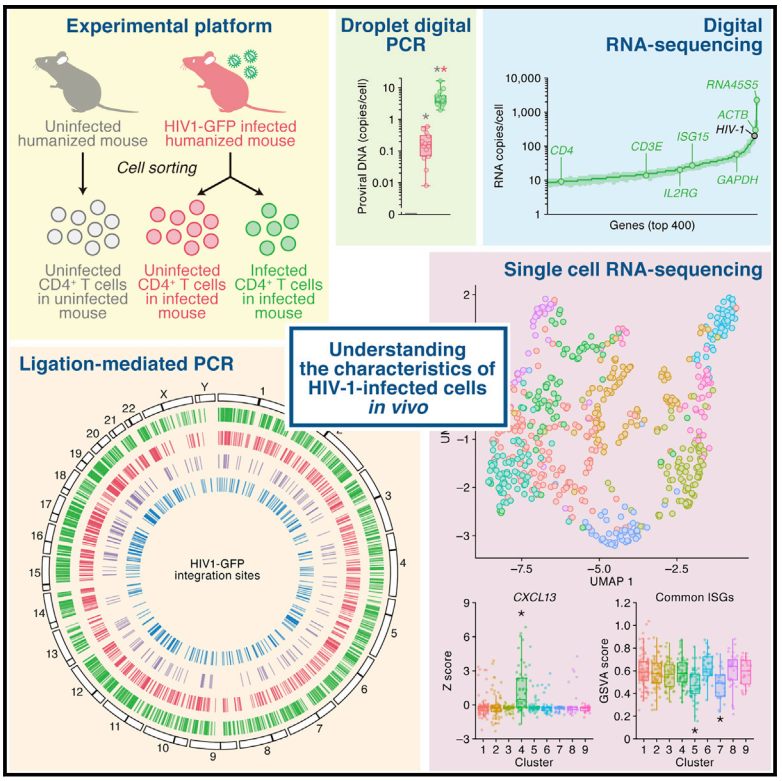
Report

Cell Reports

# Multiomics Investigation Revealing the Characteristics of HIV-1-Infected Cells *In Vivo*

## Graphical Abstract




## Highlights

- HIV-1-producing cells *in vivo* are purified and used for multiomics analyses
- Productively infected cells are characterized by high levels of viral mRNA
- Productive infection is favored by integration in transcriptionally active regions
- CXCL13$^{high}$ and ISG$^{low}$ subpopulations potentially contribute to productive infection

## Authors

Hirofumi Aso, Shumpei Nagaoka, Eiryo Kawakami, ..., Yorifumi Satou, Yoshio Koyanagi, Kei Sato

## Correspondence

ksato@ims.u-tokyo.ac.jp

## In Brief

For eradication of HIV-1 infection, it is important to gain an in-depth understanding of the wide-ranging characteristics of HIV-1-infected cells *in vivo*. Aso et al. purify HIV-1-infected cells from humanized mice and apply "multiomics" techniques to comprehensively characterize the features of HIV-1-infected cell *in vivo*.

Aso et al., 2020, Cell Reports *32*, 107887
July 14, 2020 
https://doi.org/10.1016/j.celrep.2020.107887

CellPress

Cell Reports

CellPress
OPEN ACCESS

Report

# Multiomics Investigation Revealing the Characteristics of HIV-1-Infected Cells *In Vivo*

Hirofumi Aso,[1,2,3,14] Shumpei Nagaoka,[1,14] Eiryo Kawakami,[4,5,14] Jumpei Ito,[1] Saiful Islam,[6,7] Benjy Jek Yang Tan,[6,7] Shinji Nakaoka,[8,9] Koichi Ashizaki,[4] Katsuyuki Shiroguchi,[10,11] Yutaka Suzuki,[12] Yorifumi Satou,[6,7] Yoshio Koyanagi,[2,3] and Kei Sato[1,13,15,*]

[1]Division of Systems Virology, Department of Infectious Disease Control, International Research Center for Infectious Diseases, Institute of Medical Science, The University of Tokyo, Minato-ku, Tokyo 1088639, Japan
[2]Institute for Frontier Life and Medical Sciences, Kyoto University, Kyoto 6068507, Japan
[3]Graduate School of Pharmaceutical Sciences, Kyoto University, Kyoto 6068501, Japan
[4]RIKEN Medical Sciences Innovation Hub Program, Yokohama, Kanagawa 2300045, Japan
[5]Artificial Intelligence Medicine, Graduate School of Medicine, Chiba University, Chiba 2608670, Japan
[6]International Research Center for Medical Sciences, Kumamoto University, Kumamoto 8600811, Japan
[7]Joint Research Center for Human Retrovirus Infection, Kumamoto University, Kumamoto 8600811, Japan
[8]Faculty of Advanced Life Science, Hokkaido University, Sapporo, Hokkaido 0600810, Japan
[9]PRESTO, Japan Science and Technology Agency, Kawaguchi, Saitama 3320012, Japan
[10]RIKEN Center for Biosystems Dynamics Research, Suita, Osaka 5650874, Japan
[11]RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa 2300045, Japan
[12]Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Chiba 2778561, Japan
[13]CREST, Japan Science and Technology Agency, Kawaguchi, Saitama 3320012, Japan
[14]These authors contributed equally
[15]Lead Contact
*Correspondence: ksato@ims.u-tokyo.ac.jp
https://doi.org/10.1016/j.celrep.2020.107887

## SUMMARY

For eradication of HIV-1 infection, it is important to elucidate the detailed features and heterogeneity of HIV-1-infected cells *in vivo*. To reveal multiple characteristics of HIV-1-producing cells *in vivo*, we use a hematopoietic-stem-cell-transplanted humanized mouse model infected with GFP-encoding replication-competent HIV-1. We perform multiomics experiments using recently developed technology to identify the features of HIV-1-infected cells. Genome-wide HIV-1 integration-site analysis reveals that productive HIV-1 infection tends to occur in cells with viral integration into transcriptionally active genomic regions. Bulk transcriptome analysis reveals that a high level of viral mRNA is transcribed in HIV-1-infected cells. Moreover, single-cell transcriptome analysis shows the heterogeneity of HIV-1-infected cells, including *CXCL13*$^{high}$ cells and a subpopulation with low expression of interferon-stimulated genes, which can contribute to efficient viral spread *in vivo*. Our findings describe multiple characteristics of HIV-1-producing cells *in vivo*, which could provide clues for the development of an HIV-1 cure.

## INTRODUCTION

For eradication of HIV-1 infection, it is important to gain an in-depth understanding of the wide-ranging characteristics of HIV-1-infected cells *in vivo*. Recently developed "omics" analyses can be powerful tools to identify the characteristics of HIV-1-infected cells. For instance, genome-wide HIV-1 integration sites (ISs) in infected individuals were investigated, and the result suggested that viral ISs are potentially associated with the destiny of infected cells (e.g., efficient viral production, expansion, latency) (Cohn et al., 2015; Hughes and Coffin, 2016; Maldarelli et al., 2014; Simonetti et al., 2016; Wagner et al., 2014). Transcriptome analyses, such as microarrays and RNA sequencing (RNA-seq), were also performed using samples isolated from HIV-1-infected individuals (Buggert et al., 2018; Cohn et al., 2018; Farhadian et al., 2018; Sedaghat et al., 2008) and showed dynamic changes in gene expression patterns triggered by viral infections. Similarly, we recently performed RNA-seq analysis using splenic human CD4$^+$ T cells from HIV-1-infected humanized mice transplanted with human hematopoietic stem cells (HSCs) and showed that the interferon (IFN)-related immune responses are induced by HIV-1 infection (Yamada et al., 2018). However, it should be noted that a large majority of the CD4$^+$ T cells in infected individuals and animal models are uninfected, and therefore, the transcriptional profiles of

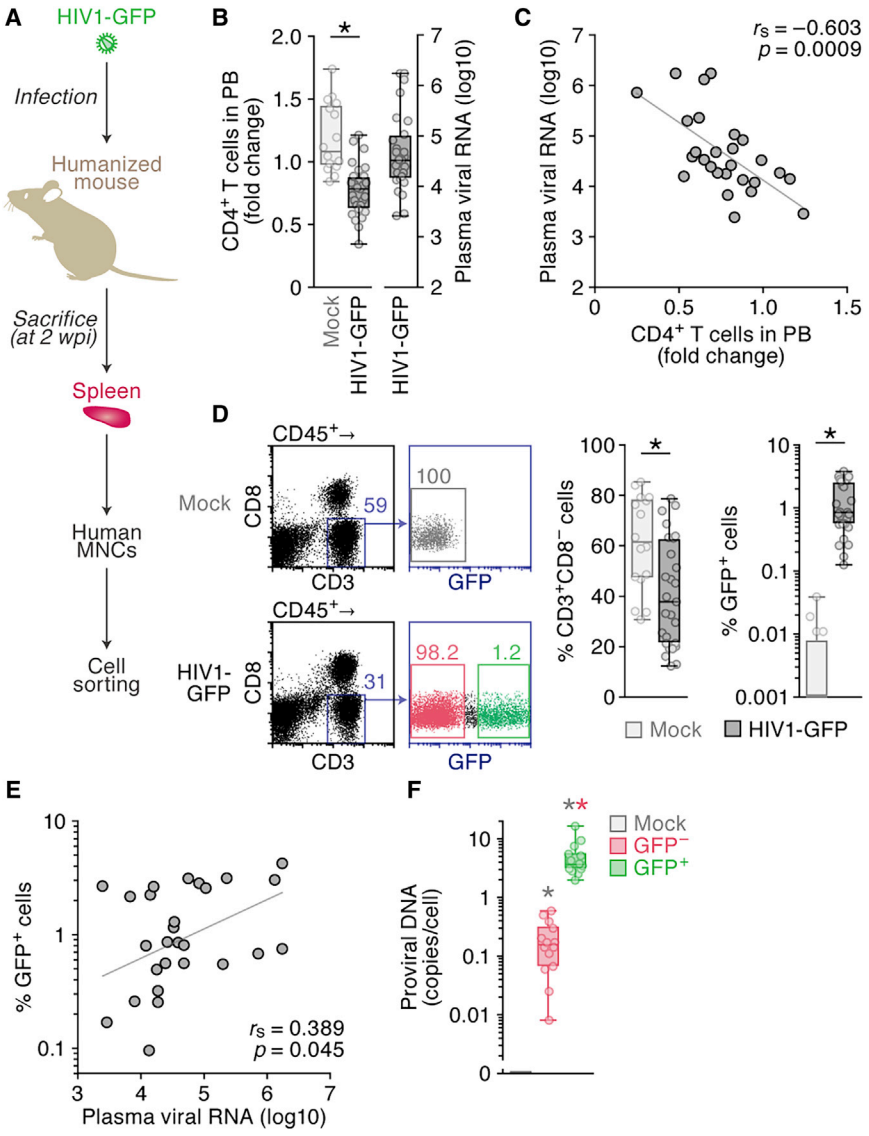



**Figure 1. Characterization of the HIV1-GFP-Infected Humanized Mice**

(A) Experimental setup. The humanized mice were inoculated with HIV1-GFP (n = 27) or RPMI1640 (for mock infection; n = 16). These mice were euthanized and sacrificed at 2 wpi, and the spleens were collected.

(B) Fold change in the level of peripheral $CD4^+$ T cells ($CD45^+CD3^+CD4^+$ cells) (left) and the levels of plasma vRNA (right) analyzed.

(C) Negative correlation between the level of peripheral $CD4^+$ T cells (x axis) and plasma viral RNA (y axis).

(D) Flow cytometry of the splenic human leukocytes. Representative dot plots (left), the percentage of human $CD4^+$ T cells ($CD45^+CD3^+CD8^-$ cells) in splenic human leukocytes ($CD45^+$ cells) (middle), and the percentage of $GFP^+$ cells in human $CD4^+$ T cells (right) are shown. On the left, the numbers in the dot plot indicate the percentage of the gated cells.

(E) Positive correlation between plasma viral RNA (x axis) and the percentage of $GFP^+$ cells in human $CD4^+$ T cells (y axis).

(F) Viral DNA in each population. The absolute copy numbers of viral DNA in the $CD4^+$ T cells of mock-infected mice (mock; n = 6), $GFP^-CD4^+$ T cells of infected mice (n = 8), and $GFP^+CD4^+$ T cells of infected mice (n = 8) were quantified using ddPCR. In (B), (D) and (F), significant differences ($p < 0.05$ by Mann-Whitney U test) compared with mock-infected mice (B and D), uninfected cells (F), or $GFP^-$ cells (F) are indicated by black, gray, or red asterisks, respectively. NS, no statistical significance. See also Figure S1. In (C) and (E), Spearman's rank correlation coefficient ($r_s$) is applied to determine statistically significant correlations.

bulk $CD4^+$ T cells *in vivo* do not reflect that of pure HIV-1-producing cells.

To purify HIV-1-infected cells and reveal their features, previous studies used recombinant viruses encoding fluorescent proteins (e.g., GFP) as markers in an *in vitro* cell culture system and sorted the fluorescent protein-positive cells as the infected cell population (Battivelli et al., 2018; Bradley et al., 2018; Golumbeanu et al., 2018; Kuo et al., 2018). However, for efficient HIV-1 replication *in vitro*, human peripheral $CD4^+$ T cells should be artificially activated with mitogens. With regard to the cellular condition, only a small portion of $CD4^+$ T cells are activated *in vivo* (Janeway et al., 2005; Yamada et al., 2015), and therefore, the expression patterns of cellular genes in HIV-1-infected cells *in vitro* (i.e., artificially activated with mitogens) might be different from those *in vivo*. For better elucidation of HIV-1-infected cells, it is important to reveal the characteristics of HIV-1-infected cells *in vivo*.

Here we use a human HSC-transplanted humanized mouse model (Sato et al., 2013; Yamada et al., 2015) with a replication-competent recombinant HIV-1 expressing GFP (designated HIV1-GFP) (Miura et al., 2001). By performing multiomics analyses, we characterize multiple features of HIV-1-producing cells *in vivo* in terms of genomics and transcriptomics.

## RESULTS

### Characteristics of HIV1-GFP Infection in Humanized Mice

To investigate global characteristics of HIV-1-infected cells *in vivo*, we used an HSC-transplanted humanized mouse model. Additionally, to obtain HIV-1-producing cells *in vivo*, we used HIV1-GFP, encoding the *GFP* gene in the viral genome (Figure S1A). HIV1-GFP is replication competent and expresses all viral proteins, similar to the parental wild-type virus (Figures S1B and S1C). We inoculated HIV1-GFP into the humanized mice and sacrificed the mice at 2 weeks postinfection (wpi) (Figure 1A). A significant decrease in the peripheral blood (PB) $CD4^+$ T cells

CellPress
OPEN ACCESS

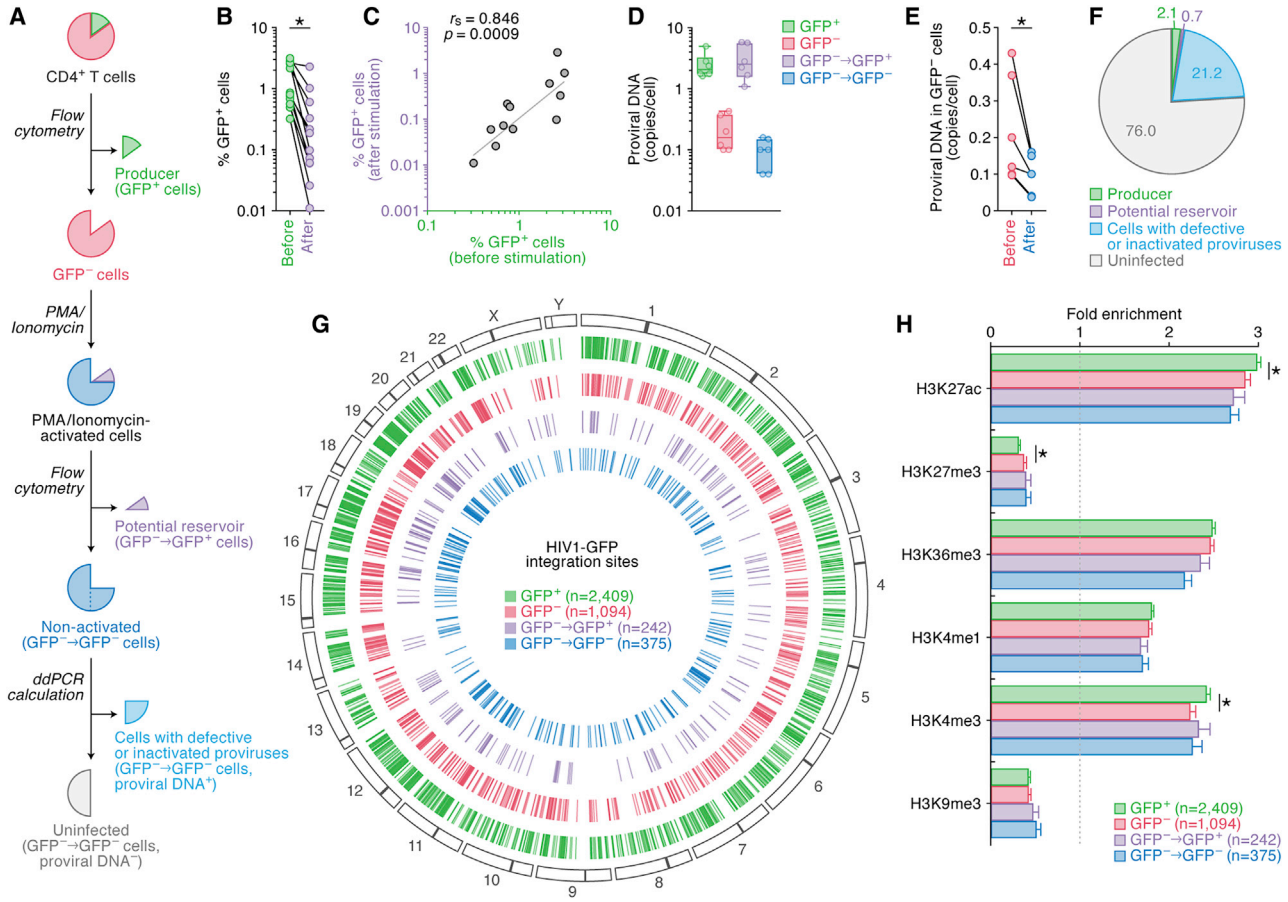



**Figure 2. Composition of the Infected Cells *In Vivo* and the Global Landscape of HIV-1 ISs**

(A) Scheme of the classification of CD4+ T cells in infected mice.

(B) Percentage of the GFP+ cells in human CD4+ T cells (CD45+CD3+CD8− cells; n = 12) before (green) and after (purple) PMA+I-mediated stimulation.

(C) Positive correlation of the percentage of GFP+ cells before (x axis; identical to that at sacrifice) and after (y axis) PMA+I-mediated stimulation.

(D) Viral DNA in each population. The absolute copy numbers of viral DNA in each population were quantified using ddPCR.

(E) Copy number of viral DNA in the GFP− cells before (red) and after (blue) PMA+I-mediated stimulation.

(F) Pie chart of the proportion of each population of CD4+ T cells in infected mice.

(G) Circos plot showing the genomic distribution of the HIV-1 ISs in each population. The number in parenthesis indicates the number of unique ISs detected.

(H) Enrichment of the HIV-1 ISs in the genomic regions with specific histone markers.

In (B), (E) and (H), significant differences ($p < 0.05$ by paired t test for B and E, $p < 0.05$ by Fisher's exact test for H) are indicated by asterisks. See also Figure S2. In (C), Spearman's rank correlation coefficient ($r_s$) is applied to determine a statistically significant correlation.

(Figure 1B, left) as well as a high level of viremia in the plasma were observed in HIV1-GFP-infected mice (Figure 1B, right). The levels of peripheral CD4+ T cells and the plasma viral load were significantly and negatively correlated (Figure 1C). Additionally, a significant decrease in the splenic CD4+ T cells (Figure 1D, middle) caused by HIV1-GFP infection was detected, and importantly, GFP+ cells were specifically detected in infected mice (Figure 1D, right). Moreover, the percentage of GFP+ cells was significantly correlated with the plasma viral load (Figure 1E).

We then sorted the three cell populations (i.e., CD4+ T cells from uninfected mice and GFP+ and GFP− CD4+ T cells) from infected mice, using a cell sorter (Figure S1D). Droplet digital PCR (ddPCR), which can quantify the absolute copy number of HIV-1 DNA in certain cell population even at a very low level, revealed that viral DNA was present in not only GFP+ cells (2.42 ± 0.39 copies per cell) but also GFP− cells (0.20 ± 0.05 copies per cell) at statistically significant levels (Figure 1F).

## Proportion of the Infected Cells with Different Properties *In Vivo*

The results shown in Figure 1F suggest the presence of non-virus-producing infected cells, including potential reservoirs, which can produce virions after activation stimulation, in the GFP− population (Figure 2A). To induce the reactivation of potential reservoirs, we stimulated the sorted GFP− cells with phorbol 12-myristate 13-acetate and ionomycin (PMA+I) *ex vivo*. Two days after stimulation, the stimulated cells were sorted into potential reservoir cells (i.e., GFP+ cells after stimulation) and other cells (Figure 2A). As shown in Figure 2B, 0.47% ± 0.24% of the GFP− cells became positive for GFP after stimulation,

suggesting that potential reservoirs were reactivated by PMA+I. Given that the percentages of GFP$^+$ cells before and after stimulation were significantly correlated (Figure 2C), the size of potential reservoir may closely associate with the level of viral replication *in vivo*. Then, the absolute copy numbers of viral DNA in GFP$^+$ (potential reservoir) and GFP$^-$ (i.e., the mixture of uninfected cells and the cells with defective or inactivated proviral DNA) cells either before or after PMA+I-mediated stimulation were quantified using ddPCR. Approximately 10% of the GFP$^-$ cells after PMA+I-mediated stimulation contained defective or inactivated proviruses (Figure 2D), but the proviral copy number in the GFP$^-$ cells after *ex vivo* stimulation was significantly lower than that before stimulation (Figure 2E). Assuming that multiple infection is ignorable, we could calculate the proportion of cells with defective or inactivated proviruses in non-activated GFP$^-$ cells (Figure 2A). On average, 21.2% ± 5.6% of the splenic CD4$^+$ T cells of infected mice were composed of cells with defective or inactivated proviruses (Figure 2F; see also Figure S2A), while 76.0% ± 6.4% were uninfected.

### HIV-1 Integration Landscape *In Vivo*

Previous works reported that HIV-1 is preferentially integrated into transcriptionally active chromosomes (Han et al., 2004; Schröder et al., 2002; Wang et al., 2007) and that the HIV-1 ISs affect viral production and reactivation (Jordan et al., 2003). To address this issue, we investigated the genomic landscape of the HIV-1 ISs in these cell populations by ligation-mediated PCR (LM-PCR) (Figure 2G). Consistent with a previous finding (Wang et al., 2007), integrative analysis with public chromatin immunoprecipitation sequencing data of histone modifications in CD4$^+$ T cells revealed that the ISs of infected mice were enriched in the genomic regions with active histone modifications (H3K27ac, H3K36me3, and H3K4me1/3) but were depleted in the regions with suppressive histone modifications (H3K27me3 and H3K9me3) (Figure 2H). These results suggest that viral integration in the region close to active histone modification is partly associated with productive viral infection. However, the frequencies of the sense and antisense proviruses in the gene bodies of both the GFP$^+$ cells and the GFP$^-$ cells were comparable (Figure S2B).

### Difference in the Global Transcriptome between the GFP$^+$ and GFP$^-$ Cells *In Vivo*

We next performed global transcriptomic analyses using the samples obtained from the infected humanized mice. However, because the numbers of the cells obtained from humanized mice, particularly those of GFP$^+$ cells, were quite low (Figure 1D), it is technically difficult to investigate the global transcriptome by conventional RNA-seq. To address this issue, we used a technique called digital RNA-seq (dRNA-seq), which can accurately measure the number of RNA molecules genome wide, even from a small number of cells due to low noise and bias during the amplification and sequencing processes. Although viral DNA was detectable in the GFP$^-$ cells (Figure 1F), dRNA-seq revealed that the HIV-1 transcript was undetectable in the CD4$^+$ T cells of mock-infected mice and the GFP$^-$ cells of infected mice, while a high level of viral RNA was detected in GFP$^+$ cells (226.1 ± 67.1 copies per cell; Figure 3A). Notably, 0.83% of the total transcripts in GFP$^+$ cells were HIV-1 RNA, which represented the top fifth of all transcripts and was higher than the transcript level of *GAPDH*, a housekeeping gene (Figure 3B).

In principal component analysis based on the differentially expressed genes (DEGs; listed in Table S1), the transcriptomic data were separated according to GFP expression and infection status (Figure 3C): 98 and 19 genes were differentially expressed in the GFP$^+$ and GFP$^-$ cells, respectively, compared with the CD4$^+$ T cells of uninfected mice (Figure 3D, top). Consistent with our previous reports (Sato et al., 2014; Yamada et al., 2018), IFN-stimulated genes (ISGs) were upregulated in both the GFP$^+$ and GFP$^-$ cells (Table S1). Seventeen genes were differentially expressed between the GFP$^+$ and GFP$^-$ cells (Figure 3D, bottom). Eleven genes were downregulated and six genes were upregulated in the GFP$^+$ cells compared with the GFP$^-$ cells (Figure 3E). A previous study showed that *NEAT1* is upregulated by HIV-1 infection (Zhang et al., 2013), and our dRNA-seq results confirmed these findings. We then performed Gene Ontology (GO) enrichment analysis of the DEGs. Interestingly, no common GO processes were upregulated in either the GFP$^+$ or GFP$^-$ cells, but some notable GO processes, such as "biotic stimulus" and "response to other organism," were specifically and significantly upregulated in the GFP$^+$ cells (Figure 3F; see also Table S2). These observations suggest that HIV1-GFP infection triggers the cellular gene expression of several genes. However, various GO processes were significantly downregulated in both the GFP$^+$ and GFP$^-$ cells (Figure 3G; see also Table S2), suggesting that HIV-1 infection status strongly affects cellular gene expression.

To identify the transcription factors (TFs) and/or DNA-binding proteins that specifically play certain roles in the GFP$^+$ cells compared with the GFP$^-$ cells, our transcriptomic data were used for the TF target enrichment analysis. This analysis predicted that histone lysine methyltransferase 2A (KMT2A) is the most highly active factor, specifically in the GFP$^+$ cells (Figure 3H; see also Table S3). The second factor, AFF1, was previously identified as a co-factor of HIV-1 Tat-mediated transactivation (Lu et al., 2014). Again, this result suggests that our transcriptomic data are fairly consistent with the findings from related investigations. Moreover, our analysis suggested that PSIP1 (also known as LEDGF), a partner of HIV-1 integrase for the targeting of viral integration into transcriptionally active sites (Cherepanov et al., 2003; Llano et al., 2006; Maertens et al., 2003), is active in the GFP$^+$ population (Figure 3H).

### Heterogeneity of the HIV-1-Infected Cells *In Vivo*

As only 17 DEGs were detected between the GFP$^+$ and GFP$^-$ cells (Figure 3E), it is reasonable to assume that CD4$^+$ T cells are heterogeneous *in vivo* regardless of HIV-1 infection, and therefore, the effect of HIV-1 infection on cellular gene expression may not be fully captured by bulk RNA-seq. To reveal the heterogeneity of infected cells *in vivo*, we performed single-cell RNA-seq (scRNA-seq) analysis. Consistent with our previous reports (Aso et al., 2019; Nakano et al., 2017; Yamada et al., 2018) and the dRNA-seq data (Figure 2; Table S1), ISG expression was upregulated in both GFP$^+$ and GFP$^-$ cells

CellPress
OPEN ACCESS

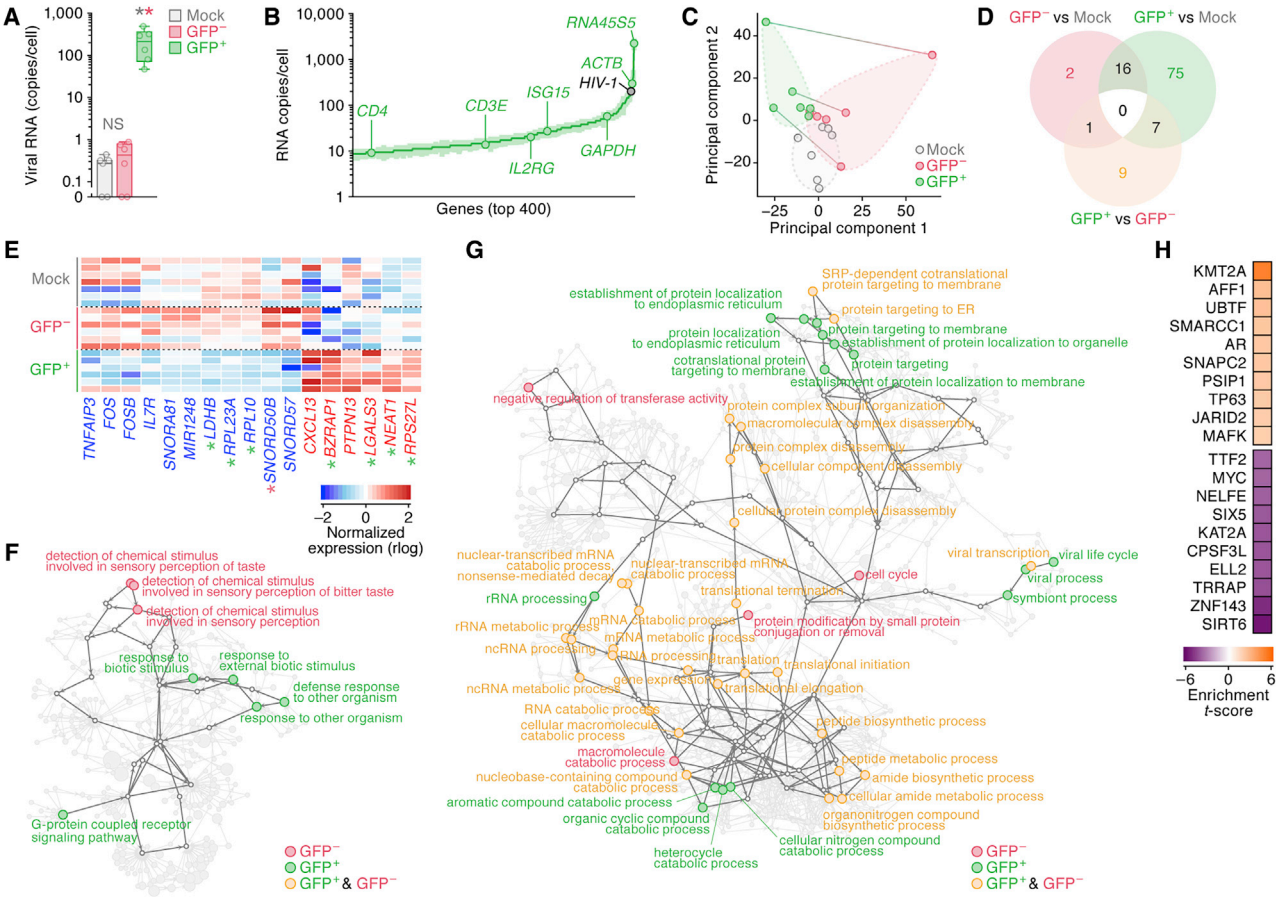



**Figure 3. Global Transcriptomic Analyses of the HIV1-GFP-Infected Cells**

(A) Viral RNA in each population. The absolute copy numbers of viral RNA in the CD4$^+$ T cells of mock-infected mice, the GFP$^-$CD4$^+$ T cells of infected mice, and the GFP$^+$CD4$^+$ T cells of infected mice (n = 6 each) were quantified using dRNA-seq. Significant differences ($p < 0.05$ by Mann-Whitney U test) compared with uninfected cells and GFP$^-$ cells are indicated by gray and red asterisks.

(B) The absolute copy numbers of the top 400 genes (x axis) expressed in the GFP$^+$ cells of infected mice (n = 8).

(C) Principal component analysis of the global transcriptome. Each dot represents the result from the corresponding mouse, and the lines connect the results from identical mice.

(D) Venn diagram of the DEGs in each population.

(E) Heatmap of the 17 DEGs between the GFP$^+$ and GFP$^-$ cells. The DEGs in GFP$^+$ versus mock-infected and GFP$^-$ versus mock-infected cell groups are indicated by green and red asterisks, respectively.

(F and G) GO enrichment analysis of the global transcriptome. The upregulated (F) and downregulated (G) GO terms are visualized in a hierarchical structure. The node size indicates the false discovery rate of enrichment.

(H) Gene set enrichment analysis for evaluating the effects of TFs on their binding target genes. The estimated effects of TFs are presented in a heatmap as the enrichment *t* scores.

See also Figure S3 and Tables S1, S2, and S3.

compared with the CD4$^+$ T cells of mock-infected mice (Figure S4A). Dimensionality reduction analyses according to gene expression showed that GFP$^+$ and GFP$^-$ cells tended to display distinct patterns of gene expression (Figure 4A, top; Figure S4B). On the basis of the single-cell global transcriptomic data, the splenic CD4$^+$ T cells can be classified into nine clusters (Figure 4A, bottom). These nine clusters included uninfected CD4$^+$ T cells of mock-infected mice (Figure 4B; see also Figure S4C), suggesting that there were no nascent clusters produced by HIV1-GFP infection. Notably, the proportion of GFP$^+$ cells was relatively high in clusters 4, 5, and 8 but was relatively low in clusters 3, 7, and 9 (Figure 4B). The expression levels of viral RNA in the GFP$^+$ cells of each cluster were also different (Figure 4C), and interestingly, the percentage of GFP$^+$ cells in each cluster was significantly and positively correlated with the expression level of HIV-1 RNA in the GFP$^+$ cells of each cluster (Figure 4D). These findings suggest that the virological properties of the respective clusters are different. To characterize the respective clusters, we extracted the genes that were strongly up- or down-regulated in each cluster compared with the other clusters (Figure 4E; see also Table S4). On the basis of these genes,

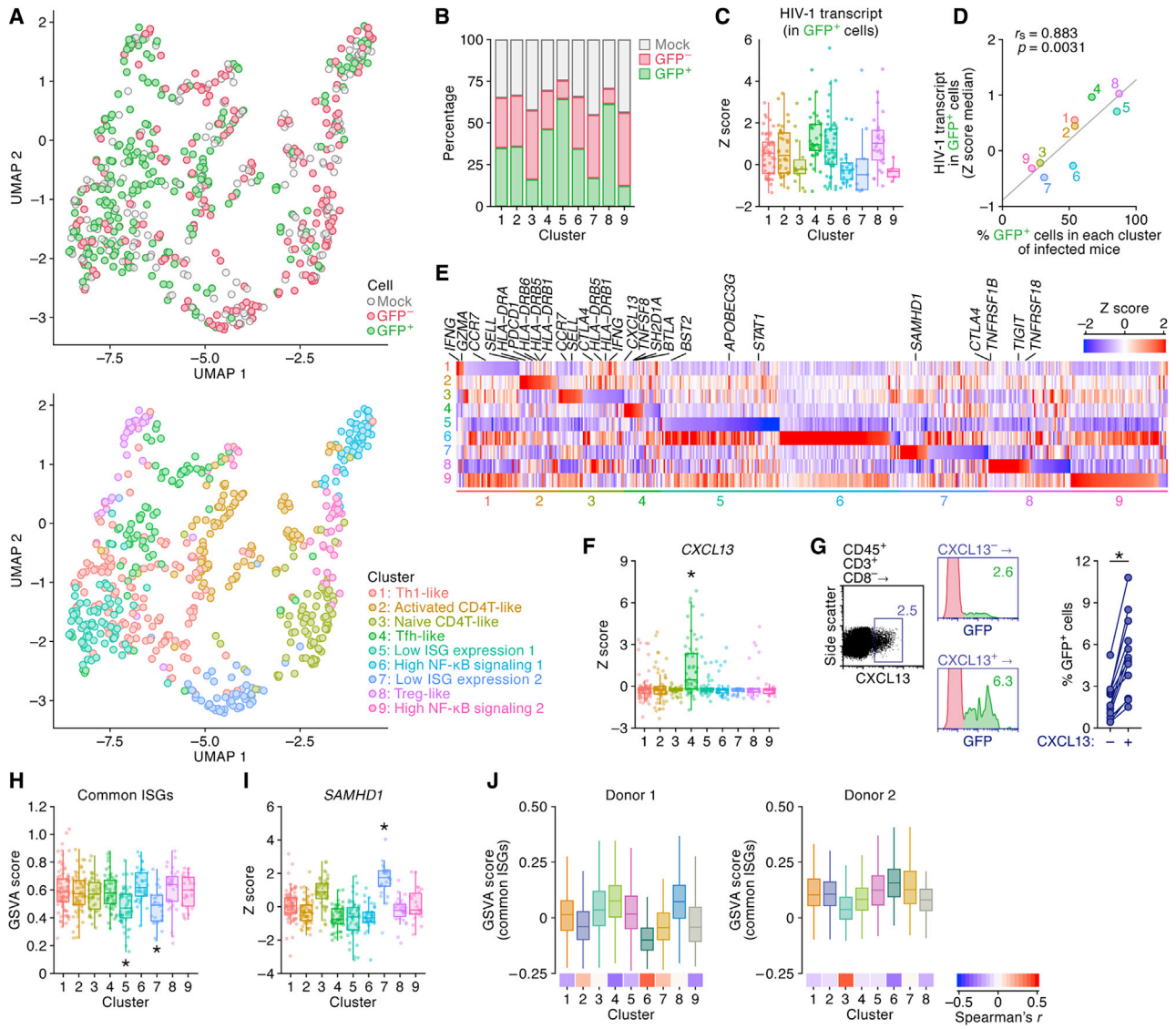



**Figure 4. Heterogeneity of the HIV1-GFP-Infected Cells**

(A) Uniform manifold approximation and projection (UMAP) plots representing the gene expression patterns of the cells (GFP$^+$ cells, 241 cells; GFP$^-$ cells, 193 cells; and mock-infected CD4$^+$ T cells, 235 cells). Each dot is colored according to the cell category (top) and the cluster information (bottom).

(B) Proportions of GFP$^+$ cells, GFP$^-$ cells, and mock-infected CD4$^+$ T cells in each cluster.

(C) Normalized expression level (*Z* score) of the HIV-1 transcripts in GFP$^+$ cells in each cluster.

(D) Association between the HIV-1 expression level and the proportion of GFP$^+$ cells in the corresponding clusters. The x axis indicates the proportion of GFP$^+$ cells in each cluster of the CD4$^+$ T cells of infected mice (i.e., GFP$^+$ and GFP$^-$ cells), and the y axis indicates the median normalized expression level of HIV-1 transcripts in GFP$^+$ cells in each cluster. Each dot indicates a cluster.

(E) Heatmap representing the expression levels of signature genes in each cluster.

(F) Normalized expression level of *CXCL13* in the CD4$^+$ T cells of infected mice.

(G) Proportion of the GFP$^+$ cells in CXCL13$^+$ and CXCL13$^-$ CD4$^+$ T cells. CD4$^+$ T cells (CD45$^+$CD3$^+$CD8$^-$ cells; n = 12) were classified into CXCL13$^+$ and CXCL13$^-$ cells (left), and the percentage of GFP$^+$ cells in each cell population was analyzed (middle and right). Each dot represents the result from the corresponding mouse, and the lines connect the results from identical mice.

(H and I) Gene set variation analysis (GSVA) score of "common ISGs" (H) and normalized expression level of *SAMHD1* (I) and in the CD4$^+$ T cells of infected mice in each cluster.

(J) Identification of cluster 5 (ISG$^{low}$)-like subpopulation in an scRNA-seq dataset of human primary CD4$^+$ T cells from two healthy donors. The subpopulation structure of the human CD4$^+$ T cells and the defined clusters are shown in Figure S4I. Top: GSVA score of "common ISGs" in each cluster. Bottom heatmap representing the similarity of gene expression pattern between each cluster and cluster 5 of humanized mice. Color indicates Spearman's rank correlation coefficient of gene expression between two clusters.

In (F), (H), and (I), asterisks denote significant differences ($p < 0.05$) tested using the Wilcoxon rank-sum test. In (G), an asterisk denotes a significant difference ($p < 0.0001$) tested using a paired t test. See also Figure S4 and Table S4. In (D), Spearman's rank correlation coefficient ($r_s$) is applied to determine a statistically significant correlation.

clusters 1, 2, 3, 4, and 8 were composed of Th1-like cells, activated $CD4^+$ T-like cells, naive $CD4^+$ T-like cells, follicular helper T (Tfh)-like cells, and regulatory $CD4^+$ T (Treg)-like cells (Figure 4E). Consistent with previous findings (Jiang et al., 2008; Sato et al., 2013), cluster 8 (Treg-like cells) is highly susceptible to HIV-1 infection and highly productive (Figure 4D).

Interestingly, *CXCL13* was uniquely and highly expressed in cluster 4 (Figure 4F), in which the $GFP^+$ cells were dominant (Figure 4B). This finding shows that the upregulation of *CXCL13* in the $GFP^+$ cells demonstrated using dRNA-seq (Figure 3E) can be attributed to this cluster. *CXCL13* expression was higher not only in the $GFP^+$ cells but also in the $GFP^-$ cells in cluster 4 than in the other clusters (Figure S4D), suggesting that this $CXCL13^{high}$ cluster did not emerge because of HIV-1 infection but was present in the infected humanized mice. Flow cytometry analysis further revealed that the percentage of $GFP^+$ cells in the $CXCL13^+CD4^+$ T cells was significantly higher than that in the $CXCL13^-CD4^+$ T cells (Figure 4G). These data suggest that the $CXCL13^{high}$ cell population contributes to productive infection *in vivo*. To further assess whether CXCL13-mediated signaling positively contributes to productive virus infection, we prepared a $CD4^+$ T cell line stably expressing CXCR5, the receptor for CXCL13 (Vinuesa et al., 2016), and analyzed the effect of CXCL13 on HIV-1 infection. However, CXCL13 treatment did not affect viral infection (Figure S4E), suggesting that CXCL13 itself does not induce a proviral state and may be a marker of the cells highly susceptible to HIV-1 infection and/or highly virus producing cells *in vivo*.

In cluster 5, in which the viral transcript level was high (Figure 4D), low *STAT1* expression was detected (Figure 4E). Given that STAT1 is a pivotal TF leading to the induction of ISG expression (Soper et al., 2018), we hypothesized that the expression levels of ISGs are relatively low in this cluster. To test this hypothesis, we assessed the expression levels of 107 "common ISGs," which are robustly upregulated upon type I IFN stimulation in multiple cell types, including $CD4^+$ T cells (Aso et al., 2019). The expression levels of "common ISGs," including two well-known anti-HIV-1 ISGs, *BST2* and *APOBEC3G*, were significantly low in clusters 5 and 7 (Figures 4E and 4H). These findings suggest that there is a $CD4^+$ T cell subset that does not express anti-HIV-1 ISGs, because of low *STAT1* expression, which positively contributes to viral spread *in vivo*. However, in cluster 7, the expression level of the HIV-1 transcript in the $GFP^+$ cells was relatively low, while cluster 5 exhibited a higher expression of viral transcripts (Figure 4H). With regard to this, we revealed that the expression level of *SAMHD1*, a known anti-HIV-1 ISG (Hrecka et al., 2011; Laguette et al., 2011), was specifically and significantly high in cluster 7 (Figure 4I; see also Figure S4H), suggesting that *SAMHD1* restricts the efficient HIV-1 replication in this cluster.

Finally, to assess the presence of the $CD4^+$ T cell subpopulation with low ISG expression, which is similar to cluster 5, in humans, we analyzed the scRNA-seq data of the human $CD4^+$ T cells obtained from the lymph nodes of two healthy individuals (Szabo et al., 2019). As shown in Figure 4J, we detected the subpopulation expressing relatively low "common ISGs" (cluster 6 in donor 1, and cluster 3 in donor 2; see also Figure S4I) in humans. Moreover, the transcriptional profiles of these clusters in the two healthy individuals were similar to that of cluster 5 in humanized mice. These findings suggest that the $CD4^+$ T cell subpopulation with low ISG expression is present in humans and that our humanized mice recapitulate the conditions in humans.

## DISCUSSION

Some previous studies performed scRNA-seq using HIV-1-infected samples (Bradley et al., 2018; Cohn et al., 2018; Golumbeanu et al., 2018). However, these previous studies are different in terms of (1) the aim of the study, (2) the virus used, and (3) the use of artificial activation stimulation before omics analyses. First, whereas the previous investigators aimed to analyze HIV-1 latently infected cells and tried to reveal the characteristics of the HIV-1 reservoirs, we aimed to characterize the HIV-1-producing cells *in vivo*. Second, whereas the viruses used by Bradley et al. (2018) and Golumbeanu et al. (2018) were replication incompetent, and seven of the nine viral genes were defective (NL4-3-Δ6-drEGFP), we used a replication-competent virus expressing all viral genes. Third, whereas the previous studies used additional activation stimulation to allow efficient HIV-1 infection *in vitro* (Bradley et al., 2018; Golumbeanu et al., 2018) or induce the expression of viral proteins (Cohn et al., 2018), we directly used the samples obtained from the infected humanized mice for multiomics analyses without any additional stimulation *in vitro* and described the heterogeneity of HIV-producing cells *in vivo*. Notably, the activation status of each human $CD4^+$ T cell subset in humanized mice is comparable with that in humans (Sato et al., 2013), and artificial activation stimulation drastically changes the cellular transcriptome profile (Imbeault et al., 2012; Rato et al., 2017; Yoder et al., 2017). Therefore, our results recapitulate the complexed heterogeneity of HIV-1-producing cells *in vivo*.

Consistent with previous findings (Han et al., 2004; Schröder et al., 2002), HIV-1 was preferentially integrated into open chromatin regions. Similar to recent findings in an *in vitro* cell culture system (Battivelli et al., 2018), the HIV-1 ISs in the $GFP^+$ cells were significantly more highly enriched in the active chromatin regions than those in the $GFP^-$ cells. However, because this difference is relatively minor, our findings suggest that efficient viral production and replication are not necessarily determined only by preferential HIV-1 integration into the active chromatin region. With regard to the cells that harbored viral DNA but were not activated by PMA+I stimulation, defective proviruses (Einkauf et al., 2019; Imamichi et al., 2016) and proviruses with APOBEC3-induced hypermutations (Simon et al., 2005) were detected in HIV-1-infected individuals. Moreover, the presence of infected cells with intact proviruses that were not induced by latency-reversing agents in patients has been reported (Hiener et al., 2017; Ho et al., 2013). Altogether, our findings suggest that viral production and silencing are determined by multiple factors *in vivo*.

Using the dRNA-seq data, we performed TF target enrichment analysis and identified KMT2A as the most active DNA-binding protein in the $GFP^+$ cells. Intriguingly, KMT2A enzymatically methylates histone H3K4 (Del Rizzo and Trievel, 2011), and a previous study demonstrated that H3K4me3 modification in HIV-1 long terminal repeat, a viral intrinsic promoter for HIV-1,

positively regulates viral transcription (Matsuda et al., 2015). Additionally, HIV-1 reactivation is suppressed by *KMT2A* knockdown (Boehm et al., 2017). Given that global IS analysis suggested that HIV-1 is integrated into the genomic region close to H3K4me3 specifically in GFP$^+$ cells, our multiomics findings suggest that KMT2A-mediated H3K4me3 modification positively affects efficient viral production *in vivo*. Moreover, another factor of interest, PSIP1, was also identified by TF target enrichment analysis. PSIP1 is a chromatin-associated protein and a co-factor for HIV-1 integration for targeting HIV-1 integration into transcriptionally active sites (Ciuffi et al., 2005; Llano et al., 2006). Given that PSIP1 interacts with KMT2A (Yokoyama and Cleary, 2008) and H3K4me3 (Tsutsui et al., 2011), our results suggest that KMT2A and PSIP1 may cooperate and contribute to efficient viral production *in vivo*.

Through dRNA-seq analysis, we revealed that viral transcripts were highly expressed in the GFP$^+$ cells. Additionally, scRNA-seq analysis revealed that the GFP$^+$ cells *in vivo* are heterogeneous and can be classified into at least nine clusters. In particular, we identified three subpopulations, clusters 4, 5, and 8, that express relatively high levels of viral transcripts. In addition to viral transcripts, cluster 4 expressed a high level of *CXCL13*, which explains the increased expression of *CXCL13* in the GFP$^+$ cells detected using dRNA-seq. Moreover, we revealed that the CXCL13$^+$CD4$^+$ T cells are highly productive for HIV-1 in the humanized mice. The presence of this *CXCL13$^{high}$* subpopulation in the HIV-1-infected cells is consistent with a recent report (Morou et al., 2019). These observations suggest that this *CXCL13$^{high}$* cluster contains Tfh-like cells. Intriguingly, previous reports suggested that the Tfh cells in the lymph nodes of infected individuals are the major source of infectious viruses (Banga et al., 2016; Perreau et al., 2013). CXCR5 is the receptor of CXCL13 and is expressed on Tfh cells (Vinuesa et al., 2016). Therefore, our data suggest that this *CXCL13$^{high}$* subpopulation (cluster 4) may recruit bystander Tfh cells via the CXCL13-CXCR5 axis and accelerates efficient viral spread *in vivo*.

Upregulation of various ISGs, particularly "common ISGs" (Aso et al., 2019), *in vivo* has been observed in HIV-1-infected individuals (Rotger et al., 2010; Sedaghat et al., 2008) and humanized mice (Cheng et al., 2017; Nakano et al., 2017; Yamada et al., 2018). Intriguingly, in cluster 5, the expression level of *STAT1*, the key TF for ISG expression, was low, and relatively high levels of viral transcripts were detected. Furthermore, a similar subpopulation of CD4$^+$ T cells with low expression of ISGs was detected in healthy individuals. Rato et al. (2017) reported the presence of a subpopulation that was highly permissive to HIV-1 infection in the primary human CD4$^+$ T cells activated with mitogens *in vitro*, and interestingly, this subpopulation exhibited lower ISG expression. However, the relevance of such subpopulation(s) permissive to HIV-1 infection *in vivo* remains unknown. Our findings describe the presence of HIV-1-producing subpopulations vulnerable to HIV-1 infection, presumably because of low anti-HIV-1 ISG expression *in vivo*.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:



## SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.celrep.2020.107887.

## ACKNOWLEDGMENTS

We would like to thank Naoko Misawa and Kotubu Misawa (inFront, Kyoto University), Mai Suganami (IMSUT), Yuuta Kuze and Kiyomi Imamura (Graduate School of Frontier Sciences, University of Tokyo), and Kaori Fukuhara (RIKEN BDR) for experimental, technical, and dedicated support. We also thank the University of California, Los Angeles (UCLA) Center for AIDS Research (CFAR) Gene and Cellular Therapy Core Facility for providing human HSCs (NIH/National Institute of Allergy and Infectious Diseases [NIAID] grant 5P30 AI028697). The super-computing resource SHIROKANE was provided by the Human Genome Center, IMSUT, Japan. This study was supported by Agency for Medical Research and Development (AMED) J-PRIDE 19fm0208006 (to E.K. and K. Sato); AMED Research Program on HIV/AIDS 19fk0410014 (to Y. Satou, Y.K., and K. Sato) and 19fk0410019 (to K. Sato); AMED Emerging/Re-emerging Infectious Diseases 20fk0108146h0001 (to K. Sato); Japan Science and Technology Agency (JST) Core Research for Evolutional Science and Technology (CREST) (to K. Sato); Japan Society for the Promotion of Science (JSPS) KAKENHI PAGS 16H06279 (to Y. Suzuki and K. Sato); JSPS Scientific Research B 18H02662 (to S. Nakaoka, Y. Suzuki, K. Shiroguchi, and K.

Sato); JSPS Scientific Research on Innovative Areas 16H06429 (to K. Sato), 16K21723 (to K. Sato), 17H05813 (to K. Sato), and 19H04826 (to K. Sato); JSPS Early-Career Scientists 20K15767 (to J.I.); JSPS Research Fellow DC1 20J23299 (to H.A.), DC1 19J22802 (to S. Nagaoka), and PD 19J01713 (to J.I.); the Joint Usage/Research Center program of inFront, Kyoto University (to K. Sato); the Institute of Medical Science, University of Tokyo (IMSUT) Joint Research Project (to Y.K.); the Takeda Science Foundation (to K. Sato); the Lotte Foundation (to K. Sato); the Mochida Memorial Foundation for Medical and Pharmaceutical Research (to K. Sato); the Daiichi Sankyo Foundation of Life Science (to K. Sato); the Sumitomo Foundation (to K. Sato); and the Uehara Foundation (to K. Sato).

## AUTHOR CONTRIBUTIONS

H.A. and S. Nagaoka mainly performed the experiments and analyzed the data. J.I., B.J.Y.T., S.I., and Y. Satou conducted global IS analysis. H.A., E.K., S. Nakaoka, K. Shiroguchi, and Y. Suzuki conducted transcriptome analysis. E.K., J.I., and K.A. supported bioinformatic analysis. Y.K. provided reagents. K. Sato conceived and designed the experiments. K. Sato mainly wrote the manuscript. All authors reviewed and edited the manuscript.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 16, 2019
Revised: February 6, 2020
Accepted: June 8, 2020
Published: July 14, 2020

## REFERENCES

Aso, H., Ito, J., Koyanagi, Y., and Sato, K. (2019). Comparative description of the expression profile of interferon-stimulated genes in multiple cell lineages targeted by HIV-1 infection. Front. Microbiol. *10*, 429.

Banga, R., Procopio, F.A., Noto, A., Pollakis, G., Cavassini, M., Ohmiti, K., Corpataux, J.M., de Leval, L., Pantaleo, G., and Perreau, M. (2016). PD-1(+) and follicular helper T cells are responsible for persistent HIV-1 transcription in treated aviremic individuals. Nat. Med. *22*, 754–761.

Battivelli, E., Dahabieh, M.S., Abdel-Mohsen, M., Svensson, J.P., Tojal Da Silva, I., Cohn, L.B., Gramatica, A., Deeks, S., Greene, W.C., Pillai, S.K., and Verdin, E. (2018). Distinct chromatin functional states correlate with HIV latency reactivation in infected primary CD4+ T cells. eLife *7*, e34655.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B Stat. Methodol. *57*, 289–300.

Boehm, D., Jeng, M., Camus, G., Gramatica, A., Schwarzer, R., Johnson, J.R., Hull, P.A., Montano, M., Sakane, N., Pagans, S., et al. (2017). SMYD2-mediated histone methylation contributes to HIV-1 latency. Cell Host Microbe *21*, 569–579.e6.

Bradley, T., Ferrari, G., Haynes, B.F., Margolis, D.M., and Browne, E.P. (2018). Single-cell analysis of quiescent HIV infection reveals host transcriptional profiles that regulate proviral latency. Cell Rep. *25*, 107–117.e3.

Buggert, M., Nguyen, S., Salgado-Montes de Oca, G., Bengsch, B., Darko, S., Ransier, A., Roberts, E.R., Del Alcazar, D., Brody, I.B., Vella, L.A., et al. (2018). Identification and characterization of HIV-specific resident memory CD8+ T cells in human lymphoid tissue. Sci. Immunol. *3*, eaar4526.

Cheng, L., Yu, H., Li, G., Li, F., Ma, J., Li, J., Chi, L., Zhang, L., and Su, L. (2017). Type I interferons suppress viral replication but contribute to T cell depletion and dysfunction during chronic HIV-1 infection. JCI Insight *2*, e94366.

Cherepanov, P., Maertens, G., Proost, P., Devreese, B., Van Beeumen, J., Engelborghs, Y., De Clercq, E., and Debyser, Z. (2003). HIV-1 integrase forms stable tetramers and associates with LEDGF/p75 protein in human cells. J. Biol. Chem. *278*, 372–381.

Ciuffi, A., Llano, M., Poeschla, E., Hoffmann, C., Leipzig, J., Shinn, P., Ecker, J.R., and Bushman, F. (2005). A role for LEDGF/p75 in targeting HIV DNA integration. Nat. Med. *11*, 1287–1289.

Cohn, L.B., Silva, I.T., Oliveira, T.Y., Rosales, R.A., Parrish, E.H., Learn, G.H., Hahn, B.H., Czartoski, J.L., McElrath, M.J., Lehmann, C., et al. (2015). HIV-1 integration landscape during latent and active infection. Cell *160*, 420–432.

Cohn, L.B., da Silva, I.T., Valieris, R., Huang, A.S., Lorenzi, J.C.C., Cohen, Y.Z., Pai, J.A., Butler, A.L., Caskey, M., Jankovic, M., and Nussenzweig, M.C. (2018). Clonal CD4+ T cells in the HIV-1 latent reservoir display a distinct gene profile upon reactivation. Nat. Med. *24*, 604–609.

Del Rizzo, P.A., and Trievel, R.C. (2011). Substrate and product specificities of SET domain methyltransferases. Epigenetics *6*, 1059–1067.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Douek, D.C., Brenchley, J.M., Betts, M.R., Ambrozak, D.R., Hill, B.J., Okamoto, Y., Casazza, J.P., Kuruppu, J., Kunstman, K., Wolinsky, S., et al. (2002). HIV preferentially infects HIV-specific CD4+ T cells. Nature *417*, 95–98.

Ebina, H., Kanemura, Y., Misawa, N., Sakuma, T., Kobayashi, T., Yamamoto, T., and Koyanagi, Y. (2015). A high excision potential of TALENs for integrated DNA of HIV-based lentiviral vector. PLoS ONE *10*, e0120047.

Einkauf, K.B., Lee, G.Q., Gao, C., Sharaf, R., Sun, X., Hua, S., Chen, S.M., Jiang, C., Lian, X., Chowdhury, F.Z., et al. (2019). Intact HIV-1 proviruses accumulate at distinct chromosomal positions during prolonged antiretroviral therapy. J. Clin. Invest. *129*, 988–998.

Farhadian, S.F., Mehta, S.S., Zografou, C., Robertson, K., Price, R.W., Pappalardo, J., Chiarella, J., Hafler, D.A., and Spudich, S.S. (2018). Single-cell RNA sequencing reveals microglia-like cells in cerebrospinal fluid during virologically suppressed HIV. JCI Insight *3*, e121718.

Golumbeanu, M., Cristinelli, S., Rato, S., Munoz, M., Cavassini, M., Beerenwinkel, N., and Ciuffi, A. (2018). Single-cell RNA-seq reveals transcriptional heterogeneity in latent and reactivated HIV-infected cells. Cell Rep. *23*, 942–950.

Han, Y., Lassen, K., Monie, D., Sedaghat, A.R., Shimoji, S., Liu, X., Pierson, T.C., Margolick, J.B., Siliciano, R.F., and Siliciano, J.D. (2004). 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. *78*, 6122–6133.

Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics *14*, 7.

Hiener, B., Horsburgh, B.A., Eden, J.S., Barton, K., Schlub, T.E., Lee, E., von Stockenstrom, S., Odevall, L., Milush, J.M., Liegler, T., et al. (2017). Identification of genetically intact HIV-1 proviruses in specific CD4+ T cells from effectively treated participants. Cell Rep. *21*, 813–822.

Ho, Y.C., Shan, L., Hosmane, N.N., Wang, J., Laskey, S.B., Rosenbloom, D.I., Lai, J., Blankson, J.N., Siliciano, J.D., and Siliciano, R.F. (2013). Replication-competent noninduced proviruses in the latent reservoir increase barrier to HIV-1 cure. Cell *155*, 540–551.

Hrecka, K., Hao, C., Gierszewska, M., Swanson, S.K., Kesik-Brodacka, M., Srivastava, S., Florens, L., Washburn, M.P., and Skowronski, J. (2011). Vpx relieves inhibition of HIV-1 infection of macrophages mediated by the SAMHD1 protein. Nature *474*, 658–661.

Hughes, S.H., and Coffin, J.M. (2016). What integration sites tell us about HIV persistence. Cell Host Microbe *19*, 588–598.

Imamichi, H., Dewar, R.L., Adelsberger, J.W., Rehm, C.A., O'Doherty, U., Paxinos, E.E., Fauci, A.S., and Lane, H.C. (2016). Defective HIV-1 proviruses produce novel protein-coding RNA species in HIV-infected patients on combination antiretroviral therapy. Proc. Natl. Acad. Sci. U S A *113*, 8783–8788.

Imbeault, M., Giguère, K., Ouellet, M., and Tremblay, M.J. (2012). Exon level transcriptomic profiling of HIV-1-infected CD4(+) T cells reveals virus-induced genes and host environment favorable for viral replication. PLoS Pathog. *8*, e1002861.

Ito, M., Hiramatsu, H., Kobayashi, K., Suzue, K., Kawahata, M., Hioki, K., Ueyama, Y., Koyanagi, Y., Sugamura, K., Tsuji, K., et al. (2002). NOD/SCID/$\gamma_{(c)}^{null}$ mouse: an excellent recipient mouse model for engraftment of human cells. Blood *100*, 3175–3182.

Iwase, S.C., Miyazato, P., Katsuya, H., Islam, S., Yang, B.T.J., Ito, J., Matsuo, M., Takeuchi, H., Ishida, T., Matsuda, K., et al. (2019). HIV-1 DNA-capture-seq is a useful tool for the comprehensive characterization of HIV-1 provirus. Sci. Rep. *9*, 12326.

Janeway, C.A., Travers, P., Walport, M., and Shlomchik, M.J. (2005). T cell-mediated immunity. Immunobiology: The Immune System in Health and Disease (New York: Garland Science), pp. 25–51.

Jiang, Q., Zhang, L., Wang, R., Jeffrey, J., Washburn, M.L., Brouwer, D., Barbour, S., Kovalev, G.I., Unutmaz, D., and Su, L. (2008). FoxP3+CD4+ regulatory T cells play an important role in acute HIV-1 infection in humanized Rag2-/-gammaC-/- mice in vivo. Blood *112*, 2858–2868.

Jordan, A., Bisgrove, D., and Verdin, E. (2003). HIV reproducibly establishes a latent infection after acute infection of T cells *in vitro*. EMBO J. *22*, 1868–1877.

Kawakami, E., Nakaoka, S., Ohta, T., and Kitano, H. (2016). Weighted enrichment method for prediction of transcription regulators from transcriptome and global chromatin immunoprecipitation data. Nucleic Acids Res. *44*, 5010–5021.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. *14*, R36.

Kuo, H.H., Ahmad, R., Lee, G.Q., Gao, C., Chen, H.R., Ouyang, Z., Szucs, M.J., Kim, D., Tsibris, A., Chun, T.W., et al. (2018). Anti-apoptotic protein BIRC5 maintains survival of HIV-1-infected CD4$^+$ T cells. Immunity *48*, 1183–1194.e5.

Laguette, N., Sobhian, B., Casartelli, N., Ringeard, M., Chable-Bessia, C., Ségéral, E., Yatim, A., Emiliani, S., Schwartz, O., and Benkirane, M. (2011). SAMHD1 is the dendritic- and myeloid-cell-specific HIV-1 restriction factor counteracted by Vpx. Nature *474*, 654–657.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930.

Linderman, G.C., Zhao, J., and Kluger, Y. (2018). Zero-preserving imputation of scRNA-seq data using low-rank approximation. bioRxiv. https://doi.org/10.1101/397588.

Llano, M., Saenz, D.T., Meehan, A., Wongthida, P., Peretz, M., Walker, W.H., Teo, W., and Poeschla, E.M. (2006). An essential role for LEDGF/p75 in HIV integration. Science *314*, 461–464.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Lu, H., Li, Z., Xue, Y., Schulze-Gahmen, U., Johnson, J.R., Krogan, N.J., Alber, T., and Zhou, Q. (2014). AFF1 is a ubiquitous P-TEFb partner to enable Tat extraction of P-TEFb from 7SK snRNP and formation of SECs for HIV transactivation. Proc. Natl. Acad. Sci. U S A *111*, E15–E24.

Luo, W., Friedman, M.S., Shedden, K., Hankenson, K.D., and Woolf, P.J. (2009). GAGE: generally applicable gene set enrichment for pathway analysis. BMC Bioinformatics *10*, 161.

Maertens, G., Cherepanov, P., Pluymers, W., Busschots, K., De Clercq, E., Debyser, Z., and Engelborghs, Y. (2003). LEDGF/p75 is essential for nuclear and chromosomal targeting of HIV-1 integrase in human cells. J. Biol. Chem. *278*, 33528–33539.

Maldarelli, F., Wu, X., Su, L., Simonetti, F.R., Shao, W., Hill, S., Spindler, J., Ferris, A.L., Mellors, J.W., Kearney, M.F., et al. (2014). HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. Science *345*, 179–183.

Martin, M. (2011). CUTADAPT removes adapter sequences from high-throughput sequencing reads. EMBnetjournal *17*, 10–12.

Matsuda, Y., Kobayashi-Ishihara, M., Fujikawa, D., Ishida, T., Watanabe, T., and Yamagishi, M. (2015). Epigenetic heterogeneity in HIV-1 latency establishment. Sci. Rep. *5*, 7701.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: uniform manifold approximation and projection for dimension reduction. arXiv, arXiv:1802.03426. https://arxiv.org/abs/1802.03426.

Miura, Y., Misawa, N., Maeda, N., Inagaki, Y., Tanaka, Y., Ito, M., Kayagaki, N., Yamamoto, N., Yagita, H., Mizusawa, H., and Koyanagi, Y. (2001). Critical contribution of tumor necrosis factor-related apoptosis-inducing ligand (TRAIL) to apoptosis of human CD4$^+$ T cells in HIV-1-infected hu-PBL-NOD-SCID mice. J. Exp. Med. *193*, 651–660.

Miyoshi, H., Blömer, U., Takahashi, M., Gage, F.H., and Verma, I.M. (1998). Development of a self-inactivating lentivirus vector. J. Virol. *72*, 8150–8157.

Morou, A., Brunet-Ratnasingham, E., Dubé, M., Charlebois, R., Mercier, E., Darko, S., Brassard, N., Nganou-Makamdop, K., Arumugam, S., Gendron-Lepage, G., et al. (2019). Altered differentiation is central to HIV-specific CD4$^+$ T cell dysfunction in progressive disease. Nat. Immunol. *20*, 1059–1070.

Nakano, Y., Misawa, N., Juarez-Fernandez, G., Moriwaki, M., Nakaoka, S., Funo, T., Yamada, E., Soper, A., Yoshikawa, R., Ebrahimi, D., et al. (2017). HIV-1 competition experiments in humanized mice show that APOBEC3H imposes selective pressure and promotes virus adaptation. PLoS Pathog. *13*, e1006348.

Ogawa, T., Kryukov, K., Imanishi, T., and Shiroguchi, K. (2017). The efficacy and further functional advantages of random-base molecular barcodes for absolute and digital quantification of nucleic acid molecules. Sci. Rep. *7*, 13576.

Perreau, M., Savoye, A.L., De Crignis, E., Corpataux, J.M., Cubas, R., Haddad, E.K., De Leval, L., Graziosi, C., and Pantaleo, G. (2013). Follicular helper T cells serve as the major CD4 T cell compartment for HIV-1 infection, replication, and production. J. Exp. Med. *210*, 143–156.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Rato, S., Rausell, A., Muñoz, M., Telenti, A., and Ciuffi, A. (2017). Single-cell analysis identifies cellular markers of the HIV permissive cell. PLoS Pathog. *13*, e1006678.

Reed, L.J., and Muench, H. (1938). A simple method of estimating fifty per cent endpoints. Am. J. Hyg. *27*, 493–497.

Rotger, M., Dang, K.K., Fellay, J., Heinzen, E.L., Feng, S., Descombes, P., Shianna, K.V., Ge, D., Günthard, H.F., Goldstein, D.B., and Telenti, A.Swiss HIV Cohort Study; Center for HIV/AIDS Vaccine Immunology (2010). Genome-wide mRNA expression correlates of viral control in CD4$^+$ T-cells from HIV-1-infected individuals. PLoS Pathog. *6*, e1000781.

Sato, K., Misawa, N., Fukuhara, M., Iwami, S., An, D.S., Ito, M., and Koyanagi, Y. (2012). Vpu augments the initial burst phase of HIV-1 propagation and downregulates BST2 and CD4 in humanized mice. J. Virol. *86*, 5000–5013.

Sato, K., Misawa, N., Iwami, S., Satou, Y., Matsuoka, M., Ishizaka, Y., Ito, M., Aihara, K., An, D.S., and Koyanagi, Y. (2013). HIV-1 Vpr accelerates viral replication during acute infection by exploitation of proliferating CD4$^+$ T cells *in vivo*. PLoS Pathog. *9*, e1003812.

Sato, K., Takeuchi, J.S., Misawa, N., Izumi, T., Kobayashi, T., Kimura, Y., Iwami, S., Takaori-Kondo, A., Hu, W.S., Aihara, K., et al. (2014). APOBEC3D and APOBEC3F potently promote HIV-1 diversification and evolution in humanized mouse model. PLoS Pathog. *10*, e1004453.

Satou, Y., Katsuya, H., Fukuda, A., Misawa, N., Ito, J., Uchiyama, Y., Miyazato, P., Islam, S., Fassati, A., Melamed, A., et al. (2017). Dynamics and mechanisms of clonal expansion of HIV-1-infected cells in a humanized mouse model. Sci. Rep. *7*, 6913.

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics *27*, 863–864.

Schröder, A.R., Shinn, P., Chen, H., Berry, C., Ecker, J.R., and Bushman, F. (2002). HIV-1 integration in the human genome favors active genes and local hotspots. Cell *110*, 521–529.

Sedaghat, A.R., German, J., Teslovich, T.M., Cofrancesco, J., Jr., Jie, C.C., Talbot, C.C., Jr., and Siliciano, R.F. (2008). Chronic CD4$^+$ T-cell activation

and depletion in human immunodeficiency virus type 1 infection: type I interferon-mediated disruption of T-cell dynamics. J. Virol. *82*, 1870–1883.

Shiroguchi, K., Jia, T.Z., Sims, P.A., and Xie, X.S. (2012). Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. Proc. Natl. Acad. Sci. U S A *109*, 1347–1352.

Simon, V., Zennou, V., Murray, D., Huang, Y., Ho, D.D., and Bieniasz, P.D. (2005). Natural variation in Vif: differential impact on APOBEC3G/3F and a potential role in HIV-1 diversification. PLoS Pathog. *1*, e6.

Simonetti, F.R., Sobolewski, M.D., Fyne, E., Shao, W., Spindler, J., Hattori, J., Anderson, E.M., Watters, S.A., Hill, S., Wu, X., et al. (2016). Clonally expanded $CD4^+$ T cells can produce infectious HIV-1 *in vivo*. Proc. Natl. Acad. Sci. U S A *113*, 1883–1888.

Soper, A., Kimura, I., Nagaoka, S., Konno, Y., Yamamoto, K., Koyanagi, Y., and Sato, K. (2018). Type I interferon responses by HIV-1 infection: association with disease progression and control. Front. Immunol. *8*, 1823.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. Cell *177*, 1888–1902.e21.

Suzuki, Y., Koyanagi, Y., Tanaka, Y., Murakami, T., Misawa, N., Maeda, N., Kimura, T., Shida, H., Hoxie, J.A., O'Brien, W.A., and Yamamoto, N. (1999). Determinant in human immunodeficiency virus type 1 for efficient replication under cytokine-induced $CD4(^+)$ T-helper 1 (Th1)- and Th2-type conditions. J. Virol. *73*, 316–324.

Suzuki, A., Matsushima, K., Makinoshima, H., Sugano, S., Kohno, T., Tsuchihara, K., and Suzuki, Y. (2015). Single-cell analysis of lung adenocarcinoma cell lines reveals diverse expression patterns of individual cells invoked by a molecular target drug treatment. Genome Biol. *16*, 66.

Szabo, P.A., Levitin, H.M., Miron, M., Snyder, M.E., Senda, T., Yuan, J., Cheng, Y.L., Bush, E.C., Dogra, P., Thapa, P., et al. (2019). Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. Nat. Commun. *10*, 4706.

Tenno, M., Shiroguchi, K., Muroi, S., Kawakami, E., Koseki, K., Kryukov, K., Imanishi, T., Ginhoux, F., and Taniuchi, I. (2017). Cbfβ2 deficiency preserves Langerhans cell precursors by lack of selective TGFβ receptor signaling. J. Exp. Med. *214*, 2933–2946.

Tsutsui, K.M., Sano, K., Hosoya, O., Miyamoto, T., and Tsutsui, K. (2011). Nuclear protein LEDGF/p75 recognizes supercoiled DNA by a novel DNA-binding domain. Nucleic Acids Res. *39*, 5067–5081.

Vinuesa, C.G., Linterman, M.A., Yu, D., and MacLennan, I.C. (2016). Follicular helper T cells. Annu. Rev. Immunol. *34*, 335–368.

Wagner, T.A., McLaughlin, S., Garg, K., Cheung, C.Y., Larsen, B.B., Styrchak, S., Huang, H.C., Edlefsen, P.T., Mullins, J.I., and Frenkel, L.M. (2014). HIV latency. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. Science *345*, 570–573.

Wang, G.P., Ciuffi, A., Leipzig, J., Berry, C.C., and Bushman, F.D. (2007). HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. Genome Res. *17*, 1186–1194.

Wei, X., Decker, J.M., Liu, H., Zhang, Z., Arani, R.B., Kilby, J.M., Saag, M.S., Wu, X., Shaw, G.M., and Kappes, J.C. (2002). Emergence of resistant human immunodeficiency virus type 1 in patients receiving fusion inhibitor (T-20) monotherapy. Antimicrob. Agents Chemother. *46*, 1896–1905.

Yamada, E., Yoshikawa, R., Nakano, Y., Misawa, N., Koyanagi, Y., and Sato, K. (2015). Impacts of humanized mouse models on the investigation of HIV-1 infection: illuminating the roles of viral accessory proteins *in vivo*. Viruses *7*, 1373–1390.

Yamada, E., Nakaoka, S., Klein, L., Reith, E., Langer, S., Hopfensperger, K., Iwami, S., Schreiber, G., Kirchhoff, F., Koyanagi, Y., et al. (2018). Human-specific adaptations in Vpu conferring anti-tetherin activity are critical for efficient early HIV-1 replication *in vivo*. Cell Host Microbe *23*, 110–120.e7.

Yoder, A.C., Guo, K., Dillon, S.M., Phang, T., Lee, E.J., Harper, M.S., Helm, K., Kappes, J.C., Ochsenbauer, C., McCarter, M.D., et al. (2017). The transcriptome of HIV-1 infected intestinal $CD4^+$ T cells exposed to enteric bacteria. PLoS Pathog. *13*, e1006226.

Yokoyama, A., and Cleary, M.L. (2008). Menin critically links MLL proteins with LEDGF on cancer-associated target genes. Cancer Cell *14*, 36–46.

Zhang, Q., Chen, C.Y., Yedavalli, V.S., and Jeang, K.T. (2013). NEAT1 long noncoding RNA and paraspeckle bodies modulate HIV-1 posttranscriptional expression. MBio *4*, e00596-e12.

## STAR★METHODS

### KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Antibodies | | |
| PerCP/Cy5.5-conjugated anti-CD45 antibody | Biolegend | Cat# 368504; RRID: AB_2566352 |
| PE-conjugated anti-CD3 antibody | BD Biosciences | Cat# 555333; RRID: AB_395740 |
| APC-conjugated anti-CD8 antibody | Dako | Cat# C722701; RRID: AB_578594 |
| APC-conjugated anti-CXCL13 antibody | Thermo Fisher Scientific | Cat# MA5-23629; RRID: AB_2610225 |
| APC-conjugated anti-CXCR5 antibody | BioLegend | Cat# 35908; RRID: AB_2561817 |
| Anti-HIV-1 p24 antibody | Virostat | Cat# 1951 |
| Anti-Vif antibody | NIH AIDS Reagent Program | Cat# 6459 |
| Anti-Vpr antibody (clone 8D1) | Cosmo Bio | Cat# NCG-M01 |
| Anti-Vpu antibody | NIH AIDS Reagent Program | Cat# 969 |
| Anti-Env antibody | NIH AIDS Reagent Program | Cat# 12559 |
| Anti-Nef antibody (clone 3D12) | Thermo Fisher Scientific | Cat# MA1-71501; RRID: AB_962113 |
| Anti-PR antibody (clone 1696) | Thermo Fisher Scientific | Cat# MA1-19015; RRID: AB_1075640 |
| Anti-RT antibody | NIH AIDS Reagent Program | Cat# 11338 |
| Anti-IN antibody (clone IN-2) | Abcam | Cat# ab66645; RRID: AB_1139533 |
| Anti-GFP antibody | Sigma-Aldrich | Cat# G6795; RRID: AB_563117 |
| Anti-alpha-Tubulin antibody | Sigma-Aldrich | Cat# T9026; RRID: AB_477593 |
| Bacterial and Virus Strains | | |
| HIV1-GFP (strain NLSCFV3-GFP) | (Miura et al., 2001) | N/A |
| HIV-1 (strain NLSCFV3) | (Suzuki et al., 1999) | N/A |
| Biological Samples | | |
| Human CD34$^{+}$ hematopoietic stem cells | UCLA CFAR Gene and Cellular Therapy Core Facility | https://www.uclahealth.org/aidsinstitute/cfar/gene-and-cellular-therapy-core |
| Human PBMCs | This study | N/A |
| Chemicals, Peptides, and Recombinant Proteins | | |
| Dulbecco's modified Eagle's medium | Sigma-Aldrich | Cat# D6046-500ML |
| RPMI 1640 | Sigma-Aldrich | Cat# R8758-500ML |
| Fetal calf serum (FCS) | Sigma-Aldrich | Cat# 172012-500ML |
| Penicillin streptomycin | Sigma-Aldrich | Cat# P4333-100ML |
| L-glutamate | Thermo Fisher Scientific | Cat# 25030081 |
| Ficoll-Paque | GE Healthcare | Cat# 17-1440-03 |
| Phytohemagglutinin (PHA) | Sigma-Aldrich | Cat# 11082132001 |
| Phorbol 12-myristate 13-acetate (PMA) | Sigma-Aldrich | Cat# P-8139 |
| Ionomycin | Sigma-Aldrich | Cat# I-0634 |
| Nelfinavir | Sigma-Aldrich | Cat# PZ0013 |
| Blasticidin | Invivogen | Cat# ant-bl-1 |
| Recombinant CXCL13 | Funakoshi | Cat# 801-CX-025 |
| EcoRI | Takara | Cat# 1040A |
| BamHI | Takara | Cat# 1010A |
| ddPCR Supermix for probes | Bio-Rad | Cat# 1863010 |
| NEBNext Ultra II End Repair/dA-Tailing module | New England BioLabs | Cat# E7546 |
| NEBNext Ultra II ligation module | New England BioLabs | Cat# E7595 |
| Agencourt AMPure XP | Beckman Coulter | Cat# A63880 |
| C1 single-cell Auto Prep Reagent kit for mRNA Seq | Fluidigm | Cat# 100-6201 |

(*Continued on next page*)

*Continued*

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| C1 single-cell Auto Prep IFC for mRNA Seq (5-10 μm) | Fluidigm | Cat# 100-5759 |
| SMART-Seq v4 ultra low RNA kit for the Fluidigm C1 system | Takara Bio | Cat# 635026 |
| Nextera XT DNA sample preparation kit | Illumina | Cat# FC-131-1096 |
| Nextera XT DNA sample preparation index kit | Illumina | Cat# FC-131-1002 |
| High Sensitivity DNA Chips and Reagents | Agilient Technologies | Cat# 5067-4626 |
| SuperScript II Reverse Transcriptase | Thermo Fisher Scientific | Cat# 18064022 |
| SUPERase In RNase Inhibitor | Thermo Fisher Scientific | Cat# AM2694 |
| Ambion RNase Inhibitor, cloned | Thermo Fisher Scientific | Cat# AM2682 |
| RNasin Plus RNase Inhibitor | Promega | Cat# N2611 |
| Deoxynucleotide (dNTP) Solution Mix | NEB | Cat# N0447S |
| Lysis Buffer for PCR | TaKaRa | Cat# 9170A |
| SingleCellProtect | AVIDIN | Cat# SCP-250 |
| PEG 8000 Powder, Molecular Biology Grade | Promega | Cat# V3011 |
| Exonuclease I (*E. coli*) | NEB | Cat# M0293S |
| Seq Amp DNA Polymerase | Clontech | Cat# 638504 |
| Agencourt AMPure XP | Beckman coulter | Cat# A63880 |
| Critical Commercial Assays | | |
| HIV-1 p24 antigen ELISA kit | ZetroMetrix | Cat# 0801111 |
| Gal-Screen β-galactosidase reporter assay system | Thermo Fisher Scientific | Cat# T1028 |
| Miseq Reagent kit v3 | Illumina | Cat# MS-102-3001 |
| Deposited Data | | |
| Global IS data (LM-PCR) | This study | GEO: GSE137962 |
| Bulk transcriptome data (dRNA-Seq) | This study | GEO: GSE137644 |
| Single cell transcriptome data (scRNA-Seq) | This study | GEO: DRA008999-DRA009013 |
| Single cell transcriptome data (scRNA-Seq) | (Szabo et al., 2019) | GEO: GSE126030 |
| Experimental Models: Cell Lines | | |
| Human: HEK293T cells | ATCC | CRL-1573 |
| Human: TZM-bl cells | NIH AIDS Reagent Program | Cat# 8129 |
| Human: Jurkat-CCR5 cells | (Ebina et al., 2015) | N/A |
| Experimental Models: Organisms/Strains | | |
| Mouse: NOD.Cg-*Prkdc*$^{scid}$ *Il2rg*$^{tm1Sug}$/Jic (NOG) | Central Institute for Experimental Animals | https://www.ciea.or.jp/en/laboratory_animal/nog.html |
| Oligonucleotides | | |
| *Gag* forward primer for ddPCR: 5′-GGT GCG AGA GCG TCG GTA TTA AG-3′ | (Iwase et al., 2019) | N/A |
| *Gag* reverse primer for ddPCR: 5′-AGC TCC CTG CTT GCC CAT A-3′ | (Iwase et al., 2019) | N/A |
| *ALB* forward primer for ddPCR: 5′-TGC ATG AGA AAA CGC CAG TAA-3′ | (Iwase et al., 2019) | N/A |
| *ALB* reverse primer for ddPCR: 5′-ATG GTC GCC TGT TCA CCA A-3′ | (Iwase et al., 2019) | N/A |
| *Gag* probe for ddPCR: 5′-/56-FAM/AAA ATT CGG/ZEN/TTA AGG CCA GGA GGA AAG AA/3IABkFQ/-3′ | (Iwase et al., 2019) | N/A |
| *ALB* probe for ddPCR: 5′-/5HEX/TGA CAG AGT/ZEN/CAC CAA ATG CTG CAC AGA A/3IABkFQ/-3′ | (Iwase et al., 2019) | N/A |
| Long linker for LM-PCR: 5′-TCA TAT AAT GGG ACG ATC ACA AGC AGA AGA CGG CAT ACG AGA TNN NNN NNN CGG TCT CGG CAT TCC TGC TGA ACC GCT CTT CCG ATC T-3′ | (Satou et al., 2017) | N/A |

(*Continued on next page*)

*Continued*

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Short linker for LM-PCR: 5′-p-GAT CGG AAG AGC GAA AAA AAA AAA AA-3′ | (Satou et al., 2017) | N/A |
| Primer (B3) for 1st LM-PCR: 5′-GCT TGC CTT GAG TGC TTC AAG TAG TGT G-3′ | (Satou et al., 2017) | N/A |
| Primer (B4) for 1st LM-PCR: 5′-TCATGATCAATG GGACGATCA-3′ | (Satou et al., 2017) | N/A |
| Primer (P5B5) for 2nd LM-PCR: 5′-AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG TGC CCG TCT GTT GTG TGA CTC TGG-3′ | (Satou et al., 2017) | N/A |
| Primer (P7) for 2nd LM-PCR: 5′-CAA GCA GAA GAC GGC ATA CGA GAT-3′ | (Satou et al., 2017) | N/A |
| Sequencing primer for LM-PCR (‘Read1’ targeting HIV-1): 5′-ATC CCT CAG ACC CTT TTA GTC AGT GTG GAA AAT CTC-3′ | (Satou et al., 2017) | N/A |
| Sequencing primer for LM-PCR (‘Read2′ targeting human genome): 5′-CGG TCT CGG CAT TCC TGC TGA ACC GCT CTT CCG ATC T-3′ | (Satou et al., 2017) | N/A |
| Sequencing primer for LM-PCR (‘Index1’ targeting adaptor): 5′-GAT CGG AAG AGC GGT TCA GCA GGA ATG CCG AGA CCG-3′ | (Satou et al., 2017) | N/A |
| Oligonucleotide for dRNA-Seq (polyT1): 5′-ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT NNN NNN ANN CNN TNN GNN ANN CNN TTT TTT TTT TTT TTV-3′ | IDT | N/A |
| Oligonucleotide for dRNA-Seq (polyT2): 5′-ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT NNN NNN TNN GNN ANN CNN TNN GNN TTT TTT TTT TTT TTV-3′ | IDT | N/A |
| Oligonucleotide for dRNA-Seq (polyT3): 5′-ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT NNN NNN GNN TNN CNN ANN GNN ANN TTT TTT TTT TTT TTV-3′ | IDT | N/A |
| Oligonucleotide for dRNA-Seq (polyT4): 5′-ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT NNN NNN CNN ANN GNN TNN CNN TNN TTT TTT TTT TTT TTV-3′ | IDT | N/A |
| Oligonucleotide for dRNA-Seq (TS-Oligo; [ ], RNA): 5′-AAG CAG TGG TAT CAA CGC [AGA GUA CAU GGG]-3′ | Hokkaido System Science Co. | N/A |
| Oligonucleotide for dRNA-Seq (PCR primer 1; XXXXXXXX, index): 5′-AAT GAT ACG GCG ACC ACC GAG ATC TAC ACX XXX XXX ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT-3′ | IDT | N/A |
| Oligonucleotide for dRNA-Seq (PCR primer 2; XXXXXXXX, index): 5′-CAA GCA GAA GAC GGC ATA CGA GAT XXX XXX XTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATC TAA GCA GTG GTA TCA ACG CAG AGT ACA TGG G-3′ | IDT | N/A |
| Primer for *CXCR5* cloning: CXCR5_F1, 5′-CGG GGA GCC TCT CAA CAT AA-3′ | This study | N/A |
| Primer for *CXCR5* cloning: CXCR5_R1, 5′-CCC TTA GGA TCC CAG CTC CT-3′ | This study | N/A |
| Primer for *CXCR5* cloning: CXCR5_F2, 5′-AGA GAG AAT TCC CAC CAT GAA CTA CCC GCT AAC GCT-3′ | This study | N/A |

(*Continued on next page*)

*Continued*

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Primer for *CXCR5* cloning: CXCR5_R2, 5′-ATA TAG GAT CCC TAG AAC GTG GTG AGA GAG G-3′ | This study | N/A |
| Recombinant DNA | | |
| Plasmid: pNLCSFV3-EGFP (infectious molecular clone of HIV1-GFP) | (Miura et al., 2001) | N/A |
| Plasmid: pNLCSFV3 (infectious molecular clone of HIV-1 strain NLCSFV3) | (Suzuki et al., 1999) | N/A |
| Plasmid: pCSII-CMV-MCS-IRES2-Bsd | Kindly provided by Dr. Hiroyuki Miyoshi | Cat# RDB04385 |
| Plasmid: pCSII-CMV-CXCR5-IRES2-Bsd | This study | N/A |
| Plasmid: pCAG-HIVgp | (Miyoshi et al., 1998) | Cat# RDB04394 |
| Plasmid: pCMV-VSV-G-RSV-Rev | (Miyoshi et al., 1998) | Cat# RDB04393 |
| Software and Algorithms | | |
| BD FACS Software | BD Biosciences | https://www.bdbiosciences.com/jp/instruments/facsjazz/features/software.jsp |
| FlowJo | Tree Star | https://www.flowjo.com |
| R Statistical Computing software | The R Foundation | https://www.r-project.org |
| QuantaSoft software | Bio-Rad | https://www.bio-rad.com/en-bd/sku/1864011-quantasoft-software-regulatory-edition?ID=1864011 |
| DESeq2 | (Love et al., 2014) | https://bioconductor.org/packages/release/bioc/html/DESeq2.html |
| TopHat2 | (Kim et al., 2013) | https://ccb.jhu.edu/software/tophat/index.shtml |
| GAGE | (Luo et al., 2009) | http://gage.cbcb.umd.edu |
| Cutadapt (v1.16) | (Martin, 2011) | https://github.com/marcelm/cutadapt/tree/v1.16 |
| prinseq-lite (v0.20.4) | (Schmieder and Edwards, 2011) | https://sourceforge.net/projects/prinseq/files/standalone/ |
| STAR (v2.5.3a) | (Dobin et al., 2013) | https://github.com/alexdobin/STAR/releases/tag/2.5.3a |
| featureCounts (v1.6.2) | (Liao et al., 2014) | https://sourceforge.net/projects/subread/files/subread-1.6.2/ |
| Seurat (v3.0.2) | (Stuart et al., 2019) | https://github.com/satijalab/seurat/releases/tag/v3.0.2 |
| GSVA (v1.32.0) | (Hänzelmann et al., 2013) | http://bioconductor.org/packages/release/bioc/html/GSVA.html |
| Python | Python Software Foundation | https://www.python.org |
| BWA (v0.7.17-r1188) | (Li and Durbin, 2009) | http://bio-bwa.sourceforge.net/ |
| Other | | |
| 0.45-μm-pore-size filter | Merck Millipore | Cat# SLHV033RB |

## RESOURCE AVAILABILITY

### Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Kei Sato (ksato@ims.u-tokyo.ac.jp).

### Materials Availability

All unique reagents generated in this study are listed in the Key Resources Table and available from the Lead Contact with a completed Materials Transfer Agreement.

## Data and Code Availability

The accession numbers for the datasets of global ISs (Figure 2), dRNA-seq (Figure 3), and scRNA-Seq (Figure 4) reported in this paper are GEO: GSE137962, GEO: GSE137644, and GEO: DRA008999-DRA009013. The code used in this study is available at GitHub (https://github.com/TheSatoLab/Multi_omics_analysis_HIV1-infected_mice).

## EXPERIMENTAL MODEL AND SUBJECT DETAILS

### Ethics Statement

All animal studies were conducted following the guidelines for the Care and Use of Laboratory Animals of the Ministry of Education, Culture, Sports, Science and Technology, Japan. The authors received approval from the Institutional Animal Care and Use Committees (IACUC)/ethics committee of the institutional review board of Kyoto University (protocol number A16-3-2). All protocols involving human subjects were reviewed and approved by the Kyoto University institutional review board. All human subjects provided written informed consent.

### Humanized Mice

NOD.Cg-*Prkdc*$^{scid}$ *Il2rg*$^{tm1Sug}$/Jic (NOG) mice (Ito et al., 2002) were obtained from the Central Institute for Experimental Animals (Kanagawa, Japan). The mice were maintained under specific-pathogen-free conditions and were handled in accordance with the Regulation on Animal Experimentation at Kyoto University. Human CD34$^+$ HSCs were isolated from human fetal liver and provided by the UCLA CFAR Gene and Cellular Therapy Core Facility, USA. The humanized mice were generated as previously described (Nakano et al., 2017; Sato et al., 2013; Sato et al., 2014; Yamada et al., 2018). Briefly, newborn (aged 0 to 2 days) NOG mice were irradiated (10 cGy per mouse) by an RX-650 X-ray cabinet system (Faxitron X-ray Corporation) and were then intrahepatically injected with the obtained human CD34$^+$ cells (1.0 to 2.5 × $10^5$ cells). In this study, 21 male and 22 female NOG mice were used for the recipient of human HSC transplantation.

### Cells and Viruses

HEK293T cells (a human embryonic kidney cell line; ATCC CRL-1573) and TZM-bl cells (obtained through the NIH AIDS Research and Reference Reagent Program) (Wei et al., 2002) were maintained in Dulbecco's modified Eagle's medium (Sigma-Aldrich, Cat# D6046-500ML) containing 10% fetal calf serum (FCS; Sigma-Aldrich, Cat# 172012-500ML), 2 mM L-glutamate (Thermo Fisher Scientific, Cat# 25030081) and 1% penicillin streptomycin (Sigma-Aldrich, Cat# P4333-100ML). Human peripheral blood mononuclear cells (PBMCs) were isolated from human PB using Ficoll-Paque (GE Healthcare) according to the manufacturer's protocol.

The infectious molecular clones (IMCs) of NLCSFV3-GFP (a GFP-expressing HIV-1 strain NLCSFV3 [HIV1-GFP]; Figure S1A) (Miura et al., 2001) and wild-type NLCSFV3 (Suzuki et al., 1999) were prepared in our previous studies. For virus production, HEK293T cells were transfected with IMC using the calcium phosphate method. At 48 hours post-transfection, the culture supernatants were harvested, centrifuged, and then filtered through a 0.45-μm-pore-size filter. The titration of virus infectivity was performed as previously described (Sato et al., 2012). Briefly, human PBMCs (1 × $10^7$ cells) were stimulated by PHA for 3 days. Then, the prepared virus was serially diluted, and 100 μL of the diluted virus was inoculated onto PHA-stimulated human PBMCs (2 × $10^5$ cells) in a 96-well plate in triplicate. Half of the culture media was refreshed for a few days, and the endpoint was determined using an HIV-1 p24 antigen enzyme-linked immunosorbent assay (ELISA) kit (ZetroMetrix) at 14 days postinfection. Virus infectivity was calculated as the 50% tissue culture infectious dose ($TCID_{50}$) according to the Reed-Muench method (Reed and Muench, 1938). In some experiments, the infectious HIV-1 yield was determined using TZM-bl indicator cells as previously described (Yamada et al., 2018). Briefly, 5,000 cells were seeded in 96-well plates and infected in triplicate with cell culture supernatants. At 48 h postinfection, the infection rates were measured using a Galacto-Star mammalian reporter gene assay system (Roche) and a 2030 ALBO X multilabel counter instrument (PerkinElmer) according to the manufacturers' procedure.

## METHOD DETAILS

### Western Blotting

Western blotting was performed as previously described (Yamada et al., 2018) by using the following antibodies: anti-HIV-1 p24 antibody (Virostat, Cat# 1951), anti-Vif antibody (obtained from the NIH AIDS Reagent Program, Cat# 6459), anti-Vpr antibody (Cosmo Bio, Cat# NCG-M01), anti-Vpu antibody (obtained from the NIH AIDS Reagent Program, Cat# 969), anti-Env antibody (obtained from the NIH AIDS Reagent Program, Cat# 12559), anti-Nef antibody (Thermo Fisher Scientific, Cat# MA1-71501), anti-PR antibody (Thermo Fisher Scientific, Cat# MA1-19015), anti-RT antibody (obtained from the NIH AIDS Reagent Program, Cat# 11338), anti-IN antibody (Abcam, Cat# ab66645), anti-GFP antibody (Sigma-Aldrich, Cat# G6795), and anti-alpha-Tubulin (Sigma-Aldrich, Cat# T9026).

### HIV1-GFP Infection of Humanized Mice

HIV1-GFP (500,000 $TCID_{50}$ [equivalent to 39 ng p24 antigen]) was intraperitoneally inoculated into the humanized mice. Twenty seven (13 male; 14 female) humanized mice were inoculated with HIV1-GFP, while sixteen (8 male; 8 female) humanized mice were

inoculated with RPMI1640 (Sigma-Aldrich, Cat# R8758-500ML; for mock infection). PB was collected at 0 and 2 wpi from the retro-orbital venous plexus under anesthesia. At 2 wpi, these mice were euthanized and sacrificed, and the spleens were collected as previously described (Yamada et al., 2018). For collection of splenic human mononuclear cells (MNCs), the spleens were crushed and rubbed on a steal mesh with 1-mm grids to generate single cell suspensions in RPMI 1640 (Sigma-Aldrich) supplemented with 4% FCS (Sigma-Aldrich). The splenic cell suspension was separated using Ficoll-Paque (GE Healthcare) as previously described (Yamada et al., 2018). The amount of HIV-1 RNA in 50 μL of plasma was quantified by Bio Medical Laboratories, Inc. (the detection limit of HIV-1 RNA is 800 copies/ml).

### Hematology Analysis, Flow Cytometry and Cell Sorting

Hematology analysis was performed with a Celltac α MEK-6450 (Nihon Kohden Co.) as previously described (Yamada et al., 2018). Flow cytometry was performed with FACSCalibur and FACSJazz systems (BD Biosciences) as previously described (Yamada et al., 2018), and the obtained data were analyzed with BD FACS Software (BD Biosciences) and FlowJo (Tree Star, Inc.). For flow cytometry analysis, anti-CD45-PerCP antibody (clone HI30; Biolegend), anti-CD3-PE (clone HIT3a; Biolegend), anti-CD8-APC (clone DK25; Dako), and anti-CXCL13-APC (Thermo Fisher Scientific, Cat# MA5-23629) were used. Splenic human $CD4^+$ T cells ($CD45^+CD3^+CD8^-$ cells) were purified by using FACSJazz (BD Biosciences) as previously described (Yamada et al., 2018), and the purity was > 99% (see also Figures S1D and S3).

### Stimulation with PMA and Ionomycin

The sorted $GFP^-$ cells were maintained in RPMI 1640 (Sigma-Aldrich) containing FCS (Sigma-Aldrich) and 1% penicillin streptomycin (Sigma-Aldrich). For induction of viral activation, these cells were stimulated with 50 ng/ml PMA (Sigma-Aldrich, P-8139) and 1 μM ionomycin (Sigma-Aldrich, I-0634). For inhibition of new viral infection, 10 nM nelfinavir (Sigma-Aldrich, PZ0013) was supplied. At 48 hours poststimulation, the $GFP^+$ and $GFP^-$ cells were sorted as described above.

### Quantification of HIV-1 DNA by ddPCR

HIV-1 DNA was quantified by ddPCR targeting a conserved part of the HIV-1 *gag* gene and then normalized to the copy number of the *ALB* gene according to previous reports with minor modifications (Douek et al., 2002; Iwase et al., 2019). Briefly, the ddPCR reaction mixtures were loaded onto a Bio-Rad QX200 droplet generator. The generated droplets were then transferred to a 96-well PCR plate and sealed with a preheated PX1 PCR plate sealer (Bio-Rad) for 5 s at 180°C. PCR cycles were performed in a C1000 Touch thermal cycler (Bio-Rad) with the following parameters 95°C for 10 minutes followed by 39 cycles of 94°C for 30 s, 58°C for 60 s, 98°C for 10 minutes and a hold at 4°C. The plate was then placed in the Bio-Rad QX200 droplet reader for quantification of the number of positive and negative droplets based on fluorescence. The threshold value for ddPCR was determined based on the highest value of droplet fluorescence in the no template control to provide an objective cut-off with maximum sensitivity. The copy number of targets was calculated by using the manufacturer's software (Bio-Rad QuantaSoft v1.7.4), which is assumed to follow a Poisson distribution. Then, proviral load was calculated as follows: proviral load (per cell) = (copy number of gag) / [(copy number of albumin) / 2]. The following primers and probes were used: *gag*-forward, 5′-GGT GCG AGA GCG TCG GTA TTA AG-3′; *gag*-reverse, 5′-AGC TCC CTG CTT GCC CAT A-3′; *ALB*-forward, 5′-TGC ATG AGA AAA CGC CAG TAA-3′; *ALB*-reverse, 5′-ATG GTC GCC TGT TCA CCA A -3′; *gag*-probe, 5′-/56FAM/AAA ATT CGG/ZEN/TTA AGG CCA GGA GGA AAG AA/3IABkFQ/-3′; *ALB*-probe, 5′-/5HEX/TGA CAG AGT/ZEN/CAC CAA ATG CTG CAC AGA A/3IABkFQ/-3′.

### LM-PCR for Global IS Analysis

LM-PCR and next-generation sequencing were performed as previously described (Satou et al., 2017). Briefly, splenic $GFP^+$ and $GFP^-$ cells were obtained from eight infected mice, and DNA was extracted using a QIAamp DNA blood mini kit (QIAGEN). Then, up to 150 ng genomic DNA was sheared by sonication with a Picoruptor (Diagenode) instrument to a size of 300-400 bp in length. DNA end repair and addition of 3′ adenosine was performed with NEBNext Ultra II End Repair/dA-Tailing Module (New England Biolabs). Linker was then ligated to the ends using NEBNext Ultra II Ligation Module (New England Biolabs). The ligated product was amplified with primers targeting the 3′-LTR on one end and the linker on the other end. The following thermal protocol was used for both the first and second PCR: 96°C 30 s (1 cycle); 94°C 5 s, 72°C 1 minute (7 cycles); 94°C 5 s, 68°C 1 minute (13 cycles); 68°C 9 minutes (1 cycle); hold at 4°C until user stops. PCR amplicons were cleaned and pooled to form a library, which was then quantified using primers for Illumina P5 and P7. DNA libraries were sequenced using Illumina NextSeq 500 to acquire paired-end reads plus an 8 bp index read. The oligonucleotides used to perform LM-PCR and sequencing are listed in KEY RESOURCES TABLE. The analytical pipeline of LM-PCR was described in a previous study (Satou et al., 2017).

### Enrichment Analysis of HIV-1 the ISs for Specific Histone Marks

ChIP-Seq data for specific histone modifications (H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3, H3K9ac, H3K9me3) in primary T regulatory cells were downloaded from the Roadmap Epigenomics data portal (http://www.roadmapepigenomics.org/). Fold enrichments of the HIV-1 ISs in genomic regions with the specific histone modifications were calculated against the random expectations that were based on the 1000 genomic permutations generated by bedtools shuffle (Quinlan and Hall, 2010).

### Orientation-Bias Analysis of HIV-1 Integration

For analysis of the orientation of the HIV-1 ISs located within the host gene body regions, the numbers of HIV-1 ISs in the same or opposite orientations were counted using bedtools intersect (Quinlan and Hall, 2010) with -s or -S options, respectively. The orientation bias was statistically compared between $GFP^+$ and $GFP^-$ cells using two-sided Fisher's exact test.

### dRNA-Seq

For dRNA-Seq (Ogawa et al., 2017; Shiroguchi et al., 2012), a library was prepared from one hundred cells collected in a tube by cell sorting (Tenno et al., 2017). Splenic $GFP^+$ and $GFP^-$ cells were always both obtained from the same mouse (in total twelve libraries from six mice); for both $GFP^+$ and $GFP^-$ cells, one library (i.e., one tube) from each of three mice was sequenced, for a total of three libraries. Three additional libraries from each of three other mice were also subjected to sequencing for a total of nine additional libraries. Splenic $CD4^+$ T cells were obtained from seven mock-infected mice in total (13 libraries total); one library from each of four mice was sequenced for a total of four libraries. Three additional libraries from each of the three other mice were also sequenced, for a total of nine additional libraries. In a single MiSeq run (150 cycles, Illumina), five or six libraries were sequenced together using different sample indexes. Sequencing data were mapped against the human genome (hg38 assembly from the UCSC Genome Browser) with human gene annotation (hg38 refFlat from the UCSC Genome Browser) using TopHat2 (Kim et al., 2013). DEGs of dRNA-Seq were determined using DESeq2 (Love et al., 2014). Genes with a log2-fold change of $> 0.5$ or $< -0.5$ and FDR $< 0.05$ were considered as differentially expressed. *P value*s were calculated with the Wald test and were corrected for multiple testing issues using the Benjamini-Hochberg procedure.

### GO Enrichment for Transcriptome Data

Enrichment analysis of transcriptome data was based on the GO hierarchy. *P value*s for the enrichment test were calculated by the GAGE algorithm (Luo et al., 2009), and the FDR was calculated from the *p* value for multiple testing. The enrichment results (Figures 3F and 3G) were visualized using an in-house algorithm based on d3.js.

### Weighted Parametric Gene Set Analysis

Enrichment analysis to evaluate the effects of TFs on their binding target genes based on numerous published ChIP-seq data was performed as described previously (Kawakami et al., 2016).

### scRNA-Seq

Splenic $GFP^+$ and $GFP^-$ cells were obtained from four infected mice, and splenic $CD4^+$ T cells were obtained from five mock-infected mice. These cells were used for the scRNA-Seq analyses. scRNA-Seq was performed as previously described (Suzuki et al., 2015). Briefly, approximately 10,000 cells were applied to the C1 system (Fluidigm), and 96 cells were captured in the flow cells and separated into independent chambers (small IFC 5-10 μm, Fluidigm). Before scRNA-Seq library construction, we manually examined each chamber under a microscope. The RNA-Seq libraries were constructed according to the manufacturers' instructions as follows: first-strand cDNA was synthesized and further amplified using the SMARTer system (Clontech). Illumina sequencing libraries were constructed using a Nextera XT DNA Sample Preparation kit (Illumina). After the evaluation of the quality and quantity of the constructed RNA-Seq libraries using a BioAnalyzer (Agilent Technologies, Santa Clara, CA, USA), sequencing was performed on the HiSeq2500 platform with a 36-base single-end read.

### Processing of scRNA-Seq Data

Before scRNA-Seq library construction, we manually examined each chamber of the Fluidigm C1 system under a microscope. Based on this information, we only analyzed the library derived from the chamber that contains a single living cell. From the respective sequence reads, PCR adapters [AAG CAG TGG TAT CAA CGC AGA GTA $CT_{30}$ (SMARTer II A Oligonucleotide) and AAG CAG TGG TAT CAA CGC AGA GTA C (3′ SMART CDC Primer II A)] were trimmed using cutadapt (v1.16) (Martin, 2011). In the adaptor trimming, the maximum sequence error rate was set at 0.2, and the minimum overlap sequence length was set at 10. Sequence reads shorter than 18 bp were discarded in this step. Subsequently, low-quality reads (the mean sequence quality score $< 25$) were discarded using prinseq-lite (v0.20.4) (Schmieder and Edwards, 2011).

The filtered sequence reads were mapped to the custom genome sequence using STAR (v2.5.3a) (Dobin et al., 2013). The custom genome sequence comprises the sequence of the human reference genome (hg38) and a partial sequence of HIV1-GFP, in which untranscribed regions (5′ U3 and 3′ U5 sequences) were excluded. In the mapping, the following parameters were used [--outFilterMismatchNmax 2,--outSJfilterOverhangMin 12 8 8 8,--chimSegmentMin 8, and--chimJunctionOverhangMin 8]. Since there are various splicing patterns in HIV-1 transcripts, we performed two-pass mapping as follows: before the mapping, we concatenated the fastq files of all the scRNA-Seq libraries. In the first mapping, we used the concatenated data with the gene annotation GENCODE (v22) (https://www.gencodegenes.org) and identified splice junctions that are not recoded in the gene annotation. Subsequently, splicing junction information supported by $\geq 50$ reads was extracted. In the second mapping, we used each dataset of the scRNA-Seq library both with the GENCODE gene annotation and the splicing junction information detected in the first mapping.

The gene expression count matrix was generated using featureCount (v1.6.2) (Liao et al., 2014) with the GENCODE gene annotation. In addition, reads mapped to the HIV-1 genome were also counted in this step. From the expression matrix, data of the cells in

which fewer numbers (Z score < −2) of expressed genes were detected were excluded. In addition, data of the cells in which an abnormal proportion (|Z score| > 1.5) of reads was assigned to mitochondrial genes were also excluded. From the expression matrix, genes with low expression [< 0.5 reads per kilobase of exon per million mapped reads (RPKM) in > 90% cells] were excluded. In addition, mitochondrial genes and genes encoding T cell receptors were also excluded.

To normalize the expression data by adjusting batch effects among humanized mouse individuals, we used the R package "Seurat (v3.0.2)" (Stuart et al., 2019). This analysis was based on the pipeline presented in the Seurat tutorials (https://satijalab.org/seurat/). First, the expression matrix was log2-transformed. In this step, we used the data of 2,000 genes with the strongest expression (including HIV-1) to reduce the effects of dropouts on read counts, which is a characteristic of scRNA-Seq. Subsequently, we applied the imputation method adaptively thresholded low rank approximation (ALRA) (Linderman et al., 2018) to the expression matrix using the function "RunALRA." After imputation, the expression matrix was split into those of the respective mice. The genes whose expression levels were variable in each mouse were extracted using the function "FindVariableFeatures" with the parameter [selection.method = mvp]. The expression matrices of the mice were merged with adjustment of the batch effects among the mice as follows: To identify the cell subpopulations used as "anchors" (Stuart et al., 2019), the function "FindIntegrationAnchors" was applied to the expression matrix with parameters [dims = 1:25, k.filter = 20, k.score = 20, anchor.features = 2000, max.features = 2000]. Subsequently, to merge the expression matrices while considering the "integration anchors," the function "IntegrateData" was applied to those matrices with the parameter [dims = 1:25]. The merged gene expression matrix was normalized as the Z-score. The Z-scored expression matrix was used in the downstream analyses unless otherwise noted.

### Uniform Manifold Approximation and Projection (UMAP) Visualization

To visualize the subpopulation structure of the cells, we employed a dimension reduction method, UMAP (McInnes et al., 2018). In this analysis, the first 20 principal components of the expression data were used. The number of neighboring points used in local approximations of the manifold structure was set at 20, and the minimum distance was set at 0.01.

### Unsupervised Clustering of scRNA-Seq Data

Clustering of the cells was performed using Seurat (v3.0.2) (Stuart et al., 2019). This analysis was based on the pipeline presented in the Guided Clustering Tutorial (https://satijalab.org/seurat/v3.0/pbmc3k_tutorial.html). Principal component analysis was performed based on the normalized expression matrix without the HIV-1 expression data. Subsequently, the cells were clustered according to the first 20 PCs using two functions, "FindNeighbors" and "FindClusters." When running the function "FindNeighbors," two parameters were changed [dim = 1:20, k.param = 9]. When running the function "FindClusters," the resolution parameter was set at 0.9.

### Identification of Signature Genes in Each Cluster

To determine the characteristics of each cluster, we extracted signature genes, whose expression levels in cells of a cluster were higher than those in the other cells. This analysis was performed using Seurat (v3.0.2) (McInnes et al., 2018; Stuart et al., 2019). We first identified the DEGs using the following three methods: 1) logistic regression test based on the imputed expression data with adjustment for the cell type effects; 2) logistic regression test based on the nonimputed expression data with adjustment for the cell type effects; and 3) Wilcoxon rank sum test based on the imputed expression data, which was performed separately in the datasets of the $GFP^+$ cells, the $GFP^-$ cells, or the mock-infected cells. A gene was regarded as a DEG if the absolute value of the fold change was greater than 1.5 and if the FDR was less than 0.01. FDRs were calculated with the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). A gene was regarded as a signature gene if it was defined as a DEG in all three methods above. In the third method, genes that were regarded as DEGs in any of the three datasets ($GFP^+$ cells, $GFP^-$ cells, and mock-infected cells) were defined as DEGs.

Based on the DEGs, clusters 1, 2, 3, 4 and 8 were composed of Th1-like cells (*IFNG*$^{high}$, *GZMA*$^{high}$, *CCR7*$^{low}$, *SELL*$^{low}$), activated $CD4^+$ T-like cells (*HLA-DRs*$^{high}$, *PDCD1*$^{high}$), naive $CD4^+$ T-like cells (*CCR7*$^{high}$, *SELL*$^{high}$, *CTLA4*$^{low}$, *HLA-DRs*$^{low}$, *IFNG*$^{low}$), Tfh-like cells (*TNFSF8*$^{high}$, *SH2D1A*$^{high}$, *BTLA*$^{high}$) and Treg-like cells (*CTLA4*$^{high}$, *TNFRSF1B*$^{high}$, *TIGIT*$^{high}$, *TNFRSF18*$^{high}$) (Figure 4D).

### Calculation of the GSVA Score

To calculate the gene setwise expression scores, we employed GSVA (v1.32.0) (Hänzelmann et al., 2013). The algorism "ssgsea" was selected. We used the three gene sets: 1) "common ISGs," which were defined in our previous paper (Aso et al., 2019); 2) "NF-κB regulated genes," which were acquired from a website (https://bioinfo.lifl.fr/NF-KB/); and 3) "NF-κB signaling related genes," which correspond to the gene set "GO_I_KAPPAB_KINASE_NF_KAPPAB_SIGNALING" in MSigDB (v6.2) (https://www.gsea-msigdb.org/gsea/msigdb/collections.jsp). First, each GSVA score was calculated using each gene set with the expression data of the 20,000 most highly expressed genes. Subsequently, we examined Pearson's correlation between the GSVA score and the expression level of each gene member of the gene set, and the lower 1/4 genes regarding the correlation score were removed from the gene set. Finally, we again calculated each GSVA score using the refined gene sets with the expression levels of the most expressed 20,000 genes.

### Identification of the DEGs between Cluster 5 and Cluster 7

To identify the DEGs between cluster 5 and cluster 7, we performed a logistic regression test with adjustment of the cell type effects. A gene with < 0.05 FDR regarded as a DEG.

### Experimental Evaluation of the Effect of CXCL13 on HIV-1 Infection

The lentiviral vector expressing CXCR5, the receptor for CXCL13, was prepared as described previously (Yamada et al., 2018). First, the open reading frame of the *CXCR5* gene was amplified by RT-PCR using cDNA obtained from human PBMCs for the template and the following primers: CXCR5_F1, 5′-CGG GGA GCC TCT CAA CAT AA-3′; and CXCR5_R1, 5′-CCC TTA GGA TCC CAG CTC CT-3′. Then, the second PCR was performed using the first RT-PCR product as the template and the following primers: CXCR5_F2, 5′-AGA GAG AAT TCC CAC CAT GAA CTA CCC GCT AAC GCT-3′; and CXCR5_R2, 5′-ATA TAG GAT CCC TAG AAC GTG GTG AGA GAG G-3′. The second PCR product was digested with EcoRI and BamHI and inserted into pCSII-CMV-MCS-IRES2-Bsd (kindly provided by Dr. Hiroyuki Miyoshi) to construct a CXCR5-expressing plasmid (pCSII-CMV-CXCR5-IRES2-Bsd). For preparation of lentiviral vectors, pCAG-HIVgp and pCMV-VSV-G-RSV-Rev were cotransfected with pCSII-CMV-MCS-IRES2-Bsd (for empty lentiviral vector) or pCSII-CMV-CXCR5-IRES2-Bsd (for CXCR5-expressing lentiviral vector) into HEK293T cells using the calcium phosphate method (Miyoshi et al., 1998). At 48 h post-transfection, the culture supernatants were harvested, centrifuged, and then filtered through a 0.45-μm-pore-size filter. Then, the Jurkat-CCR5 cells (Ebina et al., 2015) were transduced with empty or CXCR5-expressing lentiviral vector and were cultured with RPMI 1640 containing 10% FCS (Sigma-Aldrich), 2 mM L-glutamate (Thermo Fisher Scientific), 1% penicillin streptomycin (Sigma-Aldrich), and blasticidin (10 μg/ml) (InvivoGen, Cat# ant-bl-1) for selection. After 2-3 weeks, the empty lentiviral vector-transduced cells and CXCR5-expressing lentiviral vector-transduced cells were stained with anti-CXCR5-APC antibody (BioLegend, Cat# 35908), and the expression level of surface CXCR5 was analyzed by flow cytometry.

For analysis of the effect of CXCL13 on HIV-1 infection, $2 \times 10^5$ of empty lentiviral vector-transduced or CXCR5-expressing lentiviral vector-transduced Jurkat-CCR5 cells were infected with NLCSFV3-EGFP at a multiplicity of infection 0.2 without or with recombinant CXCL13 (10-1,000 ng; Funakoshi, Cat# 801-CX-025). At 48 hours postinfection, the percentage of infected cells (i.e., $GFP^+$ cells) was analyzed by flow cytometry.

### Identification of the $ISG^{low}$ Subpopulation in the Human Primary $CD4^+$ T Cells

To validate the existence of the $ISG^{low}$ subpopulation ('cluster 5' in humanized mice) in the human primary $CD4^+$ T cells, we analyzed the scRNA-Seq dataset of human $CD4^+$ T cells obtained from lymph nodes (Szabo et al., 2019). In this analysis, we used the processed count matrix (GSE126030). Importantly, we only used the data of nonstimulated $CD4^+$ T cells derived from two donors. First, since these data consist of both $CD4^+$ T cells and $CD8^+$ T cells, the data of $CD4^+$ T cells were extracted according to the cell annotation by Szabo et al. Next, log2 transformation, imputation, and standardization of the expression data were performed. Unsupervised clustering was performed using the same genes as the humanized mouse data. In this clustering, the same procedures and parameters were used as the humanized mouse data except a number of principal components (the first 100 principal components of the expression data were used). The GSVA scores of common ISGs were calculated by same methods as the humanized mouse data. To evaluate the similarity of the gene expression pattern between each cluster in human samples and cluster 5 of humanized mouse data, we calculated Spearman's correlation coefficient of the mean normalized expression levels between the two clusters. For this calculation, only the genes used in unsupervised clustering were used. Importantly, in this calculation, common ISGs were excluded to examine the transcriptomic similarity without the effect of ISGs.

## QUANTIFICATION AND STATISTICAL ANALYSIS

Unless otherwise stated, data analyses were performed using Prism 6 software (GraphPad). The data are presented as the mean ± SEM. Statistically significant differences were determined by Mann-Whitney U tests, Student's t tests, or paired t tests. Statistical comparison of the HIV-1 ISs between the $GFP^+$ and $GFP^-$ cells (Figure 2H) was performed by Fisher's exact test. Spearman's rank correlation coefficient ($r_s$) is applied to determine statistically significant correlations.

**Cell Reports, Volume *32***

## Supplemental Information

# Multiomics Investigation Revealing the Characteristics of HIV-1-Infected Cells *In Vivo*

**Hirofumi Aso, Shumpei Nagaoka, Eiryo Kawakami, Jumpei Ito, Saiful Islam, Benjy Jek Yang Tan, Shinji Nakaoka, Koichi Ashizaki, Katsuyuki Shiroguchi, Yutaka Suzuki, Yorifumi Satou, Yoshio Koyanagi, and Kei Sato**

**Supplemental Information**

# Multiomics investigation revealing the characteristics of HIV-1-infected cells *in vivo*

Hirofumi Aso, Shumpei Nagaoka, Eiryo Kawakami, Jumpei Ito, Saiful Islam, Benjy Jek Yang Tan, Shinji Nakaoka, Koichi Ashizaki, Katsuyuki Shiroguchi, Yutaka Suzuki, Yorifumi Satou, Yoshio Koyanagi, Kei Sato

Supplemental Figures S1-S4
Supplemental Tables S1-S4

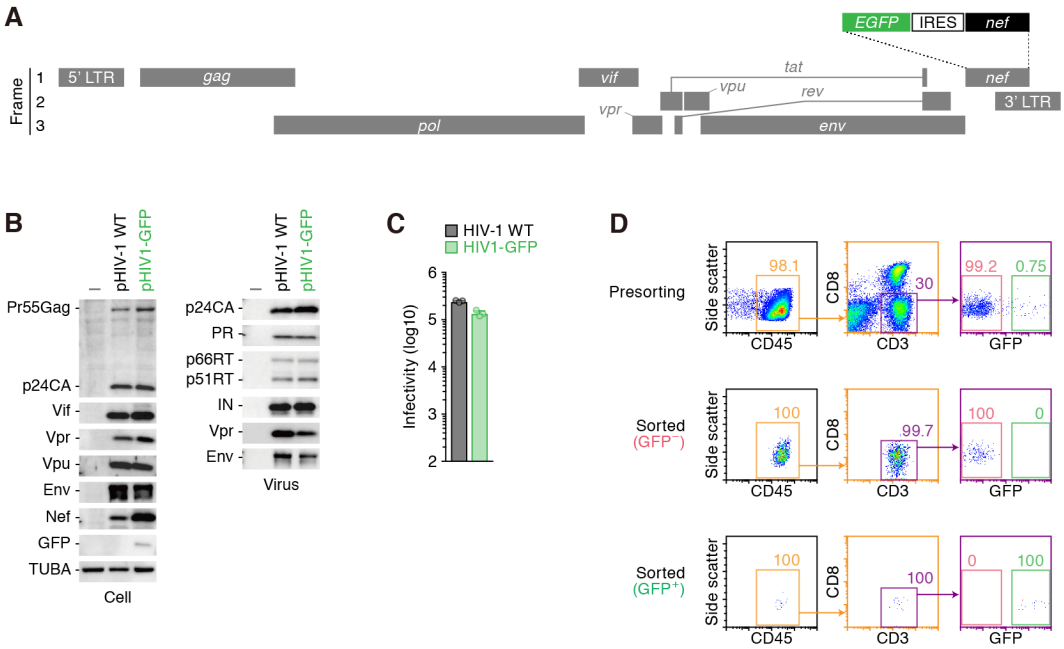



**Figure S1. Characterization of HIV1-GFP and verification of the sorted cell populations (Related to Figure 1).**

(A) Scheme of the HIV1-GFP genome.

(B and C) Viral profile of the HIV1-GFP virus used in this study. (B) Expression of viral proteins as well as GFP in the transfected cells (left) and viral supernatant (right) were analyzed by Western blotting. (C) Viral infectivity was measured by TZM-bl assay. In B and C, wild-type HIV-1 ("pHIV-1 WT", an infectious molecular clone of parental virus, strain NLCSFV3) was used as a control.

(D) Representative data showing the purity of the sorted cells. The $CD45^{+}CD3^{+}CD8^{-}$ cells were gated (left and middle) and these cells were separated into $GFP^{+}$ (green) and $GFP^{-}$ (pink) cells (right). Each number on the square in the dot plot indicates the percentage of gated cells.

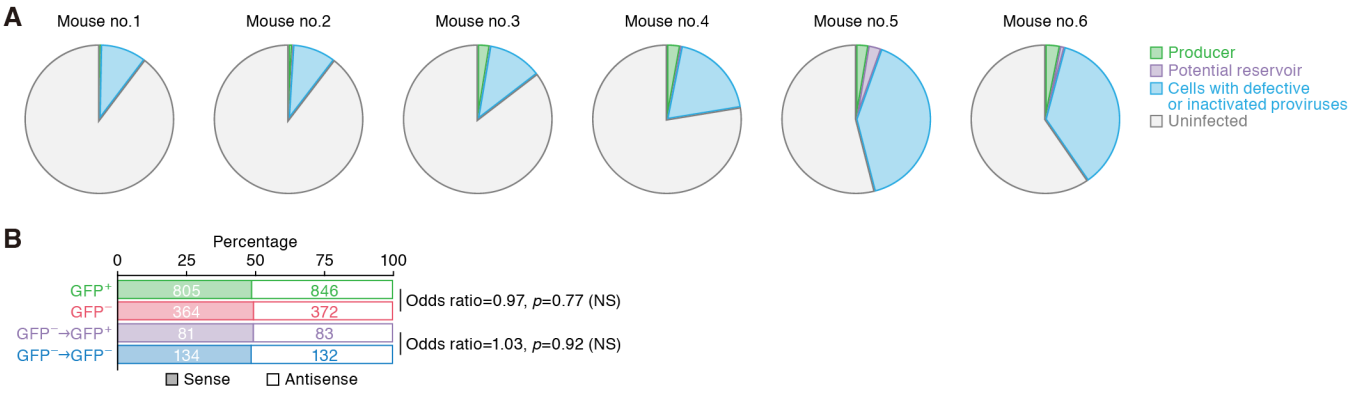



**Figure S2. Proportion of each population in the CD4⁺ T cells of HIV1-GFP-infected humanized mice (Related to Figure 2).**

(A) Pie charts of the proportion of each population (producer, potential reservoir, the cells with defective or inactivated proviruses, and uninfected cells) in the CD4⁺ T cells of six HIV1-GFP-infected humanized mice are shown.

(B) Orientation of HIV-1 integration relative to the host gene. The number in the bar indicates the number of ISs in each population. Statistical analysis was performed by two-sided Fisher's exact test. NS, no statistical significance.

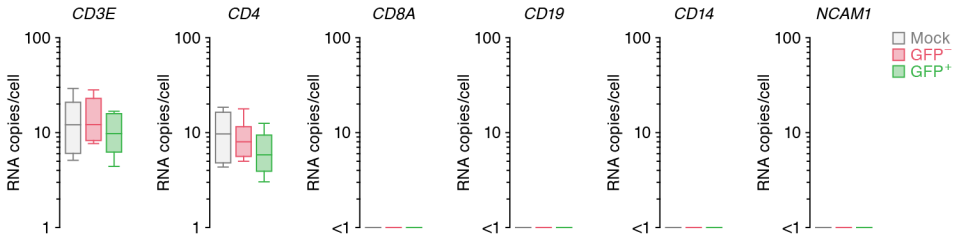



**Figure S3. Verification of the sorted cell populations by dRNA-Seq (Related to Figure 3).** Expression levels of cell lineage marker genes in each cell population. *CD3E* for T cells, *CD4* for $CD4^+$ cells, *CD8A* for $CD8^+$ T cells, *CD19* for B cells, *CD14* for monocytes, and *NCAM1* for natural killer cells. The y-axis indicates the RNA copy number per cell. Note that the transcripts of *CD8A*, *CD19*, *CD14* and *NCAM1* were undetectable.

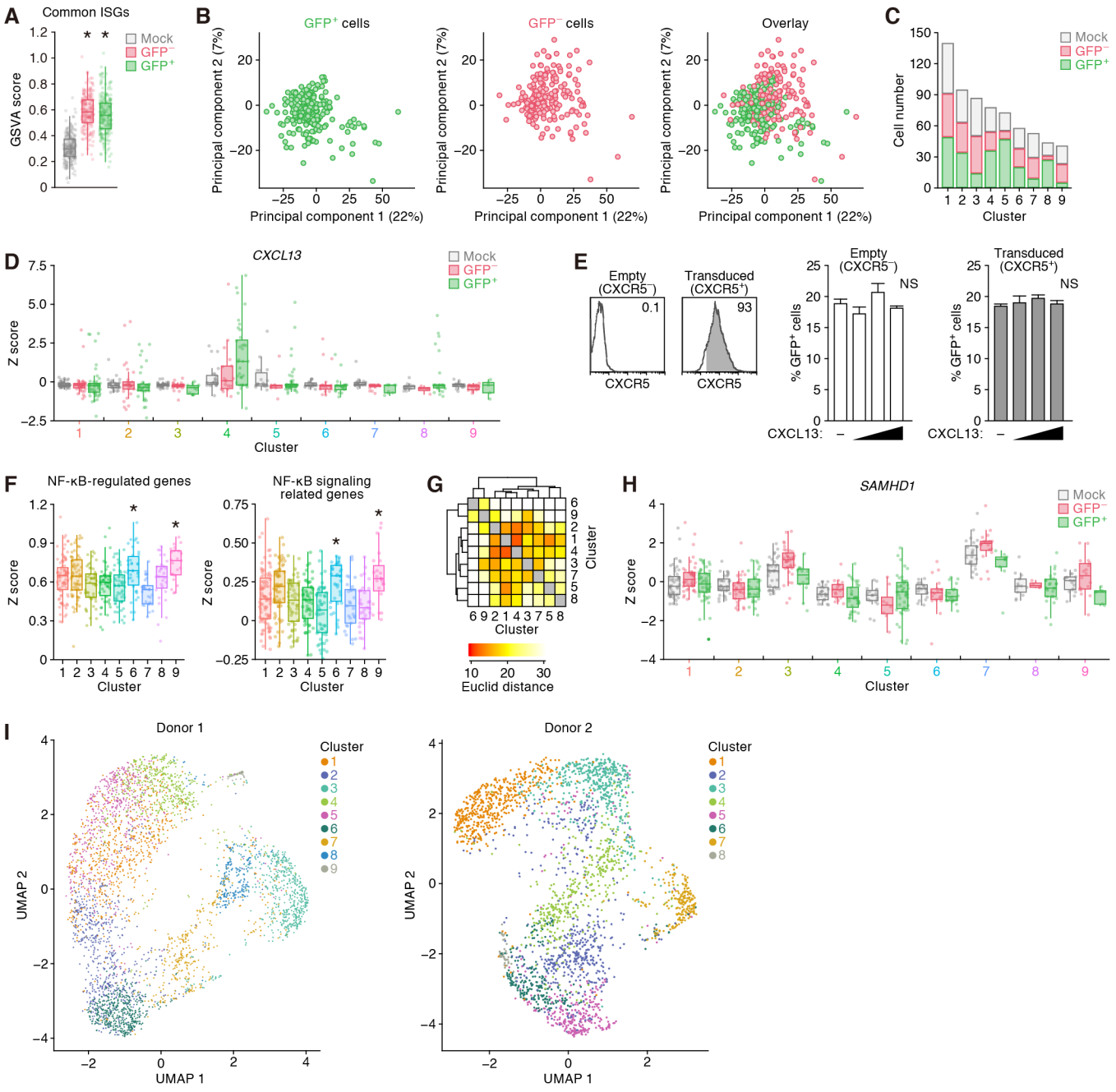



**Figure S4. scRNA-Seq analysis of the CD4⁺ T cells from HIV1-GFP infected (or mock-infected) mice (Related to Figure 4).**

(A) GSVA score of common ISGs in each cluster. Unlike **Figure 4H**, the result includes the data of GFP⁺ cells, GFP⁻ cells, and mock-infected CD4⁺ T cells.

(B) Principal component analyses representing the gene expression patterns of GFP⁺ cells (left) and GFP⁻ cells (middle). A merged plot is shown on the right. The dot is colored according to the cell category.

(C) The absolute numbers of GFP⁺ cells, GFP⁻ cells, and mock-infected CD4⁺ T cells in each cluster.

(D) Normalized expression levels of *CXCL13*. Unlike **Figure 4F**, the results were classified into GFP⁺ cells, GFP⁻ cells, and mock-infected CD4⁺ T cells.

(E) Effect of CXCL13 on HIV-1 infection. (Left) The Jurkat-CCR5 cell line stably expressing CXCR5 was prepared as described in **STAR★METHODS**, and the expression level of surface CXCR5 was analyzed by flow cytometry. Representative results of surface CXCR5 expression on empty lentiviral vector-transduced cells ["Empty ($CXCR5^-$)"] and CXCR5-expressing lentiviral vector-transduced cells ["Transduced ($CXCR5^+$)"] are shown. The number in the histogram indicates the percentage of $CXCR5^+$ cells. Note that endogenous CXCR5 is not expressed in the parental Jurkat-CCR5 cells. (Right) HIV1-GFP was inoculated into these cells at multiplicity of infection 0.2 without or with recombinant CXCL13 (10, 100, or 1,000 ng). At 48 hours postinfection, the percentage of infected cells (i.e., $GFP^+$ cells) was analyzed by flow cytometry. This assay was performed in triplicate, and the averages are shown with the standard deviation. NS, no statistical significance.
(F) GSVA scores of 'NF-κB regulated genes' (left) and 'NF-κB signaling related genes' (right) in $GFP^+$ cells and $GFP^-$ cells in each cluster.
(G) A heatmap showing pairwise Euclid distances representing the gene expression difference among clusters. The distance was calculated using PC 1–20 of the gene expression data.
(H) Normalized expression levels of *SAMHD1*. Unlike **Figure 4I**, the results were classified into $GFP^+$ cells, $GFP^-$ cells, and mock-infected $CD4^+$ T cells.
(I) UMAP plots representing the gene expression patterns in human $CD4^+$ T cells of two healthy individuals. Each dot is colored according to cell category, and the colors are identical to those in **Figure 4J**.