HIV-1 Integration Site Determines the Transcriptional Fate and Persistence of Integrated Proviruses

PAL VK[1], DANESH A[2], CANIS M[1], DILLING RT[2], MILLER GI[2], HUYNH TT[2], COPERTINO, D[2], BARROWS D[3], CARROLL T[3], HATZIIOANNOU T[1], JONES RB[2], LEE GQ[2], MUECKSCH F[1,4], BIENIASZ PD[1,5]

[1]Laboratory of Retrovirology, The Rockefeller University, New York.

[2]Division of Infectious Diseases, Department of Medicine, Weill Cornell Medical College, New York.

[3]Bioinformatics Resource Center, The Rockefeller University, New York.

[4]Department of Infectious Diseases, Virology, Heidelberg University, Heidelberg, Germany.

[5]Howard Hughes Medical Institute, The Rockefeller University, New York.

Correspondence to:
Frauke Mueksch (frauke.muecksch@uni-heidelberg.de)
Paul Bieniasz (pbieniasz@rockefeller.edu)

Summary

The mechanisms by which latent HIV-1 reservoirs persist during antiretroviral therapy is incompletely understood. Here, we derive a model system to measure clonal expansion and viral latency in which populations of human memory $CD4^+$ T cells, each bearing a single transcriptionally active HIV-1 provirus are engrafted into immunodeficient mice. Over ~2 months *in vivo*, clonal expansion and the establishment of latency occurred in subsets of engrafted infected cells. Clonal expansion *in vivo* was driven by T-cell receptor identity, but not by proviral insertional mutagenesis. The integration sites of proviruses that became latent *in vivo* were enriched on chromosome 19, in intergenic and centromeric satellite regions, and genes whose expression is atypically low. Pre-existing repressive epigenetic features were associated with latency for subsets of proviruses. Our findings suggest a confluency of genomic and epigenomic factors predispose certain genomic locations, including ZNF genes, to host proviruses that constitute the latent reservoir.

32 **Introduction**

33 During antiretroviral therapy (ART) of HIV-1 infection, a reservoir of $CD4^+$ T-cells harboring
34 intact HIV-1 proviruses persists and comprises a major barrier to achieving cures. These
35 reservoir-harboring cells undergo clonal expansion, evade host immune responses, and proviruses
36 therein provide a source of viral rebound upon ART interruption[1-6]. Understanding the
37 mechanisms underlying the establishment and maintenance of the HIV-1 proviral reservoir may
38 provide an opportunity to purge the reservoir and remove the need for lifelong ART. However,
39 there are a number of obstacles to achieving a detailed understanding of proviral persistence. The
40 extreme rarity of cells harboring intact proviruses *in vivo,* typically ~one in a million $CD4^+$ T
41 cells, poses a significant challenge to the investigation of the mechanisms that enable their
42 persistence[7,8]. Moreover, the intact proviral DNA that is intrinsically capable of initiating viral
43 rebound, is partly obscured by the existence of a larger amount of integrated viral DNA that is
44 not intact, but becomes comparatively abundant due to the selective survival of T-cell clones that
45 harbor defective, non-expressed proviruses [9-12].

46

47 Several factors may govern the establishment and maintenance of the HIV-1 reservoir. A key
48 mechanism of persistence is proviral latency, which protects infected cells from viral
49 cytopathicity and immune clearance. Whether a provirus adopts a state of latency versus active
50 transcription can be impacted by the physiological state of the infected cell, which influences the
51 levels of trans-acting factors such as NF-κB and NFAT, that are activated during T-cell activation
52 and link cellular activation to HIV-1 transcription[13,14]. Additionally, immune evasion
53 mechanisms, such as impaired antigen presentation or resistance to cytotoxic T lymphocyte
54 (CTL) killing, may further promote the survival and expansion of reservoir cells during periods
55 of activation [15-19].

56

57 The site of proviral integration into host DNA may influence reservoir persistence in two key and
58 distinct ways. First, the genomic context of the integration site may affect the accessibility of
59 proviruses to the transcriptional machinery and thus impact the propensity of a provirus to
60 establish and maintain latency. Second, by acting as an insertional mutagen, HIV-1 integration in
61 or near genes controlling cell proliferation may drive clonal expansion [4,20-29]. HIV-1 favors
62 integration within actively transcribed genes, via the interactions of viral capsid and integrase
63 proteins with CPSF-6 and LEDGF/p75, host proteins that are associated with areas of the genome,
64 or individual genes, that are actively transcribing [30-35]. In HIV-1-infected expanded cell clones
65 from ART-treated individuals, viral integration sites are often found in a small number of genes
66 whose function is related to cell survival and proliferation, suggesting that clonal expansion might
67 be driven in part by viral integration. However, these clones typically harbor defective proviruses,
68 and the mechanism by which these insertions drive clonal expansion, and whether they contribute
69 to intact proviral reservoir persistence, remains unclear [5,27,36]. In contrast, recent studies have
70 shown that expanded clones of cells from infected individuals on long-term ART that carry intact
71 proviruses, do so predominantly with proviral integration sites in regions comprising gene deserts,
72 intergenic, centromeric satellite DNA, and also, inexplicably, within zinc finger (ZNF) genes
73 [23,29,37].
74
75 Studies of integrated HIV-1 proviruses in cell lines have suggested the importance of the genomic
76 context in latency and its reversal [38,39]. However, currently available *in vitro* models for studying
77 HIV-1 latency in primary $CD4^+$ T cells [40-44] do not allow the evaluation of the relative
78 contributions of integration site versus the intrinsic physiological properties of T cell clones
79 harboring proviruses. Long-term *ex vivo* cell culture models in which primary $CD4^+$ T cells from
80 ART-treated individuals have been used to model latency in an authentic cell context, but cannot
81 evaluate cell clone-specific effects [45,46]. Thus, there is an unmet need for laboratory methods

capable of identifying individual HIV-1-infected cell clones, tracking their long-term clonal expansion, and assessing the effects of proviral integration site on latency in a clone-specific manner.

In this study, we present a novel 'HIV-1 latency mouse' (HIVLM) model that overcomes some of the limitations of current *in vitro* and *ex vivo* models of HIV-1 latency. We use the model to determine the impact of proviral integration sites on clonal expansion and latency in HIV-1 infected human memory $CD4^+$ T cells. Combining the HIVLM model with a method for efficient detection of integration sites and quantification of clonal expansion (PRISM-seq) [47], we determined how host-specific genomic and epigenomic features associated with each distinct HIV-1 integration site affect clonal survival and proviral expression in the absence of immunological or anti-viral drug selection pressures. We find that clonal expansion of infected cells in the HIVLM model was linked to the T cell receptor (TCR) sequence of engrafted cells. We find no evidence that clonal expansion is driven by insertional mutagenesis associated with provirus integration into specific genes. In contrast, the proviral genomic location and epigenetic status proximal to the integration site are key factors distinguishing latent from active proviruses. Latent proviruses were found at elevated frequency in intergenic regions and associated with heterochromatin. In line with previous clinical studies, latent proviruses in the HIVLM model were also found at elevated frequency in ZNF genes. Notably, our analysis reveals that the increased number of latent proviruses in ZNF genes was associated with their abundance, location, low expression levels, and heightened repressive epigenetic state.

Results

Long-term *in vivo* survival of HIV-1 infected human memory $CD4^+$ T cells in the HIVLM model

107 To track HIV-1 latency, we developed a single-cycle HIV-1 reporter virus construct, termed
108 V1/SBP-GFP, that expresses both GFP, and a cell surface streptavidin binding peptide (SBP)
109 fused to low-affinity nerve growth factor receptor. These reporters, expressed in place of Nef
110 under the control of the authentic HIV-1 promoter and splicing signals, could be used to both
111 monitor HIV-1 gene expression and as an affinity tag for one-step positive or negative selection
112 of cells with active or latent proviruses, using streptavidin-conjugated magnetic beads.
113
114 The suitability of primary human $CD4^+$ T cells for the study of HIV-1 latency over long timespans
115 is limited by their relatively short lifespan in cell culture. Following activation and infection of
116 freshly isolated $CD4^+$ T cells, these cells can be maintained in culture for only a few weeks, with
117 a need for re-stimulation to propagate them further. To enable the investigation of HIV-1 latency
118 and persistence in primary human $CD4^+$ T cells over longer time periods, we adapted a previously
119 described participant-derived xenograft (PDX) model of HIV-1 infection[48]. In this model, NSG
120 mice are engrafted with the memory ($CD45RO^+$) subset of $CD4^+$ T cells, which avoids the severe
121 graft versus host disease that rapidly occurs in mice engrafted with total human $CD4^+$ T cells[48].
122 For mouse engraftment, memory $CD4^+$ T cells were enriched from leukapheresis-derived PBMCs
123 of HIV-1-negative donors via negative selection (Fig. S1a). Cells were activated for 48 hours by
124 CD3/CD28-stimulation, resulting in >80% in $CD69^+$ cells (Fig. S1b). Activated cells were
125 infected, and approximately $4x10^9$ cells were subjected to streptavidin-bead-based magnetic cell
126 sorting to isolate V1/SBP-GFP-expressing memory $CD4^+$ T cells. Infection was done at low
127 multiplicity so that after positive selection, each cell is presumed to carry a single transcriptionally
128 active provirus at a unique integration site (Fig.1a). The sorting procedure yielded a >90 % pure
129 cell population of V1/SBP-GFP-expressing cells (Fig. S1c). V1/SBP-GFP expression was
130 confirmed two days after sorting, and cells were cryopreserved in aliquots for later engraftment
131 into mice (Fig. S1d).

132

133 The transcriptionally active HIV-1 provirus-containing (SBP^+GFP^+) human memory $CD4^+$ T
134 cells were engrafted into NSG mice by tail vein injection (~5-7×10^6 cells per mouse), and
135 successful engraftment was confirmed by the subsequent appearance of human $CD45^+$ cells in
136 peripheral blood (Fig. 1b and S1e). Engrafted human $CD4^+$ T cells were detected in all mice, with
137 ~10-100 human $CD4^+$ T cells per mm^3 blood at 35 days after engraftment (Fig. S1f). At this time
138 point, ~40 to 65% of the human $CD45^+$ cells were SBP^+GFP^+, and a time-dependent loss of
139 SBP/GFP reporter expression ensued, (Fig. 1b and S1g). The loss of SBP/GFP expression in
140 human $CD4^+$ T cells was observed in multiple tissues, including spleen, lung, and bone marrow,
141 months after engraftment (Fig. S1h-j). To determine whether the loss of SBP/GFP expression
142 over time reflected the establishment of latency or the selective outgrowth of contaminating
143 uninfected cells, we determined the *psi-gag*, *env*, and human RPP30 DNA copy numbers in
144 genomic DNA extracted from mouse blood and tissues by digital droplet PCR (ddPCR) using an
145 intact proviral DNA assay (IPDA)[49]. Detection of approximately one *psi-gag* and *env* copy per
146 human cell demonstrates the long-term survival of HIV-1 infected memory $CD4^+$ T cells and
147 shows that most human cells in mouse spleens contained a reporter provirus (Fig. S1k), despite
148 the low number of GFP-expressing cells. The observed loss of V1/SBP-GFP expression is
149 therefore not caused by outgrowth of contaminating uninfected cells, but primarily by the
150 transition to latency by a subset of proviruses in infected cells, during the months after
151 engraftment in the HIVLM model.

152

153 **Clonal expansion of HIV-1 infected human cells and proviral integration site features in the**
154 **HIVLM model**

155 To determine the distribution of integration sites in engrafted HIV-1 infected human cells, we
156 employed a method, termed PRISM-seq, in which a DNA sample is divided into 96 replicates,

157 each containing an estimated 50 proviruses, as measured by ddPCR, prior to whole genome
158 amplification and integration site determination. PRISM-seq enables the identification of >80%
159 of the integration sites in a sample, and the fraction of the 96 replicates in which a given
160 integration site is found is a measure of the clonal expansion of the T cell containing an HIV-1
161 provirus with that integration site [47]. Using this approach, we simultaneously determined HIV-1
162 integration site identity and the extent of clonal expansion for each HIV-1-infected memory $CD4^+$
163 T cell harboring a given integration site, for the pre-engraftment cell population, and for cells
164 recovered from 8 mouse spleens >2 months after engraftment of cells from the same donor. Post-
165 engraftment spleen samples were either left unsorted or were sorted into active and latent cell
166 populations, based on SBP/GFP expression, before HIV-1 integration site determination (Fig.
167 S2a,b,c and S3a,b). We identified 5239 and 3152 unique integration sites in pre-engraftment and
168 post-engraftment cell populations, respectively (Fig. 1c, S2a,b and S3a). For 57% to 84% of
169 proviruses, we were able to capture both the 5′ and 3′ LTR-host junction during integration sites
170 determination (Fig. S2b and S3a), indicating that our overall integration site detection was
171 efficient, and suggesting (along with the SBP/GFP and IPDA measurements conducted on the
172 pre- and post-engraftment samples) that the majority of proviruses were intact.
173
174 Clonal expansion of the HIV-1-infected cells in the pre-engraftment sample was minimal, with
175 3.1% of the 5239 integration sites detected in more than one replicate. This result was
176 approximately as expected given that the cells were sampled 4 days after *in vitro* infection (Fig.
177 S2a). Conversely, at > 2 months after mouse engraftment, clonal expansion of HIV-1-infected
178 cells was extensive, with 8% to 60% of integration site detected multiple times in 96 replicates,
179 and with a few HIV-1 integration sites appearing in all 96 replicates of the PRISM-seq assay
180 (representing a frequency of > 1 in 50 cells and thus exceeding the dynamic range of our clonal
181 expansion measurement) (Fig. 1c, S2a and S3a). Consistent with previously published data, HIV-

182 1 integrations were biased toward genic rather than intergenic regions in the pre-engraftment
183 sample [50]. Moreover, there was no clear difference between pre- and post-engraftment samples
184 in genic versus intergenic distribution of proviral integration sites (Fig. 1d, S2c, and S3b), The
185 proportion of proviruses integrated in genic or intergenic regions in cells recovered from mice
186 was similar to that reported for ART-treated individuals[11,23,51] (Fig. 1d). Most genic integration
187 sites were in introns, as expected (Fig. 1e) and approximately equivalent numbers of proviruses
188 were integrated in the same versus the opposite orientation with respect to transcription of the
189 gene in which the integration site was found (Fig. 1f). Overall, these results suggest that the
190 genomic features associated with HIV-1 integration sites obtained pre-engraftment, after *in vitro*
191 infection of primary $CD4^+$ T-cells, and also after 2 months of engraftment in the HIVLM model
192 were consistent with patterns observed clinically in ART-treated individuals.
193
194 **Frequent integration of HIV-1 into specific host genes does not predict clonal expansion**
195 Because HIV-1 has the potential to act as an insertional mutagen that might drive cell proliferation
196 in humans or the mouse grafts, we next examined whether HIV-1 integration into specific host
197 genes predicted the extent of expansion of the respective infected human cell clone during the ~
198 2 month engraftment period. We identified numerous genes, such as NPLOC4, PACS1,
199 NFATC3, NOSIP, NEAT1, ASH1L, and KDM2A, in which multiple distinct HIV-1 integrations
200 were found in multiple mice, following 2 months of engraftment (Fig. 2a). However, the
201 appearance of multiple integrated proviruses in a given gene across several engrafted cell
202 populations did not predict the degree of expansion of the respective cell clones containing that
203 integration site. A possible exception to this conclusion was NPLOC4, in which one out of 10
204 cell clones was highly expanded in one mouse (Fig. 2a). A better predictor of the frequency with
205 which proviral integrations were found in a given gene post engraftment, was the number of
206 integration sites found in that same gene in the pre-engraftment population (Fig. 2b). The genes

207 with increased frequency of integration were clustered near previously identified *in vitro* infection
208 defined integration 'hot spots'[52], for example on chromosome 11, 16, and 17 in both pre- and
209 post-engraftment cell populations (Fig. S4). As reported previously for *in vitro* HIV-1 infections,
210 increased gene density correlated with increased integration frequency in both pre- and post-
211 engraftment cell populations[53-55] (Fig. S4), The genes with frequent integration in our pre- and
212 post-engraftment datasets also overlap with the previously described "recurrent integration
213 genes" (RIGs)[52,56,57] (Fig. S5a and S5b).
214 Comparison of gene length and gene expression levels, determined based on previously published
215 RNA-seq data from activated central memory $CD4^+$ T cells[58], revealed that both characteristics
216 significantly correlated with an increased number of HIV-1 integrations in both pre- and post-
217 engraftment samples (Fig. 2c and 2d, respectively). Genes such as NPLOC4, PACS1, NFATC3,
218 NOSIP, NEAT1, ASH1L, and KDM2A may therefore be frequent targets of integration because
219 they are large genes and/or are highly expressed in memory $CD4^+$ T cells (Fig. 2e,f). CPSF6 and
220 LEDGF bias HIV-1 integration toward gene bodies and to nuclear speckles[32,34,54]. Using
221 published TSA-seq and DamID datasets for speckle-associated domains (SPAD) and lamina-
222 associated domains (LAD)[59], we also found enrichment of proviral integration sites in SPAD
223 (~75%) as compared to LAD (Fig. S5c) in both pre- and post-engraftment populations. Indeed,
224 genes with >1 proviral integration were more likely to be found in SPADs compared to genes
225 without proviruses (Fig. S5d and S5e). In ART-treated individuals, proviral integration sites in
226 expanded clones have been reported near host genes associated with cell growth and proliferation,
227 suggesting a possible effect of insertional mutagenesis on proliferation and survival of provirus-
228 containing cell clones[4,5]. However, comparative ontology analysis of genes with HIV-1
229 integrations in pre- and post-engraftment cell populations showed a similar distribution of
230 integration sites in genes associated with various functions, including epigenetic regulation of
231 gene expression, lymphocyte differentiation, and mitotic cell cycle regulation (Fig. S5f).

232 Additionally, in pre- and post-engraftment cells, we observed no enrichment or depletion of
233 integration sites in essential versus non-essential genes in expanded clones as compared to non-
234 expanded clones, irrespective of whether they were in RIGs or not[60] (Fig. S5g). Overall, the
235 distribution of HIV-1 integration sites among human genes was similar for *in vitro* infected
236 memory $CD4^+$ T-cells (pre-engraftment) and among clonally expanded HIV-1 infected $CD4^+$ T
237 cells in mice (post-engraftment). Because the pattern of integration sites did not differ in these
238 two populations, these results suggest an absence of selection pressure for proliferation and
239 survival that is driven by HIV-1 integration into specific genes. Rather, specific genes were found
240 to harbor proviruses in engrafted cells, including those that were clonally expanded, because they
241 are more frequent targets for provirus integration, due to their size, expression level, and nuclear
242 location, not because of putative effects of proviral integration on $CD4^+$ T-cell survival or
243 proliferation.
244
245 **T cell receptor (TCR) association with clonal expansion**
246 Mature T cells harbor uniquely rearranged TCR genes, whose sequence can be used to track the
247 expansion of individual T cell clones. While it was not possible to link individual TCR clonotypes
248 to a given HIV-1 integration site in our analyses, we determined complementarity-determining
249 region 3 (CDR3) TCR sequences in human cells pre- and post-engraftment, to assess clonal
250 expansion (Fig. S6a,b). We found hundreds of unique TCR in each engrafted mouse (mean = 300,
251 range 111 to 638), similar to the predicted number of independent T-cell clones recovered after
252 engraftment inferred by HIV-1 integration site mapping (Fig. S6b and Fig. 1c). The number of
253 unique T cell clones recovered from individual mice at 2 months post-engraftment (mean = 300)
254 was lower than the number of T cell clones recovered from the pre-engraftment sample (n =
255 38,621), indicating a bottleneck for cell survival or selective clonal expansion following
256 engraftment (Fig. S6b). While each mouse harbored a distinct TCR repertoire, and most TCR

257 sequences were recovered from only one mouse, certain TCR variants were detected in 2 or more
258 mice post-engraftment for both HIV-1 infected and uninfected engraftments (Fig. 3a and 3b).
259 Indeed, certain T cell clones were expanded similarly across multiple mice, whether or not the
260 cells were HIV-1 infected prior to engraftment (Fig. 3c-g). For example, a T-cell clone harboring
261 the CDR3 sequence CASSYGGAGNYGYTF was similarly expanded in HIV-1 infected and
262 uninfected grafts (Fig. 3c -f). In that case, the clone-CASSYGGAGNYGYTF was associated with
263 an abundant clone in the pre-engraftment sample (Fig. 3g, S6c). However, apparently selective
264 growth of some T cell clones that were either less abundant or not detected in the pre-engraftment
265 sample was evident in one or more engrafted mice (Fig. 3g and S6c). Indeed, while there was no
266 correlation between TCR sequence abundance in the pre-engraftment population and it
267 abundance post-engraftment, the mean number of times a given TCR sequence was recovered
268 from individual mice was correlated with the number of mice from which that clonotype was
269 recovered (Fig 3c,d). These findings indicate that the extent of expansion of HIV-infected and
270 uninfected cell clones post-engraftment is independent of their abundance pre-engraftment and
271 predicted to some extent by their TCR sequence. Xenoreactivity thus likely plays a role in clone
272 selection, a notion that is supported by the frequent occurrence of graft vs host disease in human
273 T-cell xenograft models[61,62], where the memory $CD4^+$T-cells are not purified before engraftment.
274 While our data do not identify specific drivers for the differential expansion of human memory
275 $CD4^+$ T cell clones in the HIVLM model, we can nevertheless conclude that a diverse set of
276 memory $CD4^+$ T-cells are expanded post-engraftment, predicted in part by TCR sequence and
277 not by HIV-1 integration site. Thus, for the majority of integration sites, HIV-1 proviruses act as
278 passengers, and not as mutagenic drivers during clonal expansion of infected cells in our model,
279 and likely in the majority of cases in humans.
280
281 **Genomic locations of transcriptionally active and latent proviruses**

282 The genomic context in which an HIV-1 provirus resides in an infected cell *in vivo* could affect
283 its transcription status and its propensity to transition to latency[11,28]. Infected human cells
284 harvested from the HIVLM model >2 months post-engraftment were sorted into populations with
285 transcriptionally active (SBP^+GFP^+) or latent (SBP^-GFP^-) proviruses. The low-template, single-
286 molecule detection efficiency of PRISM-seq, combined with multiple displacement amplification
287 performed directly on cells isolated from sorted populations, enabled reliable and comprehensive
288 determination of HIV-1 integration sites in both transcriptionally active and latent provirus-
289 harboring cell populations (Fig. S7a and Fig. S7b). This analysis identified 1,347 integration sites
290 associated with transcriptionally active proviruses, 863 sites associated with latent proviruses,
291 and 140 sites associated with both transcriptionally active and latent proviruses (Fig. 4a and Fig.
292 S7c). Note that 2 and 12 integration sites with active and latent proviruses, respectively, were
293 found in more than one mouse and are merged for downstream analysis. Clonal expansion, that
294 is detection in multiple replicate wells was evident for 139 integration sites with only
295 transcriptionally active proviruses (mean clonal size = 7.2 wells, range = 2-93) and 90 integration
296 sites with only latent proviruses (mean clonal size = 2.4 wells, range = 2-10) (Fig 4a). In the
297 comparative analyses that follow, we considered either (i) all detected integration sites or (ii) only
298 the integration sites associated with expanded clones. Restricting the analysis to expanded clones
299 reduced the number of sites that were considered but allowed more confident assignment of active
300 or latent status, as each site was observed multiple times. Additionally, to increase the robustness
301 of our comparisons, we combined all the active and latent integration sites retrieved from four
302 individual mice (Fig. 4a).
303
304 Latent proviruses were found to be more frequently integrated in intergenic regions as compared
305 to active proviruses, or proviruses exhibiting both active and latent states (Fig. 4b). This
306 differential distribution of integration sites was especially evident when expanded clones were

307 analyzed separately from non-expanded clones (Fig. S7d and S7e). Notably, integration sites
308 harboring latent proviruses, or a mixture of active and latent proviruses, were more likely to be
309 found on chromosome 19 ($p = 1.7\times10^{-6}$) and 21 ($p = 0.049$) as compared to integration sites of
310 active proviruses (Fig. 4c). This conclusion held for all integration sites found in expanded and
311 non-expanded clones, but was particularly evident for integration sites found in the expanded
312 clones (Fig. S8a-b, S9 and S10).
313
314 To ascertain whether the expression levels of a gene was associated with the propensity of
315 proviruses integrated therein to establish latency, we analyzed integration site and host gene
316 expression data in two ways: First, we assigned each gene to a category, determined by whether
317 proviruses integrated therein were active, latent or a mixture of both active and latent, regardless
318 of the number of integration sites in that gene. Second, we assigned each individual genic
319 integration site to an active, latent, or both active and latent category, regardless of the viral
320 phenotype associated with other integration sites in that gene. In the first analysis, human genes
321 with only latent proviruses had lower expression levels than those with only active proviruses (p
322 $= 0.027$) (Fig. 4d and S11a-b). In the second analysis, integration sites with latent proviruses were
323 found in genes with lower expression than integration sites with active proviruses ($p = 0.0094$)
324 (Fig. S11c). The magnitude of these differences was modest but was increased when integration
325 sites in expanded clones were considered separately (Fig. S11d-e). Thus, integration into human
326 genes that have lower expression levels increases the probability of generating a latent provirus.
327 We did not find differences in gene length for human genes with only latent proviruses as
328 compared to active proviruses (Fig. 4e). Notably, most genes in which multiple independent
329 integrations occurred had integration sites with mixtures of both active and latent proviruses.
330 However, some genes exhibited a striking skew in proviral transcriptional activity for the multiple
331 proviruses integrated therein. For example, 7/8 integration sites in PACS1 had active proviruses

332 and one provirus was found in both active and latent populations (Fig. 4f). In contrast, all 8
333 integration sites in the NOSIP gene harbored only latent proviruses (Fig. 4f, S11f-g). For
334 comparison, the overall fraction of genes with integration sites harboring only active proviruses
335 was 64.3% (number of genes with only active proviral integration sites divided by the total
336 number of genes with integrated provirus). This data suggests that (for example) integration into
337 the NOSIP gene is more likely to generate a latent provirus than integration into the PACS1 gene
338 ($p = 2.3\times10^{-4}$). The distance to the nearest host gene transcription start site (TSS) has been
339 previously shown to be correlated with expression of the integrated provirus [37]. For proviruses
340 found within genes, we found that the distance to the TSS was smaller for latent than active
341 proviruses, irrespective of orientation relative to host transcription (Fig. S12a-c). Proviral
342 orientation relative to host gene transcription did not strongly influence whether proviruses were
343 active or latent (Fig. S12d-f). Integration of proviruses into SPAD versus LAD regions did not
344 lead to significant differences in their propensity to establish latency (Fig. S13a). Similarly, there
345 was no significant difference in the proportion of active and latent proviruses when integration
346 sites in distinct nuclear sub-compartments, defined using chromatin conformation capture (Hi-C)
347 sequencing data [63], were compared (Fig. S13b). However, mapping of integration sites with
348 respect to genomic repeat elements showed that latent proviruses were more frequently integrated
349 into centromeric satellite regions (3.3% of repeat sequence integration sites) as compared to active
350 proviruses (0.31% of repeat sequence integration sites) (Fig. 4g). This difference was also evident
351 when integration sites in expanded clones or non-expanded clones were considered separately. In
352 expanded clones, 13.6% of integration sites with latent proviruses and only 2.8% of integration
353 sites with active proviruses were in centromeric satellite regions (Fig. S13c-d). Overall, the
354 anatomical nuclear region did not detectably impact the propensity of HIV-1 proviruses to
355 establish latency. However, discernible differences in the local sequence context in which

356 transcriptionally active and latent proviruses were found show that the HIV-1 integration site
357 influences the probability that a given provirus will establish latency.
358
359 **Repressive epigenetic modifications are enriched in proximity to latent proviruses**
360 Epigenetic modifications in the vicinity of the integration site might constitute the underlying
361 mechanism by which local sequence context affects proviral transcription. After identifying HIV-
362 1 integration sites associated with active and latent proviruses, we examined the levels of histone
363 modifications proximal to each integration site (up to 100 kb upstream or downstream of the
364 integration site) using a previously published ChIP-seq dataset for primary $CD4^+$ T cells[64].
365 Consistent with the fact that most integration sites are in genes, HIV-1 proviruses recovered from
366 engrafted cells were preferentially found in the vicinity of genomic regions enriched for histone
367 modifications typically found within gene bodies (H3K36me3) and enhancers associated with
368 active transcription (H3K4me1 and to a lesser extent H3K27Ac and H3K4me3) (Fig. 5a and
369 S14a). The repressive histone mark H3K27me3 was relatively depleted in the vicinity of
370 integration sites, while another repressive histone mark H3K9me3 appeared unchanged in
371 proximity to integration sites (Fig. 5a). The level of CpG DNA methylation was reduced proximal
372 to integration sites overall but was marginally higher in the proximity of latent proviruses than
373 active proviruses (Fig. 5a). These findings are in agreement with previous studies[53,65], and the
374 notion that HIV-1 integration is favored in transcriptionally active regions of the genome.
375 However, there were not striking differences in the histone modifications associated with active
376 versus latent proviruses when histone modifications were considered individually and proviral
377 populations were considered as a whole.
378
379 Proviral integration sites are flanked by chromatin bearing multiple epigenetic modifications,
380 whose effects on proviral expression may be exerted combinatorially. To integrate these signals

381 and classify genomic regions into distinct chromatin states that account for combinatorial effects,
382 we applied the generative machine-learning model ChromHMM[66]. The model uses epigenomic
383 ChIP-seq datasets and enabled us to annotate the genome in $CD4^+$ T-cells into 15 distinct
384 chromatin states based on the signal intensity of 6 histone marks (Fig. 5b). These states can be
385 grouped into four functional categories typically associated with promoter, enhancer, transcribed
386 and repressed sequences. The occurrence of active and latent proviral integration sites within
387 these chromatin-state defined categories showed that latent proviruses were depleted relative to
388 active proviruses in genomic regions associated with host gene transcription marked by
389 H3K36me3 (states 3 and 4) (Fig. 5b-c). Conversely, latent proviruses were found to be enriched
390 within chromatin states associated with heterochromatin that is transcriptionally repressed,
391 marked by H3K27me3 and H3K9me3 (state 11-15) (Fig. 5b-c).
392
393 We next compared the epigenetic features surrounding the integration sites of the 10 most
394 expanded clones in the active, latent, and mixed (active and latent) categories. The integration
395 sites were chosen based purely based the number of times they were encountered in the sorted
396 categories; by restricting analyses to a population that was sampled most intensively, the
397 transcriptional phenotype associated with each integration site could be ascribed with the greatest
398 confidence. For the active, latent, and mixed (active and latent) categories, 9/10, 5/10, and 7/10
399 of the integration sites were in genes (Fig. 5d). Several of the latent and the mixed integration
400 sites in this population were associated with intergenic genomic regions with abundant
401 H3K9me3, but none of the 10 active proviruses demonstrated this property (Fig. 5d and S14b).
402 Other epigenetic modifications did not differ noticeably among the 10 most expanded clones three
403 categories.
404
405 **Properties of ZNF genes favor the occurrence of latent HIV-1 proviruses**

406 Genes encoding ZNF proteins have recently been reported to harbor intact proviruses in
407 individuals on long-term ART[23,29]. Importantly, ZNF genes are exceptionally abundant in the
408 human genome, representing 2.6% of all genes. HIV-1 integration *per se* was not biased to ZNF
409 genes and the large numbers of integration sites found therein (3.1% of all integrations) was
410 approximately as expected based on the number of ZNF genes in the human genome ($p = 0.15$)
411 (Fig. 5e). However, integration into a ZNF gene was more likely to lead to proviral latency than
412 was integration in other genes. Among proviruses integrated within genes, ZNF genes harbored
413 a higher fraction that were latent. Indeed, 5.2% of latent proviruses in genes but only 1.6% of
414 active proviruses in genes were integrated in ZNF genes (Fig. 5f).
415
416 Notably, the expression of ZNF genes in general was lower than that of typical non-ZNF genes
417 in $CD4^+$ T-cells (Fig. S15a). and lower that that of genes that were typically favored for
418 integration (Fig. S15 b-c). This conclusion held for ZNF genes that harbored active or latent
419 proviruses (Fig. S15b-c). There was no statistically significant difference in the expression levels
420 or the levels of promoter and gene body-associated histone marks for ZNF genes that had active
421 versus latent proviruses (Fig. S15d-e). However, ZNF genes with latent proviruses had higher
422 levels of integration site-proximal H3K9me3 compared to those of active proviruses (Fig. 5g).
423 The levels of methylated CpG DNA were higher proximal to latent proviruses with integration
424 sites in ZNF genes as compared to active proviruses (Fig. 5g) and this property was particularly
425 prominent for latent proviruses in ZNF genes that did not exhibit proximal H3K9me3 enrichment.
426 Overall, the majority of latent proviruses in ZNF genes were associated with proximally elevated
427 levels of either H3K9me3 or methylated CpG DNA (Fig. 5g). These data clearly suggest that the
428 type of epigenetic modifications found near the site of HIV-1 integration in the human genome
429 affects the transcriptional fate of the integrated provirus and drives proviral latency. ZNF genes

430 preferentially harbor latent proviruses due to their large numbers, comparatively low levels of
431 expression, and enhanced levels of repressive H3K9me3 and CpG epigenetic modifications.
432
433 **Discussion**
434 Here, we describe a model (HIVLM) to study the impact of HIV-1 integration site on proviral
435 latency. In this model, memory $CD4^+$ T-cells, each carrying a single transcriptionally active
436 provirus, are engrafted into mice. The integrated proviruses in this model are fully competent to
437 report HIV-1 gene expression but do not further replicate – no new viral particles or proviruses
438 are generated after engraftment. Non-essential genes that may be cytocidal were removed from
439 the viral genome, and the only viral gene products expressed are those that regulate gene
440 expression (Tat and Rev). Moreover, the mice into cannot mount an immune response. Therefore,
441 it is expected that little, if any, negative selective pressure that would favor transcriptional
442 silencing is applied to the proviral population. Rather, the proviral population is unencumbered
443 and can exert effects on the expansion of cell clones, and adopt transcriptionally active or latent
444 states, based purely on the physiology of the cell clone that carries it, and the site in the host
445 genome at which it is integrated.
446
447 Clonal expansion is an inherent phenomenon associated with memory T cell survival *in vivo*[67]
448 and provirus-containing cells clonally expand to varying degrees after engraftment. Notably
449 though, a central finding of our study was that the integration site in cells harvested from mouse
450 grafts mirrored the integration site profile during initial infection of $CD4^+$ T cells prior to
451 engraftment. Gene features that are key determinants of integration frequency *in vitro* including
452 size, expression level, and nuclear localization, are sufficient to explain the occurrence of
453 'recurrent integration genes' such as NPLOC4, PACS1, NOSIP, and NFATC3 in grafted cells,
454 which match those observed in human studies[11,23,51]. Conversely, TCR analysis revealed that

455 some common T cell clones are favored for survival and expansion both HIV-1-infected and
456 uninfected grafts. Overall, our results suggest that TCRs are drivers of T cell clonal expansion in
457 our model and potentially in humans, while the HIV-1 provirus acts as a passenger during cell
458 proliferation, in most cases.
459
460 The diversity of integration sites harboring latent or active proviruses sites in post-engraftment
461 cells indicates that many locations in the human genome can sustain proviruses in either an active
462 or a latent state. Nevertheless, there was a clear enrichment of transcriptionally latent proviruses
463 in genomic regions that are repressed prior to integration such as intergenic and centromeric
464 satellite regions. This observation is consistent with studies in which intact proviruses present in
465 reservoir cells ART-treated human were found to be enriched in transcriptionally repressed
466 regions in the genome[29,68,69]. The integration site profile in the natural reservoir after long-term
467 ART presumably reflects the impact of immune selection pressure on a non-replicating viral
468 population, favoring the survival of proviruses that are poorly expressed[29]. In the HIVLM model,
469 where immune selection pressure is absent, unselected engrafted cells exhibit an integration
470 profile indistinguishable from that of *in vitro* infection. Immune selection is mimicked
471 experimentally in the HIVLM model, by the artificially application of selection pressure using
472 flow cytometry or magnet-based sorting of harvested SBP^-/GFP^- versus SBP^+/GFP^+ cells. That
473 the impacts of immune selection over years in humans can apparently be mimicked in our HIVLM
474 model after a single reporter gene-based simple sort, 2 months post engraftment is notable, and
475 our approach enables many proviruses, with a great variety of integration sites to be evaluated.
476
477 A limitation of our study is that we did not directly measured epigenetic modifications at the
478 provirus promoter or in the linear proximity of provirus integration site in the infected cell clones,
479 but rather relied on the distribution of histone markers measured elsewhere in primary CD4 T^+

480 cells[64]. Nevertheless, our data identifies host epigenetic signatures near provirus integration sites
481 that predict latency. In some cases, integration associated epigenetic signatures (such as those
482 found within genes, including elevated H3K36Me3, H3K4Me1, and reduced H3K27Me3)
483 disfavor latency. Conversely, latency associated signatures include histone modifications, such
484 as H3K9Me3 as well as CpG methylation, that are generally found outside genes, and typically
485 encountered in heterochromatin.
486
487 The underlying drivers of the phenomenon, that latent intact proviruses are frequently found in
488 ZNF genes[23,28], is suggested by our data. ZNF genes are a particularly abundant gene class (3.6%
489 of all genes), and our data shows that integration in ZNF genes predisposed integrated proviruses
490 to acquire latency in the HIVLM model. Notably ZNF genes exhibit reduced expression in
491 CD4+T-cells than is typical of human genes of a whole. Moreover, genomic loci in the vicinity
492 of the subset of ZNF genes that were found to harbor latent proviruses in our experiments
493 exhibited signatures of an epigenetically repressed states. These states included with a high level
494 of H3K9me3 mark, or higher than typical levels of methylated CpG. We note that about 40% of
495 the >700 ZNF genes in the human genome reside on a single human chromosome (chromosome
496 19), and while chromosome 19 is exceptionally gene dense, ~20% of all genes on chromosome
497 19 are ZNF genes. Chromosome 19 is also relatively G+C rich and is enriched in CpG
498 dinucleotides and islands[70]. Chromosome 19 was also strikingly enriched in latent proviruses in
499 the HIV-LM model. We posit that ZNF genes represent an abundant gene class, residing in a are
500 of genome where a confluence of factors favoring HIV-1 integration such as high gene density
501 gene density and those favoring latency (low gene expression and a repressed chromatin state)
502 are balanced in such a way that the probability of generating latent proviruses approaches a
503 maximum.

504 Overall, the approaches described herein provide a model system in which epigenomic
505 mechanisms involved in latency can be studied and HIV-1 cure strategies employing approaches
506 to reverse latency in reservoir cells can be evaluated.
507
508 **Material and methods**
509
510 **Proviral construct V1/SBP-GFP**
511 A proviral plasmid pV1/SBP-GFP was generated based on a minimal viral genome (V1)[71],
512 engineered from HIV-1 NL43 harboring large deletions or inactivating mutations in *gag*, *pol*, *vif*,
513 *vpu*, and *env*, and in which *nef* was replaced by GFP. To construct the V1/SBP-GFP plasmid, an
514 DNA fragment encoding a streptavidin binding peptide appended to the extracellular domain of
515 the low affinity nerve growth factor receptor and a P2A site, flanked by NotI restriction sites was
516 synthesized, PCR-amplified, and cloned into the NotI-linearized pV1 plasmid.
517
518 **Cell lines**
519 293T cells (ATCC, #CRL-3216; RRID: CVCL_0063) and MT4-LTR-GFP cells[72] were cultured
520 at 37 °C and 5% CO2 in DMEM (293T) or RPMI (MT4-LTR-GFP) supplemented with 10 %
521 FCS and Gentamycin. All cell lines used in this study were monitored regularly to ensure the
522 absence of retroviral contamination. Cells were also monitored regularly for mycoplasma
523 contamination. Cells used here were contamination-free.
524
525 **Virus production and characterization**
526 To generate V1/SBP-P2A-GFP-derived viral stocks, $7.5x10^7$ 293T cells (ATCC, #CRL-3216;
527 RRID: CVCL_0063) were seeded per $875cm^2$ 5-layer flask the day before transfection. Cells were
528 transfected with 75 µg pV1/SBP-P2A-GFP, 75 µg pCRV1/NLGagPol, and 15 µg pVSV-G[73]

529 using polyethyleneimine. 12 h post-transfection, medium was replaced with fresh growth
530 medium. Virus-containing supernatants were harvested and filtered 48 h post-transfection. Virus
531 was pelleted by ultracentrifugation though a 20% sucrose cushion, resuspended in PBS with 10%
532 BSA to achieve 70-80-fold concentrated virus stocks, aliquoted, and stored at -80 °C. Infectious
533 titers were determined by titration on MT4-LTR-GFP cells[72] and assessment of %GFP-positive
534 cells by flow cytometry 48 hrs after infection.
535
536 **Isolation and activation of memory CD4$^+$ T cells**
537 A small-volume blood draw (50 ml whole blood) and Leukapheresis were performed at the
538 Rockefeller University Hospital from a healthy adult participant who provided written informed
539 consent before participation. Samples were processed within 4 hrs of collection. PBMCs were
540 isolated from 50 ml whole blood (test infection to determine optimal participant-specific MOI)
541 or 100 ml Leukopack (for preparation of cells for engraftment) by density gradient centrifugation
542 using lymphocyte separation medium (Corning). Memory CD4$^+$ T cells were isolated from
543 PBMCs using the EasySep Human Memory CD4$^+$ T Cell Enrichment Kit (StemCell, #19157). A
544 subset of cells was used for purity assessment of memory CD4$^+$ T cells (CD4$^+$CD45RA$^-$
545 CD45RO$^+$) by flow cytometry using the following fluorochrome-conjugated antibodies: anti-
546 human CD4 Antibody, clone OKT4 (BioLegend, #317408), anti-human CD45RA Antibody,
547 clone HI100 (BD, # 550855), and anti-human CD45RO Antibody, clone UCHL1 (BD, #555493).
548 Approximately 25 % of retrieved memory CD4$^+$ T cells from Leukopack were frozen at a density
549 of 1x10^7 cells/ml in 100% fetal bovine serum and stored at -150 °C. The remaining memory CD4$^+$
550 T cells were cultured in regular RPMI growth medium supplemented with IL-2 (RPMI
551 supplemented with 10% fetal bovine serum, gentamycin, 2 ng/ml (20 U/ml) human recombinant
552 IL-2 (PeproTech, #200-02)) and activated with 25 µl Dynabeads Human T-Activator CD3/CD28
553 (Thermo Scientific) per 1x10^6 cells. Dynabeads were removed after 48 hrs, and cells were

554 prepared for infection. Activation was assessed by flow cytometry using anti-human CD69
555 antibody (BD, #555533).
556
557 **Infection of memory CD4$^+$ T cells**
558 Following removal of CD3/CD28 T cell activator beads, 1.3x10^9 memory CD4$^+$ T cells were
559 infected using 2x10^7 IU (based on titration of MT4-LTR-GFP cells, equivalent to an approximate
560 participant-specific MOI of 0.3) V1/SBP-GFP virus in the presence of 8 µg/ml polybrene. For
561 this, the virus was diluted in 50 ml RPMI growth medium supplemented with IL-2, and polybrene
562 was added at approximately. 9x concentration. The virus/polybrene mixture was added to the
563 cells, which were adjusted for a final cell density of 3x10^6 cells/ml in fresh RPMI growth medium
564 supplemented with human recombinant IL-2. After 12 hrs, the cell density was decreased to 1x10^6
565 cells/ml using fresh RPMI growth medium supplemented with human recombinant IL-2.
566 Infection efficiency was assessed 48 hours after infection by flow cytometry (GFP).
567
568 **Antibody-free magnetic isolation of HIV-reporter-expressing memory CD4$^+$ T cells**
569 Cells were washed once with PBS and once with PBS + 2mM EDTA + 0.2 % BSA. Cells were
570 resuspended at a density of 1x10^7 cells/ml in pre-warmed DMEM + 2mM EDTA + 0.2 % BSA
571 and incubated with pre-washed Dynabeads Biotin Binder (Thermo Fisher Scientific, #11047) at
572 a GFP+ cell: bead ratio of 1:20. Following 30 min incubation shaking at 37 °C, unbound
573 (negative) cells were removed and bead-bound (positive) cells were washed twice on a magnet
574 using pre-warmed DMEM + 2mM EDTA + 0.2 % BSA. For removal of the beads from the
575 isolated positive cells, bead-bound cells were incubated in pre-warmed RPMI growth medium
576 supplemented with 2mM biotin and incubated for 15 min at 37 °C on a shaker. Beads were
577 removed using a magnet and once again incubated with biotin-containing medium. Supernatants
578 containing released cells were pooled, counted, and the cell density was adjusted to 5x10^5 cells/ml

579 in RPMI growth medium supplemented with IL-2. The efficiency of antibody-free magnetic cell
580 sorting was confirmed by analysis of input, negative, and positive cells using flow cytometry
581 (GFP).
582
583 **Preparation of cells for engraftment**
584 At 48 hours after antibody-free magnetic isolation of V1/SBP-GFP-expressing cells, the purity of
585 cells was confirmed by flow cytometry, and cells were frozen for future engraftment at a density
586 of $1x10^7$ cells/ml, 1 ml/tube in fetal bovine serum with 10% DMSO. Cells were stored at -150 °C
587 until engraftment.
588 The day before engraftment, V1/SBP-GFP-expressing memory $CD4^+$ T cells frozen in fetal
589 bovine serum were thawed and quickly resuspended in 7-fold excess volume of pre-warmed
590 RPMI growth medium and centrifuged at 300g for 10 min. Cells were adjusted to 1.6E6 cells/ml
591 in RPMI growth medium supplemented with 50 U/ml IL-2 and cultured for 14 hrs. Before
592 engraftment, cells were washed once with PBS and resuspended in HBS at 6E7 cells/ml. 100 µl
593 of this cell suspension (6E6 cells) was used per mouse for engraftment.
594
595 **Engraftment of NSG mice with V1/SBP-GFP-expressing human memory $CD4^+$ T cells**
596 Irradiated (50 rad) 6-week-old NSG mice (NOD.Cg-Prkdc^scid Il2rg^tm1Wjl/SzJ from The
597 Jackson Laboratory) were engrafted with $6x10^6$ V1/SBP-P2A-GFP-expressing memory $CD4^+$ T
598 cells resuspended in 100 µl HBS by tail vein injection. Mice were periodically bled to assess
599 human immune cell reconstitution and V1/SBP-GFP expression. Approximately 100 µl of
600 peripheral blood was collected in EDTA-coated tubes. Cell pellets were stained for flow
601 cytometric analysis. Mice were cared for by trained research animal technicians and received
602 daily wellness checks. Procedures were performed in accordance with protocols approved by the
603 Weill Cornell Medical College Institutional Animal and Use Committee and Rockefeller

604 University CBC facility (Protocol # 24016-H). At 63 days after engraftment, mice were
605 sacrificed, and blood, bone marrow, lung, and spleen were isolated. Cells were
606 immunophenotyped for human CD45, CD3, CD8, and CD4 expression, as well as V1/SBP-GFP
607 expression. Remaining harvested cells were frozen at -80 °C until DNA extraction.
608
609 **Flow cytometric analysis of human cells engrafted in NSG mice**
610 Pre-engraftment samples were analyzed for purity using a panel of the following human-specific
611 antibodies: FITC anti-CD4 (OKT4), PE anti-CD45RO (UCHL1), and APC anti-CD45RA
612 (HI100). T cell activation was analyzed using APC anti-CD69, and V1/SBP-GFP expression was
613 analyzed by detection of GFP. For engraftment samples, the mice's peripheral blood was
614 collected into EDTA-coated tubes from the mice by submandibular bleed every 2 weeks to assess
615 human T cell reconstitution. Blood was centrifuged at 2000g for 5 mins to separate plasma and
616 cell pellets. Cells were immediately stained for flow cytometric analysis. Samples were stained
617 with human-specific antibodies for the analysis of cell surface markers: CD4 BV421, CD8
618 BV605, CD3 BV711, CD45 PerCp-Cy5.5, CD45RO-PE, CD69 PE-eFluor 610, CCR7 PE-Cy-7,
619 CD27 APC, CD25 APC-eFluor780. Fluorescent counting calibration beads were added to each
620 sample to determine absolute cell counts. For post-engraftment spleen sample analysis and
621 sorting, a similar antibody panel was used in combination with Live/Dead Aqua (Invitrogen) dead
622 stain as a cell viability marker. Stained samples were run on Attune NxT cytometer analyzer
623 (Thermo Fisher Scientific), sorted on BD FACS Aria III (BD Biosciences), and data were
624 analyzed by FlowJo version 10 software (Treestar, Ashland, OR).
625
626 **Intact Proviral DNA Assay (IPDA)**
627 Genomic DNA was isolated from blood, bone marrow, lung, and spleen using the Zymo Quick-
628 DNA miniprep PLUS kit according to the manufacturer's instructions. Intact HIV-1

629 copies/million $CD4^+$ T cells or per ng DNA were determined by digital droplet PCR (ddPCR)
630 using the Intact Proviral DNA Assay (IPDA)[49], where HIV-1 and human RPP30 reactions were
631 performed independently in parallel, and copies were normalized to the amount of input DNA. In
632 each ddPCR reaction, an average of 15 ng of genomic DNA was combined with ddPCR supermix
633 for probes (no dUTP, BioRad), primers (final concentration 900 nM, Integrated DNA
634 technologies), probes (final concentration 250 nM, Thermo Fisher Scientific), and nuclease-free
635 water. Primer and probe sequences (5′–>3′) were as follows:
636 RPP30 Forward Primer- GATTTGGACCTGCGAGCG,
637 RPP30 Probe- VIC-CTGACCTGAAGGCTCT- MGBNFQ,
638 RPP30 Reverse Primer- GCGGCTGTCTCCACAAGT;
639 HIV-1 Ψ Forward Primer- CAGGACTCGGCTTGCTGAAG,
640 HIV-1 Ψ Probe- FAM- TTTTGGCGTACTCACCAGT- MGBNFQ,
641 HIV-1 Ψ Reverse Primer- GCACCCATCTCTCTCCTTCTAGC;
642 HIV-1 env Forward Primer- AGTGGTGCAGAGAGAAAAAAGAGC,
643 HIV-1 env Probe- VIC-CCTTGGGTTCTTGGGA- MGBNFQ,
644 HIV-1 env Reverse Primer- GTCTGGCCTGTACCGTCAGC.
645 Droplets were prepared using a QX200 Droplet Generator (BioRad) and cycled at 95 °C for 10
646 min; 45 cycles of (94 °C for 30 sec, 59 °C for 1 min) and 98 °C for 10 min. Droplets were analyzed
647 on a QX200 Droplet Reader (BioRad) using QuantaSoft software (BioRad). Four technical
648 replicates were run for each sample.
649
650 **Ligation-mediated PCR-based HIV-1 integration site analysis**
651 A detailed procedure for intergrtaion site determination using PRISM-seq will be published
652 elsewhere (manuscript submitted). Briefly, extracted DNA was diluted according to ddPCR
653 results, so that 50 provirus was present in each well of the 96-well plate. Subsequently, gDNA in

654 each well was subjected to multiple displacement amplification (MDA) with phi29 polymerase
655 (Qiagen REPLI-g Single Cell Kit #150345) at 30 °C for 4 h. DNA produced by whole-genome
656 amplification was used as a template. Viral-host junction regions were amplified using a nested
657 PCR strategy. 5LTR host junction was amplified with LSP1
658 (GCTTCAGCAAGCCGAGTCCTGCGTCGAG) and LSP2
659 (GCTCCTCTGGTTTCCCTTTCGCTTTCAA), and 3LTR host junction was amplified with
660 LestraLTR1 (CTTAAGCCTCAATAAAGCTTGCCTTGAG) and LestraLTR2
661 (AGACCCTTTTAGTCAGTGTGGAAAATC) as forward primers in the 1st and 2nd PCR
662 reaction in combination with reverse primers annealing with the ligated adaptor sequence, AP1
663 (GTAATACGACTCACTATAGGGC) and AP2 (ACTATAGGGCACGCGTGGT) respectively.
664 The resulting PCR products were subjected to next-generation sequencing using Illumina MiSeq.
665 MiSeq paired-end FASTQ files were demultiplexed; small reads (142 bp) were then aligned
666 sequentially to HIV-1 reference genomes HXB2 then to the human reference genome GRCh38.
667 Biocomputational identification of integration sites was performed using a house R-based
668 pipeline (BulkIntSiteR). Briefly, chimeric reads containing both human and HIV-1 sequences
669 were evaluated for mapping quality based on (i) HIV-1 coordinates mapping to the terminal
670 nucleotides of the viral genome, (ii) absolute counts of chimeric reads, and (iii) depth of
671 sequencing coverage in the host genome adjacent to the viral integration site. The final list of
672 integration sites and corresponding chromosomal annotations was obtained using Ensembl (v86,
673 www.ensembl.org), the UCSC Genome Browser (www.genome.ucsc.edu), and GENCODE (v39,
674 www.gencodegenes.org). Repetitive genomic sequences harboring HIV-1 integration sites were
675 identified using Repeat Masker (www.repeatmasker.org).
676
677 **TCR analysis of HIV-1 infected and uninfected cell populations**

678 Briefly, 1ug gDNA samples from mouse spleen samples of HIV-1 infected and uninfected
679 engrafts were outsourced to Adaptive Biotechnologies for TCR sequencing using the assay type
680 hsTCRBv4b. The data obtained was analyzed using the ImmunoSeq analyzer and an in-house
681 built-in pipeline. The clones with a read count of 1 were removed, and the CDR3 protein
682 sequences were taken to create a unique list of TCR in each individual sample. The log count per
683 million was used as a measure to indicate the degree of clonal expansion of each cell clone in an
684 individual sample. The Pearson correlation analysis was performed to compare clones that
685 expanded across the mice engrafts in HIV-1 infected and uninfected samples.
686
687 **Analysis of epigenetic modifications proximal to integrations sites.**
688 The HIV-1 integration site coordinates retrieved from cells harboring latent and active proviruses
689 were used to generate 4kb binned chromosomal regions spanning 100kb upstream and
690 downstream of the provirus. The ChiP-seq dataset was assessed from ENCODE database for
691 H3K4me1, H3K4me3, H3K27ac, H3K36me3, H3K27me3, and H3K9me3 histone marks for
692 primary $CD4^+$ T cells using the following criteria: organ, blood; cell, leukocyte; biosample,
693 GM12878 or CD4-positive, alpha-beta memory T cell; genome assembly, GRCh38; file type,
694 "bigwig"; and output type, "fold change over control." Only datasets based on two replicates are
695 used. The histone modification signal for each epigenetic mark was represented as an average
696 read signal for 100 kb upstream and downstream of the proviral integration coordinate using the
697 UCSC bigWigAverageOverBed tool.
698
699 **Quantification and statistical analysis**
700 Data are presented as pie charts, bar charts, scatter plots with individual values, or box and
701 Whisker plots (indicating the median, minimum, maximum, and interquartile ranges).
702 Differences were tested for statistical significance using Mann-Whitney U tests, Fisher's exact

703 tests, chi-square tests, and two-sided Kruskal-Wallis nonparametric tests, as appropriate. $P < 0.05$
704 was considered significant; false discovery rate (FDR) correction was performed using the
705 Benjamini-Hochberg method. Statistical details can be found in the corresponding figure legends.
706 Analyses were performed using GraphPad Prism and R software.
707

708 **Acknowledgements**

709 We thank the Rockefeller University flow cytometry core, Comparative Biosciences Center
710 (CBC), and Genomic Resource Center. We are grateful to Marina Caskey (The Rockefeller
711 University) for organizing Leukapheresis at The Rockefeller University Hospital and Emily
712 Mastrocola and Dennis Copertino for technical assistance. This work was supported by NIH
713 grants to the Research Enterprise to Advance a Cure for HIV (REACH) 5UM1AI164565 and the
714 Center for Structural Biology of HIV RNA U54 AI170660. This article is subject to HHMI's
715 Open Access to Publications policy. HHMI lab heads have previously granted a nonexclusive CC
716 BY 4.0 license to the public and a sublicensable license to HHMI in their research articles.
717 Pursuant to those licenses, the author-accepted manuscript of this article can be made freely
718 available under a CC BY 4.0 license immediately upon publication.
719

720 **References**
721

722 1 Finzi, D. *et al.* Latent infection of CD4+ T cells provides a mechanism for lifelong
723 persistence of HIV-1, even in patients on effective combination therapy. *Nature*
724 *Medicine* **5**, 512–517 (1999). https://doi.org/10.1038/8394
725 2 Wong, J. K. *et al.* Recovery of replication-competent HIV despite prolonged
726 suppression of plasma viremia. *Science* **278**, 1291–1295 (1997).
727 https://doi.org/10.1126/science.278.5341.1291
728 3 Chun, T. W. *et al.* Presence of an inducible HIV-1 latent reservoir during highly active
729 antiretroviral therapy. *Proc Natl Acad Sci U S A* **94**, 13193–13197 (1997).
730 https://doi.org/10.1073/pnas.94.24.13193
731 4 Maldarelli, F. *et al.* Specific HIV integration sites are linked to clonal expansion and
732 persistence of infected cells. *Science* **345**, 179–183 (2014).
733 https://doi.org/doi:10.1126/science.1254194

5 Wagner, T. A. *et al.* HIV latency. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. *Science* **345**, 570–573 (2014). https://doi.org/10.1126/science.1256304

6 Cohn, Lillian B. *et al.* HIV-1 Integration Landscape during Latent and Active Infection. *Cell* **160**, 420–432 (2015). https://doi.org/https://doi.org/10.1016/j.cell.2015.01.020

7 Hosmane, N. N. *et al.* Proliferation of latently infected CD4(+) T cells carrying replication-competent HIV-1: Potential role in latent reservoir dynamics. *J Exp Med* **214**, 959–972 (2017). https://doi.org/10.1084/jem.20170193

8 Siliciano, J. D. *et al.* Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4+ T cells. *Nature Medicine* **9**, 727–728 (2003). https://doi.org/10.1038/nm880

9 Ho, Y.-C. *et al.* Replication-Competent Noninduced Proviruses in the Latent Reservoir Increase Barrier to HIV-1 Cure. *Cell* **155**, 540–551 (2013). https://doi.org/10.1016/j.cell.2013.09.020

10 White, J. A. *et al.* Clonally expanded HIV-1 proviruses with 5'-leader defects can give rise to nonsuppressible residual viremia. *J Clin Invest* **133** (2023). https://doi.org/10.1172/jci165245

11 Einkauf, K. B. *et al.* Intact HIV-1 proviruses accumulate at distinct chromosomal positions during prolonged antiretroviral therapy. *J Clin Invest* **129**, 988–998 (2019). https://doi.org/10.1172/jci124291

12 Wei, Y. *et al.* Single-cell epigenetic, transcriptional, and protein profiling of latent and active HIV-1 reservoir revealed that IKZF3 promotes HIV-1 persistence. *Immunity* **56**, 2584–2601.e2587 (2023). https://doi.org/10.1016/j.immuni.2023.10.002

13 Kinoshita, S., Chen, B. K., Kaneshima, H. & Nolan, G. P. Host Control of HIV-1 Parasitism in T Cells by the Nuclear Factor of Activated T Cells. *Cell* **95**, 595–604 (1998). https://doi.org/10.1016/S0092-8674(00)81630-X

14 Hokello, J., Lakhikumar Sharma, A. & Tyagi, M. AP-1 and NF-κB synergize to transcriptionally activate latent HIV upon T-cell receptor activation. *FEBS Letters* **595**, 577–594 (2021). https://doi.org/https://doi.org/10.1002/1873-3468.14033

15 Klatt, N. R., Chomont, N., Douek, D. C. & Deeks, S. G. Immune activation and HIV persistence: implications for curative approaches to HIV infection. *Immunological Reviews* **254**, 326–342 (2013). https://doi.org/https://doi.org/10.1111/imr.12065

16 Gramatica, A. *et al.* EZH2 inhibition mitigates HIV immune evasion, reduces reservoir formation, and promotes skewing of CD8⁺ T cells toward less-exhausted phenotypes. *Cell Reports* **44** (2025). https://doi.org/10.1016/j.celrep.2025.115652

17 Huang, S. H. *et al.* Latent HIV reservoirs exhibit inherent resistance to elimination by CD8+ T cells. *J Clin Invest* **128**, 876–889 (2018). https://doi.org/10.1172/jci97555

18 Ren, Y. *et al.* BCL-2 antagonism sensitizes cytotoxic T cell-resistant HIV reservoirs to elimination ex vivo. *J Clin Invest* **130**, 2542–2559 (2020). https://doi.org/10.1172/jci132374

19 Sun, W. *et al.* Phenotypic signatures of immune selection in HIV-1 reservoir cells. *Nature* **614**, 309–317 (2023). https://doi.org/10.1038/s41586-022-05538-8

20 Hughes, Stephen H. & Coffin, John M. What Integration Sites Tell Us about HIV Persistence. *Cell Host & Microbe* **19**, 588–598 (2016). https://doi.org/10.1016/j.chom.2016.04.010

21 Zhang, T.-h. *et al.* Barcoded HIV-1 reveals viral persistence driven by clonal proliferation and distinct epigenetic patterns. *Nature Communications* **16**, 1641 (2025). https://doi.org/10.1038/s41467-025-56771-4

22 Nguyen, K. & Karn, J. The sounds of silencing: dynamic epigenetic control of HIV latency. *Current Opinion in HIV and AIDS* **19**, 102–109 (2024). https://doi.org/10.1097/coh.0000000000000850

23 Huang, A. S. *et al.* Integration features of intact latent HIV-1 in CD4+ T cell clones contribute to viral persistence. *Journal of Experimental Medicine* **218** (2021). https://doi.org/10.1084/jem.20211427

24 Wu, V. H. *et al.* Profound phenotypic and epigenetic heterogeneity of the HIV-1-infected CD4+ T cell reservoir. *Nature Immunology* **24**, 359–370 (2023). https://doi.org/10.1038/s41590-022-01371-3

25 Teixeira, A. R. *et al.* Transcription of HIV-1 at sites of intact latent provirus integration. *J Exp Med* **221** (2024). https://doi.org/10.1084/jem.20240391

26 Symons, J., Cameron, P. U. & Lewin, S. R. HIV integration sites and implications for maintenance of the reservoir. *Curr Opin HIV AIDS* **13**, 152–159 (2018). https://doi.org/10.1097/coh.0000000000000438

27 Maldarelli, F. *et al.* HIV latency. Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science* **345**, 179–183 (2014). https://doi.org/10.1126/science.1254194

28 Jiang, C. *et al.* Distinct viral reservoirs in individuals with spontaneous control of HIV-1. *Nature* **585**, 261–267 (2020). https://doi.org/10.1038/s41586-020-2651-8

29 Lian, X. *et al.* Progressive transformation of the HIV-1 reservoir cell profile over two decades of antiviral therapy. *Cell Host Microbe* **31**, 83–96.e85 (2023). https://doi.org/10.1016/j.chom.2022.12.002

30 Schröder, A. R. *et al.* HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110**, 521–529 (2002). https://doi.org/10.1016/s0092-8674(02)00864-4

31 Han, Y. *et al.* Resting CD4+ T cells from human immunodeficiency virus type 1 (HIV-1)-infected individuals carry integrated HIV-1 genomes within actively transcribed host genes. *J Virol* **78**, 6122–6133 (2004). https://doi.org/10.1128/jvi.78.12.6122-6133.2004

32 Engelman, A. N. & Singh, P. K. Cellular and molecular mechanisms of HIV-1 integration targeting. *Cell Mol Life Sci* **75**, 2491–2507 (2018). https://doi.org/10.1007/s00018-018-2772-5

33 Winans, S. *et al.* A point mutation in HIV-1 integrase redirects proviral integration into centromeric repeats. *Nature Communications* **13**, 1474 (2022). https://doi.org/10.1038/s41467-022-29097-8

34 Achuthan, V. *et al.* Capsid-CPSF6 Interaction Licenses Nuclear HIV-1 Trafficking to Sites of Viral DNA Integration. *Cell Host Microbe* **24**, 392–404.e398 (2018). https://doi.org/10.1016/j.chom.2018.08.002

35 Ciuffi, A. *et al.* A role for LEDGF/p75 in targeting HIV DNA integration. *Nature Medicine* **11**, 1287–1289 (2005). https://doi.org/10.1038/nm1329

36 Coffin, J. M. *et al.* Integration in oncogenes plays only a minor role in determining the in vivo distribution of HIV integration sites before or during suppressive antiretroviral therapy. *PLoS Pathog* **17**, e1009141 (2021). https://doi.org/10.1371/journal.ppat.1009141

37 Einkauf, K. B. *et al.* Parallel analysis of transcription, integration, and sequence of single HIV-1 proviruses. *Cell* **185**, 266–282.e215 (2022). https://doi.org/10.1016/j.cell.2021.12.011

38 Lewinski, M. K. *et al.* Genome-wide analysis of chromosomal features repressing human immunodeficiency virus transcription. *J Virol* **79**, 6610–6619 (2005). https://doi.org/10.1128/jvi.79.11.6610-6619.2005

39 Chen, H. C., Martinez, J. P., Zorita, E., Meyerhans, A. & Filion, G. J. Position effects influence HIV latency reversal. *Nat Struct Mol Biol* **24**, 47–54 (2017). https://doi.org/10.1038/nsmb.3328

40 Dobrowolski, C. *et al.* Entry of Polarized Effector Cells into Quiescence Forces HIV Latency. *mBio* **10**, 10.1128/mbio.00337–00319 (2019). https://doi.org/doi:10.1128/mbio.00337-19

41 Martins, L. J. *et al.* Modeling HIV-1 Latency in Primary T Cells Using a Replication-Competent Virus. *AIDS Res Hum Retroviruses* **32**, 187–193 (2016). https://doi.org/10.1089/aid.2015.0106

42 Bosque, A. & Planelles, V. Induction of HIV-1 latency and reactivation in primary memory CD4+ T cells. *Blood* **113**, 58–65 (2009). https://doi.org/10.1182/blood-2008-07-168393

43 Zhu, J., Yamane, H. & Paul, W. E. Differentiation of effector CD4 T cell populations (*). *Annu Rev Immunol* **28**, 445–489 (2010). https://doi.org/10.1146/annurev-immunol-030409-101212

44 Kaech, S. M., Wherry, E. J. & Ahmed, R. Effector and memory T-cell differentiation: implications for vaccine development. *Nature Reviews Immunology* **2**, 251–262 (2002). https://doi.org/10.1038/nri778

45 Kessing, C. F. *et al.* In Vivo Suppression of HIV Rebound by Didehydro-Cortistatin A, a “Block-and-Lock” Strategy for HIV-1 Treatment. *Cell Reports* **21**, 600–611 (2017). https://doi.org/10.1016/j.celrep.2017.09.080

46 Takata, H. *et al.* Modeling HIV-1 Latency Using Primary CD4(+) T Cells from Virally Suppressed HIV-1-Infected Individuals on Antiretroviral Therapy. *J Virol* **93** (2019). https://doi.org/10.1128/jvi.02248-18

47 Pal, V. K. *et al.* PRISM-Seq: An Ultra-sensitive Sequencing Approach For Mapping Lentiviral Integration Sites. *bioRxiv*, 2025.2012.2020.695659 (2025). https://doi.org/10.64898/2025.12.20.695659

48 McCann, C. D. *et al.* A participant-derived xenograft model of HIV enables long-term evaluation of autologous immunotherapies. *J Exp Med* **218** (2021). https://doi.org/10.1084/jem.20201908

49 Bruner, K. M. *et al.* A quantitative approach for measuring the reservoir of latent HIV-1 proviruses. *Nature* **566**, 120–125 (2019). https://doi.org/10.1038/s41586-019-0898-8

50 Schröder, A. R. W. *et al.* HIV-1 Integration in the Human Genome Favors Active Genes and Local Hotspots. *Cell* **110**, 521–529 (2002). https://doi.org/10.1016/S0092-8674(02)00864-4

51 Guo, S. *et al.* HIV infected CD4+ T cell clones are more stable than uninfected clones during long-term antiretroviral therapy. *PLoS Pathog* **18**, e1010726 (2022). https://doi.org/10.1371/journal.ppat.1010726

52 Marini, B. *et al.* Nuclear architecture dictates HIV-1 integration site selection. *Nature* **521**, 227–231 (2015). https://doi.org/10.1038/nature14226

53 Singh, P. K., Bedwell, G. J. & Engelman, A. N. Spatial and Genomic Correlates of HIV-1 Integration Site Targeting. *Cells* **11** (2022). https://doi.org/10.3390/cells11040655

54 Sowd, G. A. *et al.* A critical role for alternative polyadenylation factor CPSF6 in targeting HIV-1 integration to transcriptionally active chromatin. *Proc Natl Acad Sci U S A* **113**, E1054–1063 (2016). https://doi.org/10.1073/pnas.1524213113

55 Ocwieja, K. E. *et al.* HIV integration targeting: a pathway involving Transportin-3 and the nuclear pore protein RanBP2. *PLoS Pathog* **7**, e1001313 (2011). https://doi.org/10.1371/journal.ppat.1001313

56 Lucic, B. *et al.* Spatially clustered loci with multiple enhancers are frequent targets of HIV-1 integration. *Nature Communications* **10**, 4059 (2019). https://doi.org/10.1038/s41467-019-12046-3

57 Bedwell, G. J., Jang, S., Li, W., Singh, P. K. & Engelman, Alan N. rigrag: high-resolution mapping of genic targeting preferences during HIV-1 integration in vitro and in vivo. *Nucleic Acids Research* **49**, 7330–7346 (2021). https://doi.org/10.1093/nar/gkab514

58 White, C. H. *et al.* Transcriptomic Analysis Implicates the p53 Signaling Pathway in the Establishment of HIV-1 Latency in Central Memory CD4 T Cells in an In Vitro Model. *PLoS Pathog* **12**, e1006026 (2016). https://doi.org/10.1371/journal.ppat.1006026

59 Chen, Y. *et al.* Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler. *J Cell Biol* **217**, 4025–4048 (2018). https://doi.org/10.1083/jcb.201807108

60 Funk, L. *et al.* The phenotypic landscape of essential human genes. *Cell* **185**, 4634–4653.e4622 (2022). https://doi.org/https://doi.org/10.1016/j.cell.2022.10.017

61 Jeon, M. S. *et al.* Xenoreactivity of human clonal mesenchymal stem cells in a major histocompatibility complex-matched allogeneic graft-versus-host disease mouse model. *Cell Immunol* **261**, 57–63 (2010). https://doi.org/10.1016/j.cellimm.2009.11.001

62 Alhaj Hussen, K. *et al.* CD4(+)CD8(+) T-Lymphocytes in Xenogeneic and Human Graft-versus-Host Disease. *Front Immunol* **11**, 579776 (2020). https://doi.org/10.3389/fimmu.2020.579776

63 Xiong, K. & Ma, J. Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions. *Nature Communications* **10**, 5069 (2019). https://doi.org/10.1038/s41467-019-12954-4

64 Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015). https://doi.org/10.1038/nature14248

65 Sapp, N. *et al.* HIV-1 Preintegration Complex Preferentially Integrates the Viral DNA into Nucleosomes Containing Trimethylated Histone 3-Lysine 36 Modification and Flanking Linker DNA. *Journal of Virology* **96**, e01011–01022 (2022). https://doi.org/doi:10.1128/jvi.01011-22

66 Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nature Protocols* **12**, 2478–2492 (2017). https://doi.org/10.1038/nprot.2017.124

67 Soerens, A. G. *et al.* Functional T cells are capable of supernumerary cell division and longevity. *Nature* **614**, 762–766 (2023). https://doi.org/10.1038/s41586-022-05626-9

68 Armani-Tourret, M. *et al.* Selection of epigenetically privileged HIV-1 proviruses during treatment with panobinostat and interferon-α2a. *Cell* **187**, 1238–1254.e1214 (2024). https://doi.org/10.1016/j.cell.2024.01.037

69 Bittar, C. *et al.* HIV-1 expression is heterogeneous among clones of CD4+ T cells carrying authentic intact latent proviruses. *bioRxiv*, 2025.2008.2005.668730 (2025). https://doi.org/10.1101/2025.08.05.668730

70 Grimwood, J. *et al.* The DNA sequence and biology of human chromosome 19. *Nature* **428**, 529–535 (2004). https://doi.org/10.1038/nature02399

71 Zennou, V. & Bieniasz, P. D. Comparative analysis of the antiretroviral activity of APOBEC3G and APOBEC3F from primates. *Virology* **349**, 31–40 (2006). https://doi.org/10.1016/j.virol.2005.12.035

72 Kane, M. *et al.* Identification of Interferon-Stimulated Genes with Antiretroviral Activity. *Cell Host Microbe* **20**, 392–405 (2016). https://doi.org/10.1016/j.chom.2016.08.005

73 Yee, J. K., Friedmann, T. & Burns, J. C. Generation of high-titer pseudotyped retroviral vectors with very broad host range. *Methods Cell Biol* **43 Pt A**, 99–112 (1994). https://doi.org/10.1016/s0091-679x(08)60600-7

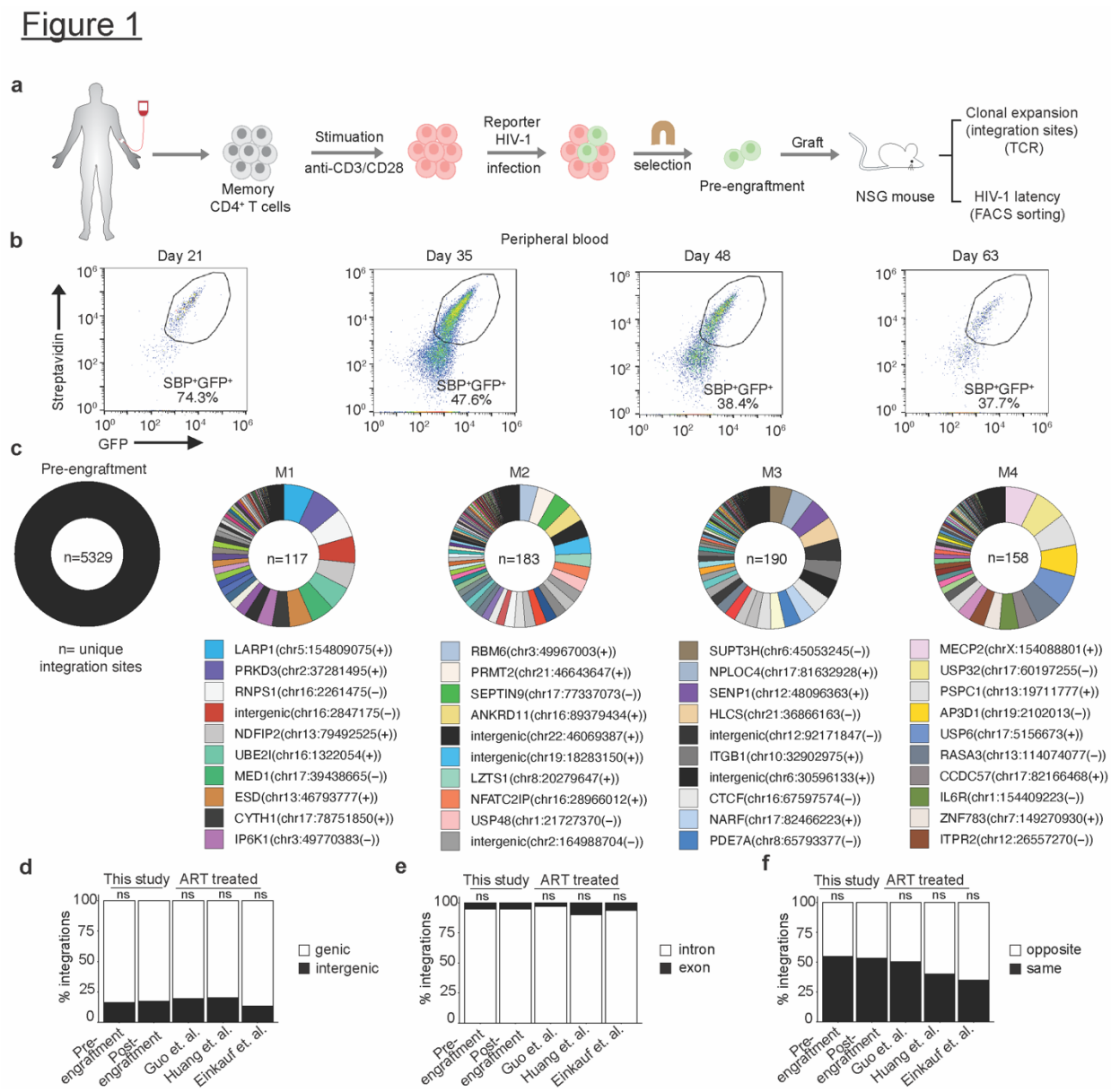


Figure 1. Tracking clonal expansion of HIV-1 infected human memory CD4$^+$ T cells *in vivo* in HIVLM model. **(a)** Isolation of human CD4$^+$ memory T cells, reporter HIV-1 (V1/SPB-GFP) infection, selection of infected cells, and engraftment in NSG mice. **(b)** Flow cytometry-based temporal analysis of GFP/SBP expression in engrafted human cells in the peripheral blood of a typical mouse. **(c)** Circos plots showing the relative expansion of HIV-1 infected cell clones in the pre- and post-engraftment (spleen) cell populations. Each clone is color-coded according to its proviral integration site, and the area of the pie represents the relative extent of clonal expansion, as measured by the number of wells in which that integration site was detected. Each integration site is labelled as the name of the gene with integrated provirus or as 'intergenic' if integration was outside of the gene, followed by the chromosome name, the genomic coordinate, and the host DNA strand (positive or negative). The n value represents the total number of unique integration sites identified in an individual pre- or post-engraftment (spleen) cell population. **(d)** Comparison of genic and intergenic distribution of unique integration sites in pre- or post-engraftment (spleen) cell populations, and a previously published clinical datasets[11,23,50]. **(e, f)** Comparison of the

proportion of genic integrations present in introns versus exons (**e**), and in same or opposite orientation as host transcription (**f**). The p-values for differences compared to post-engraftment (spleen) cell populations were determined by Fisher's exact test; ns = non-significant.

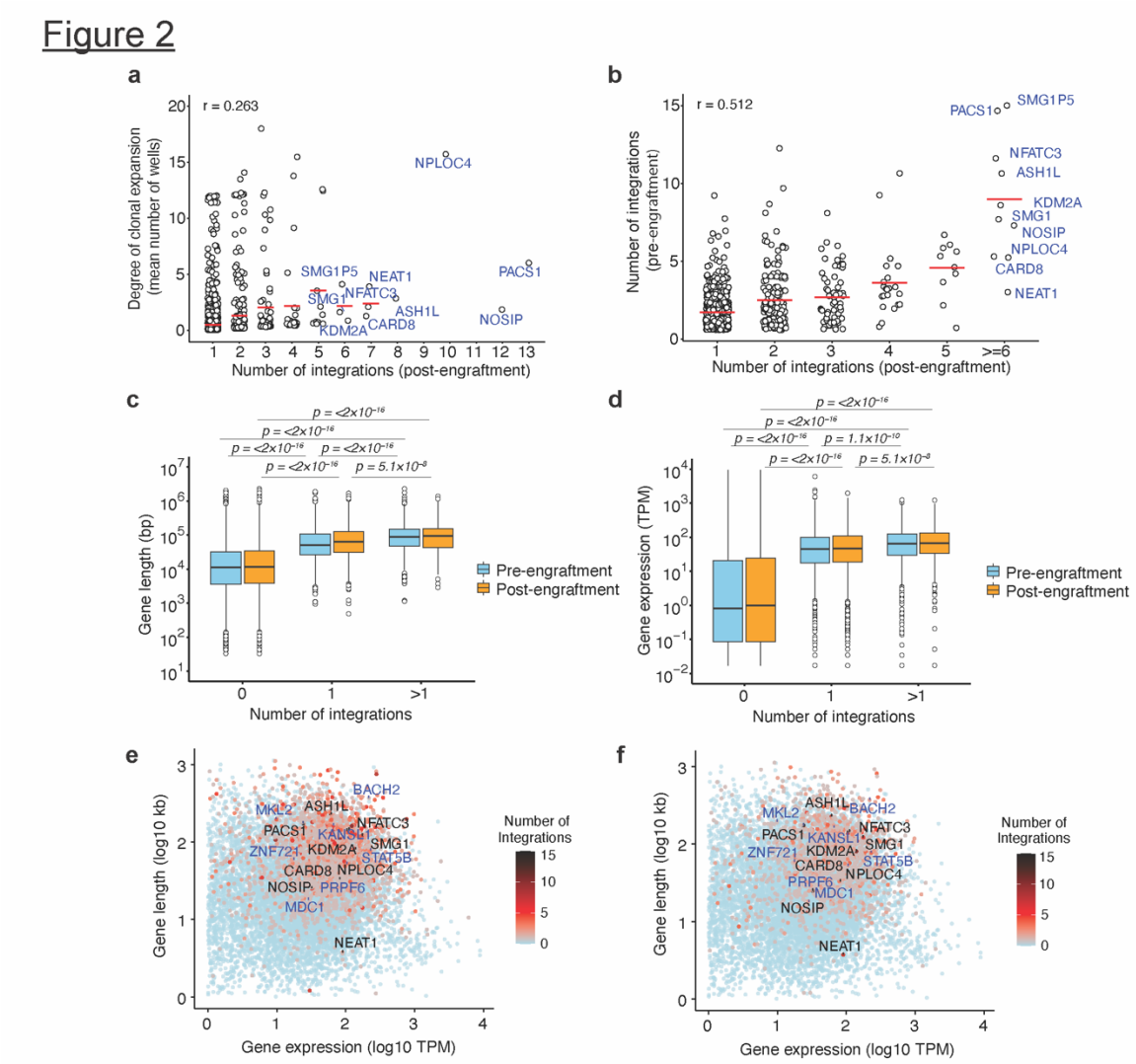


Figure 2. Identification of genes with frequent HIV-1 integration and their association with clonal expansion. **(a)** Extent of clonal expansion, as measured by the mean number of wells in which a unique integration site in a given gene was detected, plotted against the number of unique integration sites in a given gene detected post-engraftment. Each data point represents a gene with at least one integrant. **(b)** The number of unique integrations in a given gene in the pre-engraftment cell population plotted against the number of unique integration sites detected in that same gene 2 months post-engraftment. For **a** and **b**, data from four mice whose cells were not sorted is pooled. **(c, d)** Box plot of gene lengths (**c**) and level of expression (**d**) for genes with 0, 1, or >1 integrations. The plot shows median, interquartile ranges, and minimum/maximum values. P values were determined by pairwise Wilcoxon rank-sum tests with Bonferroni correction for multiple comparisons. **(e, f)** Gene length and expression level for genes without or with a varying number of integrated proviruses as indicated by the color scale in (e) pre- and (f) post-engraftment samples.

Figure 3

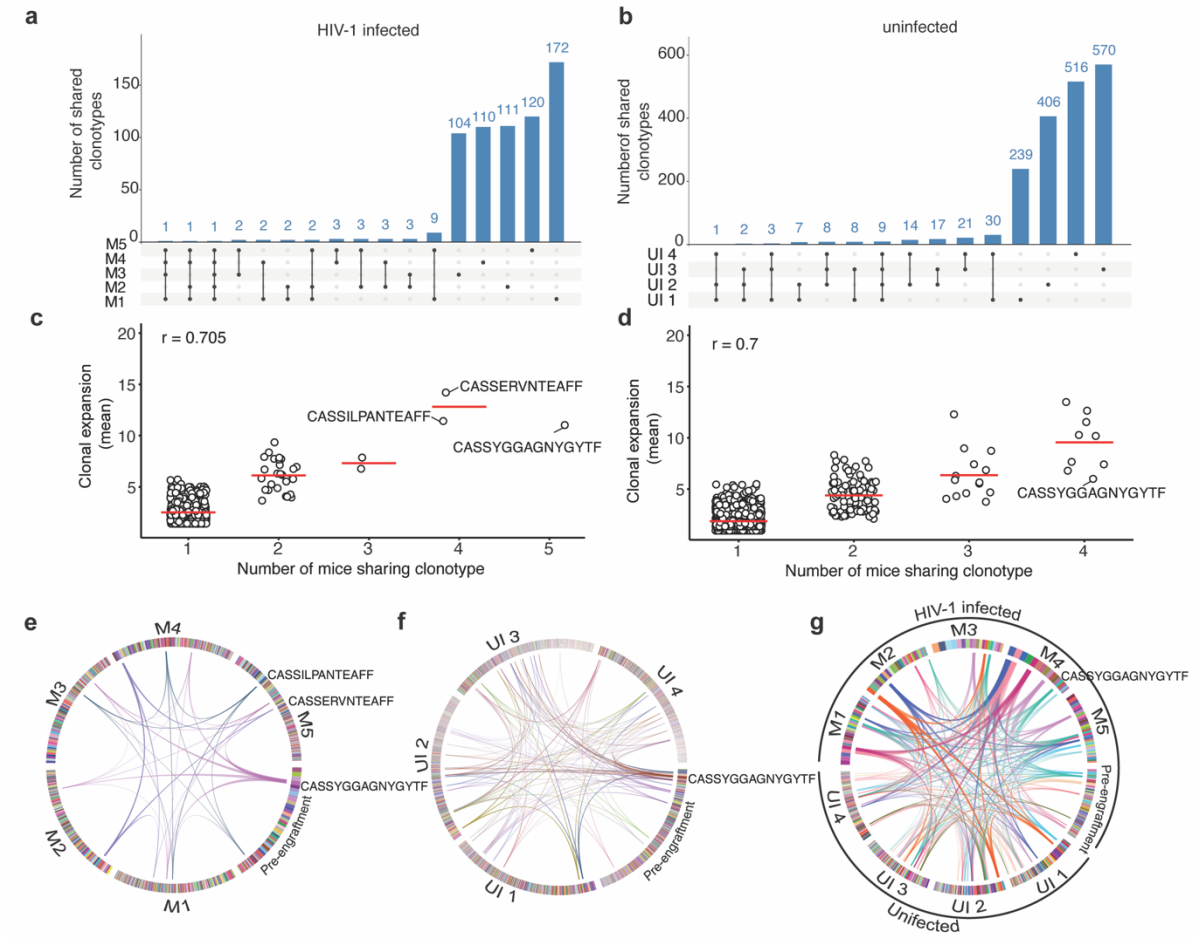


Figure 3. TCR analysis of HIV-1 infected and uninfected human memory CD4$^+$ T cells post-engraftment in mice. (a and b) Upset plots comparing unique T cell clones retrieved from multiple mice engrafted with HIV-1 infected (a) or uninfected (b) human cells from the same donor. **(c** and **d)** Correlation between the extent of clonal expansion and the number of mice from which individual T cell clonotypes were retrieved after engraftment of HIV-1 infected (c) or uninfected (d) human cells. The r value denotes the Pearson correlation coefficient value. **(e** and **f)** Circos plots showing the extent of clonal expansion of each T cell clone in pre-engraftment and post-engraftment in individual mouse samples. Arcs mark clones shared by a minimum of 4 mice for HIV-1 infected (e) and uninfected (f) grafts. Each T cell clone is color-coded and shown as boxes at the circumference of the circos plot for the respective sample. The T cell clones were arranged clockwise from most expanded to least expanded. **(g)** Circos plot showing T cell clones identified in pre-engraftment, and uninfected, HIV-1-infected post engraftment populations for a shared by a minimum of 4 mice.

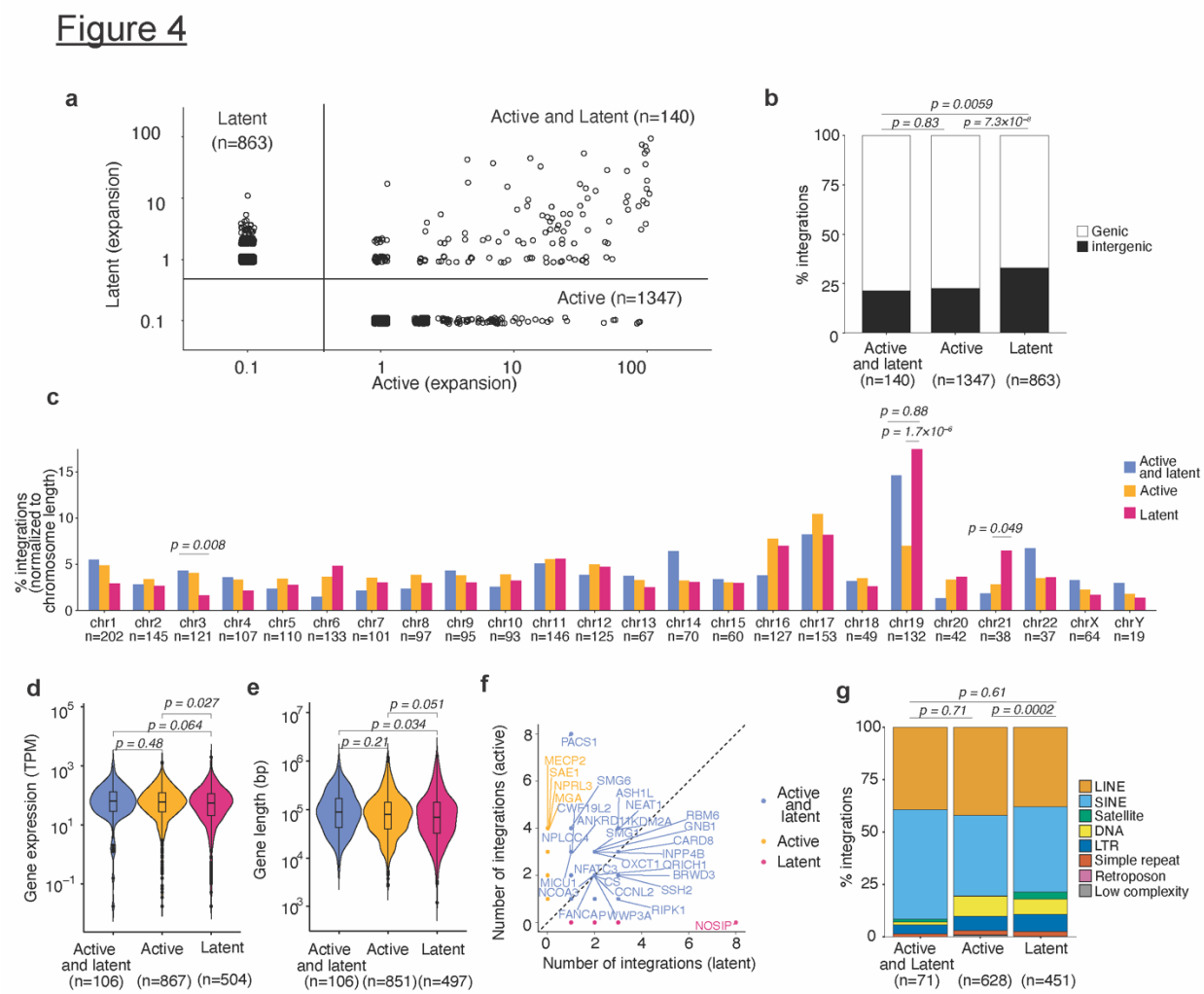


Figure 4. Genomic features associated with active and latent proviral integration sites. (a) Expansion of HIV-1-infected cell clones harboring only active, only latent, or a mixture of cells with active or latent proviruses. Axis values represent the number of wells in which an integrant was detected, integrant that were not detected are assigned a value of 0.1 for representation on a logarithmic scale. **(b)** Percentage of genic and intergenic integration site obtained in active, latent, or mixed active and latent cell clones. **(c)** Distribution of active, latent, or active and latent proviruses integration sites on human chromosomes. **(d)** Expression level of genes with only active proviruses, only latent proviruses, or both active and latent proviruses. **(e)** Length of genes with only active proviruses, only latent proviruses, or both active and latent proviruses. **(f)** Number of distinct integration sites in genes with only active proviruses, only latent proviruses, or both active and latent proviruses. The color of the dots and labels indicates if the proviruses in a given gene was found to be active, latent, or exhibited both phenotypes. **(g)** Percentage of integration sites in distinct chromosomal repeat regions in the human genome for the active, latent, or active and latent proviruses. P-values were calculated using Fisher's exact test (**b** and **c**), Wilcoxon rank-sum test with Bonferroni correction for multiple comparisons (**d** and **e**), and two proportions z-test to compare the proviruses present satellite repeat-rich region between the active and latent populations (**g**).

Figure 5

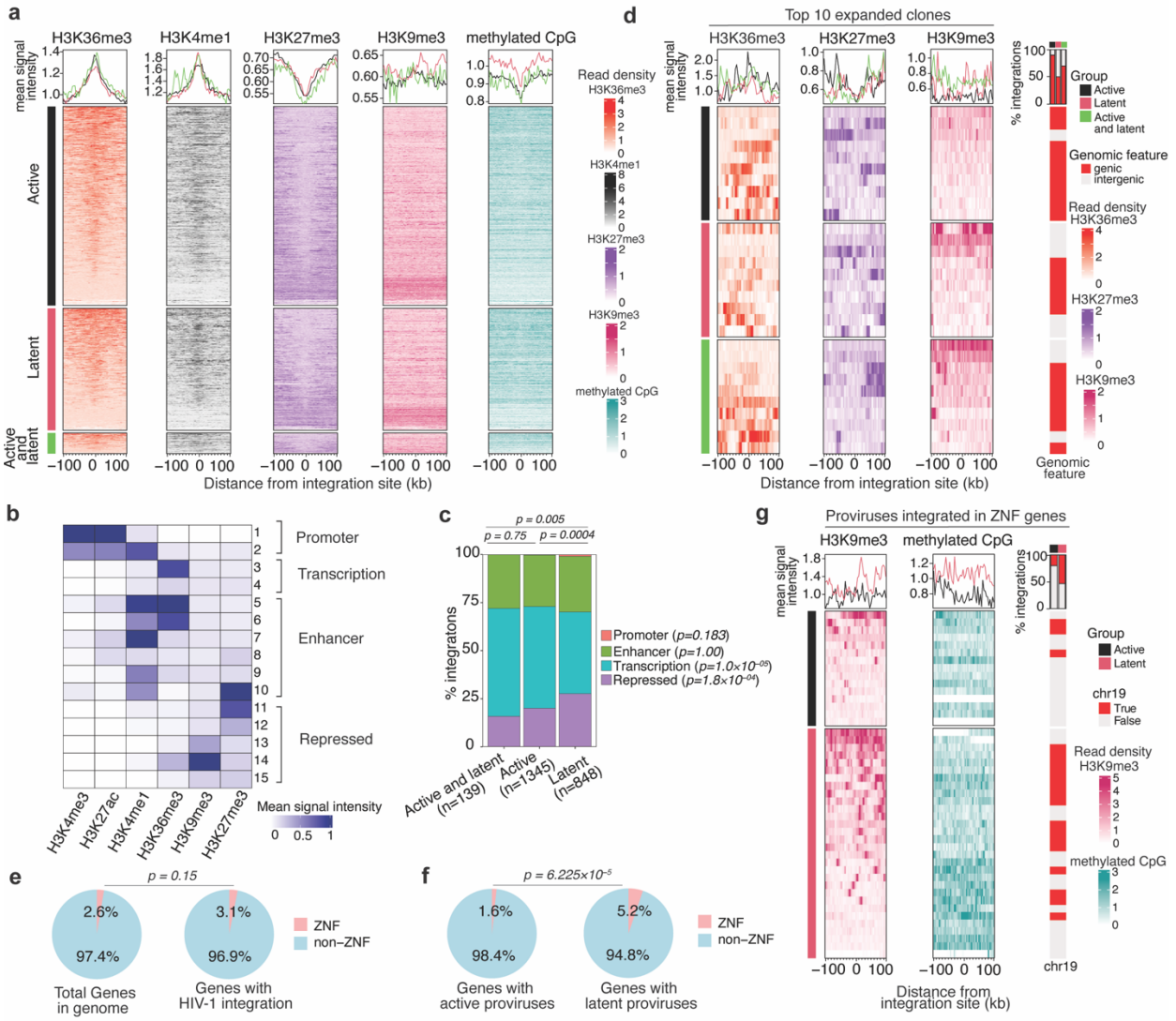


Figure 5. Distinct epigenetic signatures in the proximity of active and latent proviruses. (a) Heat map showing levels histone marks and methylated CpG found up to 100kb upstream and downstream of the integration sites of active, latent, or both (active and latent) proviruses. The line plot above the heat map shows the mean signal across all integration sites found within the active (black), latent (pink), and both (green) categories. The heatmaps for each group are arranged from top to bottom according to the highest to lowest mean signal intensity for H3K36me3 (orange) and this order is maintained for other histone marks, H3K4me1 (black), H3K27me3 (purple), H3K9me3 (pink), and methylated CpG (cyan), respectively. **(b)** Heatmap showing the mean signal intensity of various histone marks in the 15 chromatin states predicted by the ChromHMM model. **(c)** Stack bar plot showing the distribution of proviruses in distinct chromatin state groups. Global group-level statistical comparisons are shown on top of the plot using the chi-square test, and per-state level comparison between active and latent proviruses is given in in the legend using a two-proportion z-test. **(d)** The levels of H3K36me3 (orange), H3K27me3 (purple), and H3K9me3 (pink) histone marks in the proximity of the integration sites for active, latent, active and latent proviruses found in the top 10 expanded clones arranged in order arranged from top to bottom according to mean H3K9me3 signal.

Bars to the right depict whether each proviruses within a group was integrated into a genic or intergenic region. **(e)** Pie charts shows the percentage of ZNF genes in the genome and the fraction of ZNF and non ZNF genes that have HIV-1 integration. **(f)** Proportion of ZNF and non-ZNF genes harboring either active or latent proviruses. The p-values were determined by Fisher's exact test**. (g)** Histone modification levels proximal to active, latent, or active and latent proviruses found within ZNF genes.

Figure S1

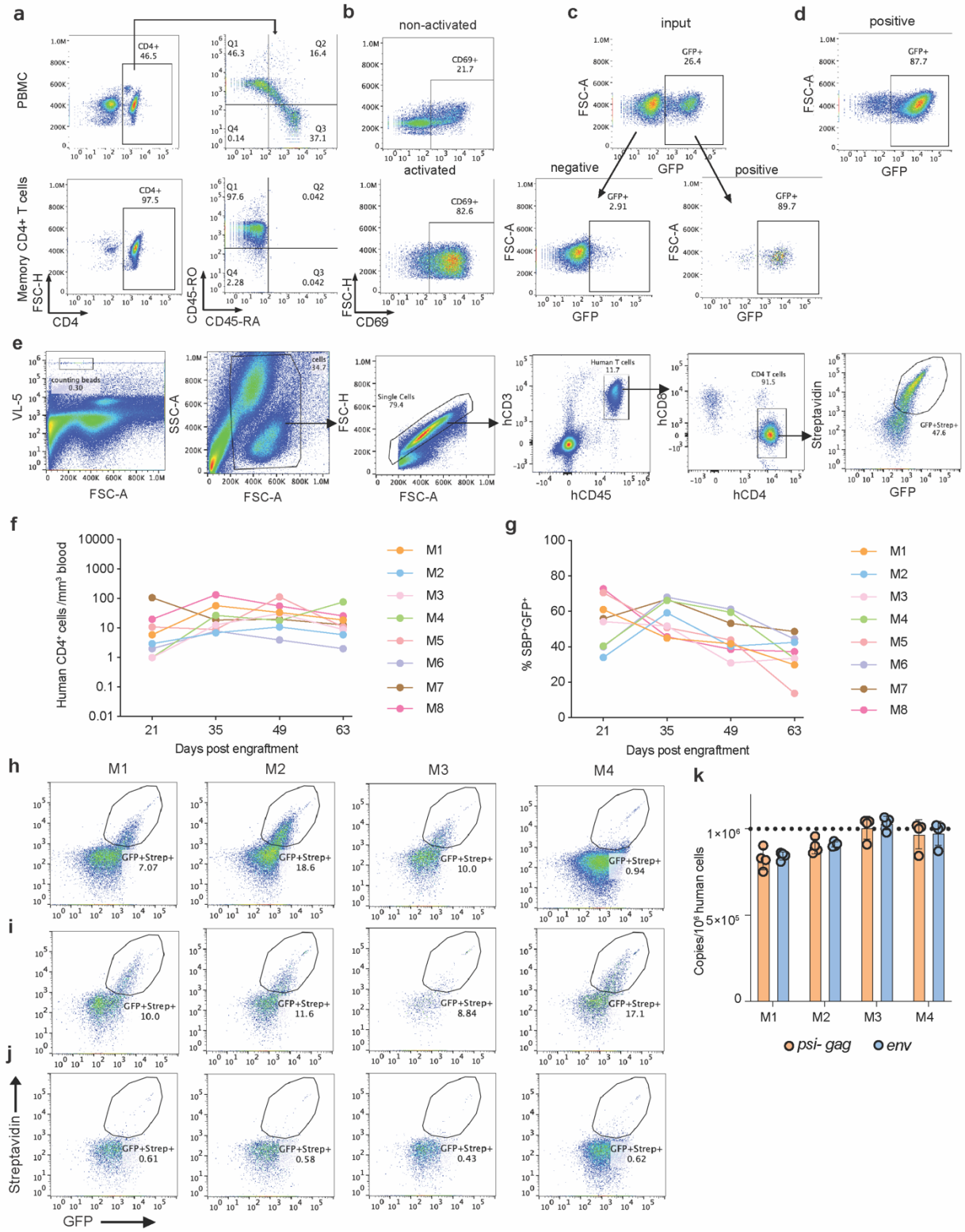


Figure S1. Infection, selection of reporter HIV-1-infected human memory CD4$^+$ T cells, and engraftment of NSG mice. **(a)** Flow cytometry plot showing the isolation of human memory CD4$^+$ T cells (lower panels)

from peripheral blood mononuclear cells (PBMCs, upper panels) of an HIV-1-uninfected human donor. **(b)** Frequency of activated human memory $CD4^+$ T cells ($CD69^+$) 4 days after activation with anti-CD3/CD28 beads and 2 days after infection with V1/SBP-GFP virus. **(c)** Streptavidin magnetic bead-based cell isolation of V1/SBP-GFP-expressing memory $CD4^+$ T cells. Bead-bound (positive) cells were separated from unbound (negative) cells 48 hours after infection with V1/SBP-GFP virus. **(d)** Percentage of V1/SBP-GFP positive cells 48 hours after sorting, immediately before freezing cells in aliquots for engraftment. **(e)** Flow cytometry gating strategy used for quantification of human $CD4^+$ T cell counts and HIV-1 expression post engraftment. Counting beads for calibration of the quantification were gated based on FSC-A and VL-5 (detector), an Attune NxT detector excited by a 405nm laser with an emission filter at 710/50nm. The figure reports the percentage of transcriptionally active HIV-1-infected (SBP^+GFP^+) cells in the $hCD3^+/hCD45^+/hCD4^+$ human cell compartment in a typical mouse peripheral blood sample. **(f, g)** Quantification of the numbers of human cells over ~2 months following mouse engraftment; total number of human cells (**f**) and percentage of transcriptionally active HIV-1 infected (SBP^+GFP^+) human cells (**g**) over time in multiple animals. **(h-j)** Percentage of transcriptionally active HIV-1 infected (SBP^+GFP^+) human cells (gated on $hCD3^+/hCD45^+/hCD4^+$ cells) in mouse spleen (**h**), lung (**i**), and bone marrow (**j**) of individual animals harvested 2 months after engraftment. **(k)** Quantification of *psi-gag* and *env* DNA copies per million human cells in spleen from each engrafted mouse.

Figure S2

a

Sample	Sample processing		Viral-host junction	Total integration sites	Unique integration sites	% clonal
Pre-infusion	unsorted		5' LTR	3789	5239	3.10%
			3' LTR	3887		
M1			5' LTR	1338	117	54.70%
			3' LTR	1258		
M2			5' LTR	1928	183	63.90%
			3' LTR	2066		
M3			5' LTR	1713	190	50%
			3' LTR	1797		
M4			5' LTR	1166	158	44.30%
			3' LTR	1087		
M5	sorted	Active	5' LTR	2191	328	23.17%
			3' LTR	2594		
		Latent	5' LTR	230	133	25.56%
			3' LTR	362		
M6		Active	5' LTR	971	85	35.30%
			3' LTR	1110		
		Latent	5' LTR	1022	46	41.30%
			3' LTR	1514		
M7		Active	5' LTR	971	235	28.51%
			3' LTR	1110		
		Latent	5' LTR	230	179	19.55%
			3' LTR	362		
M8		Active	5' LTR	2191	841	8.44%
			3' LTR	2594		
		Latent	5' LTR	1022	657	12.48%
			3' LTR	1514		

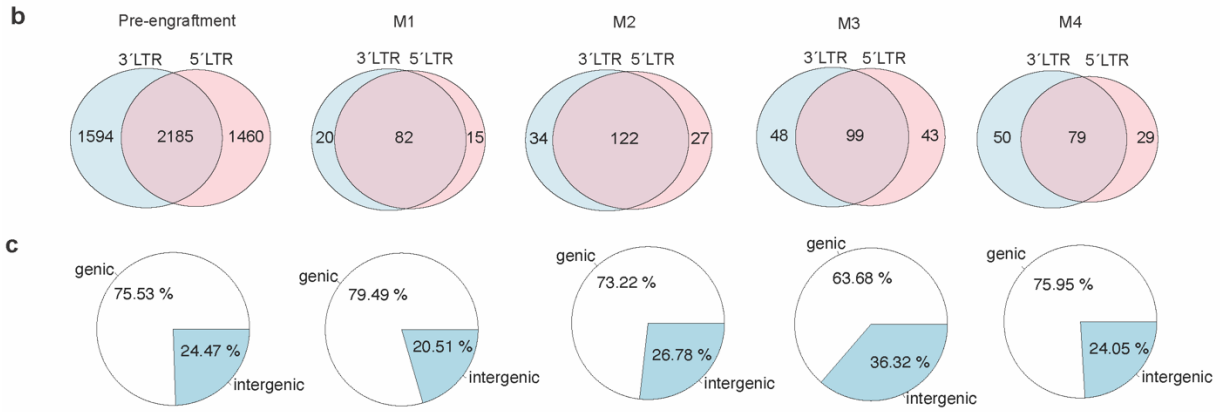


Figure S2. Identification of proviral integration sites in HIV-1 infected human memory $CD4^+$ T cell populations pre- and post-engraftment. **(a)** Table showing the number of unique integration sites retrieved from human cells harvested from the spleen of individual, engrafted mice. Cells recovered from mouse spleens for animals M1-M4 were not sorted, while human cells recovered from mice M5-M8 were sorted based on SBP/GFP expression. The 'percent clonal' column reports the proportion of the total number of unique integration sites found within a sample that were detected in more than one well out of the 96 wells tested. **(b)** Comparison and overlap in the numbers of integration sites retrieved by either 5´LTR or 3´LTR viral-host junction amplification strategies for pre- and post-engraftment cell populations. **(c)** Distribution of integrated proviruses in genes and intergenic regions in pre- engraftment cells and post-engraftment cells from individual mice.

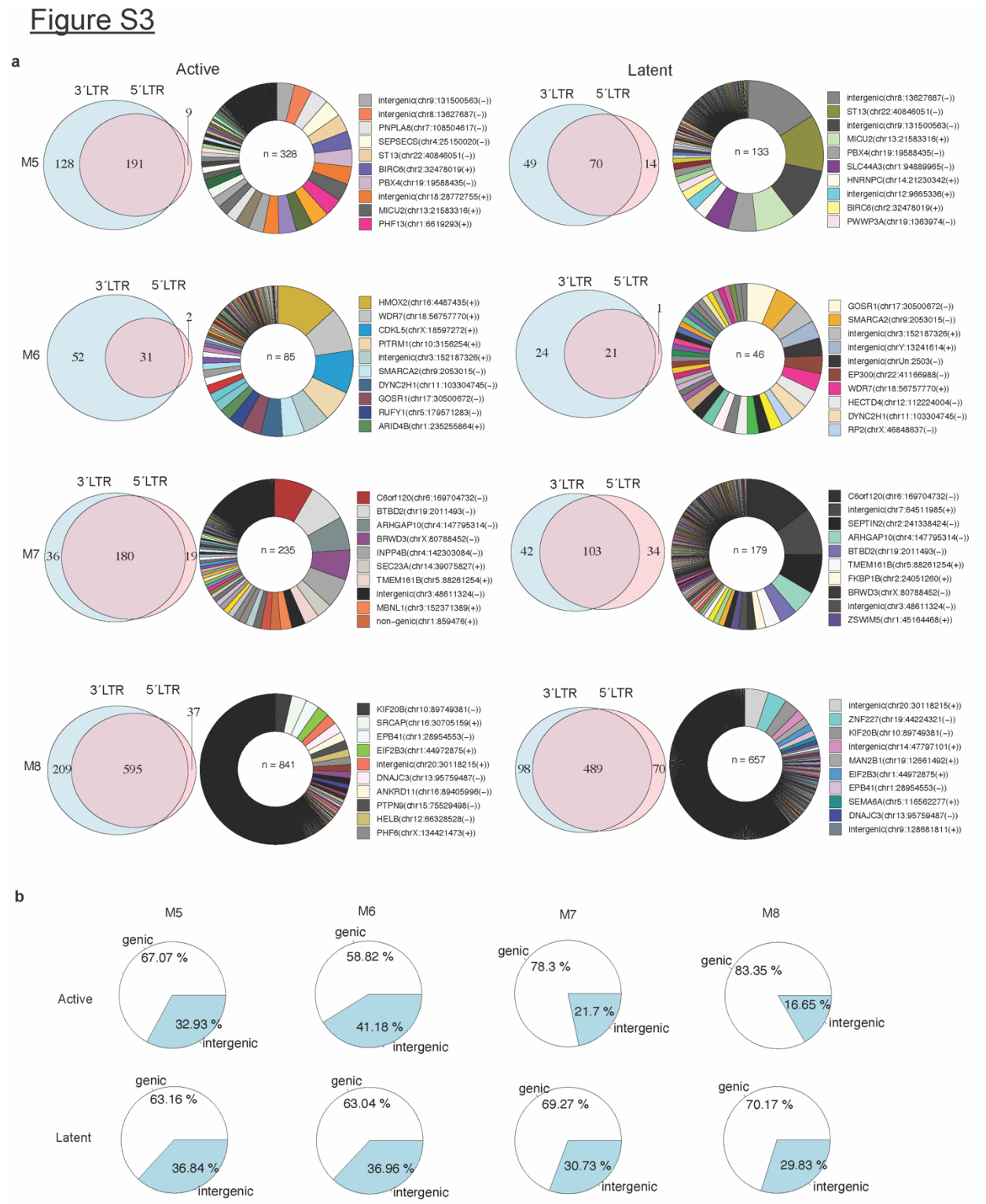


Figure S3. Integration sites identified in HIV-1 infected human memory CD4^{+} T-cells post-engraftment with active and latent proviruses. **(a)** Comparison of integration sites retrieved by 5´LTR and 3´LTR viral-host junction amplification. Circos plots showing the relative proportion of expansion of each clone in post-engraftment human cells from individual mouse spleens, sorted into populations with active and latent

proviruses. Each clone is color-coded according to its proviral integration site. Each integration site is labelled as the name of the gene with integrated provirus or as intergenic if integration was outside of the gene, followed by the chromosome name, the genomic coordinate, and the host DNA strand (positive or negative). **(b)** Distribution of integrated proviruses in genes and intergenic regions in pre- and post-engraftment samples from individual mice.

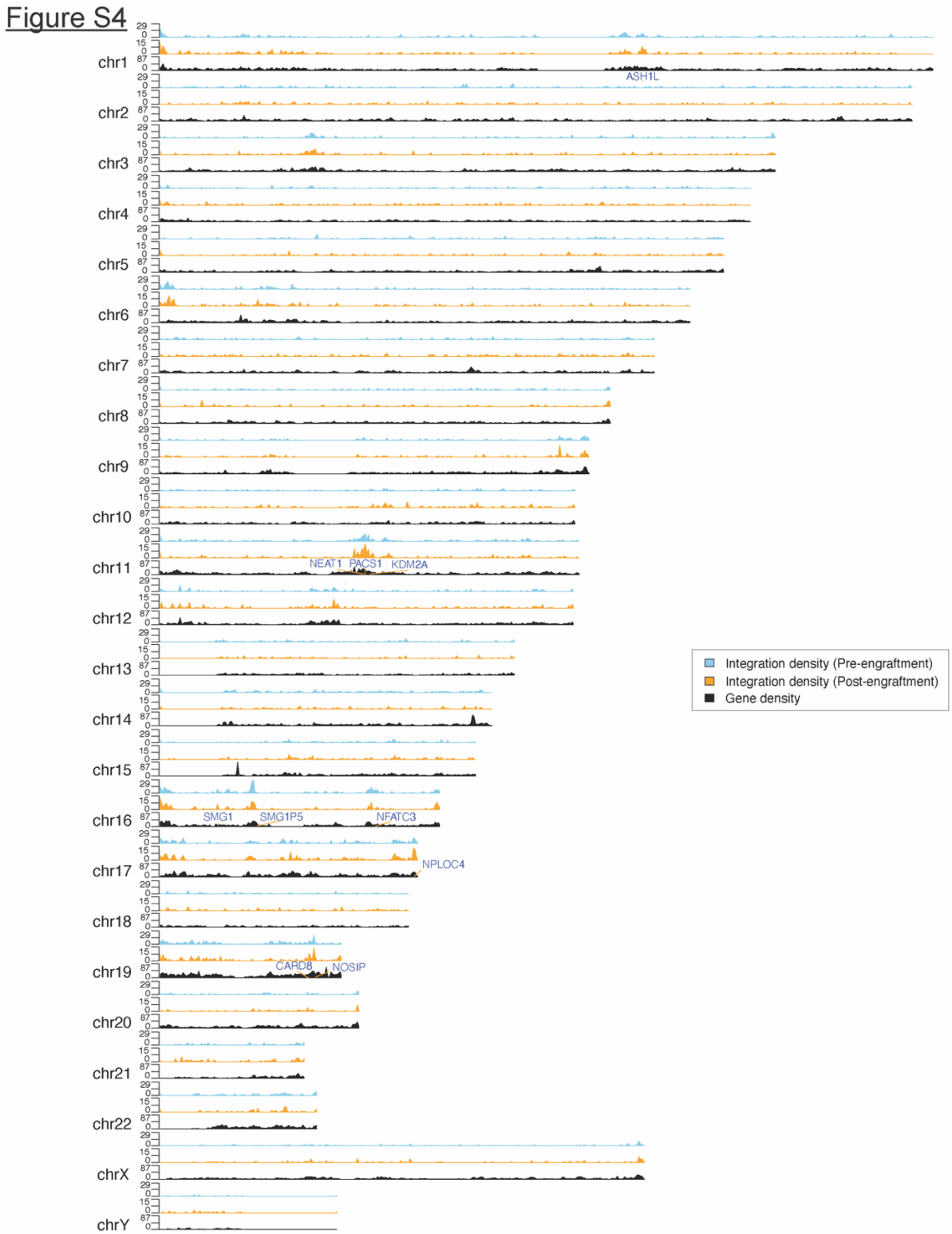


Figure S4. Gene density and integration frequency for human chromosomes and integration sites found in the pre- and post-engraftment cell populations. Genes with >5 unique integration sites post-engraftment are highlighted. The stacked y-axes represent the number of genes (0-87), the number of integrations in pre-(0-29), and post- (0-15) engraftment cell populations.

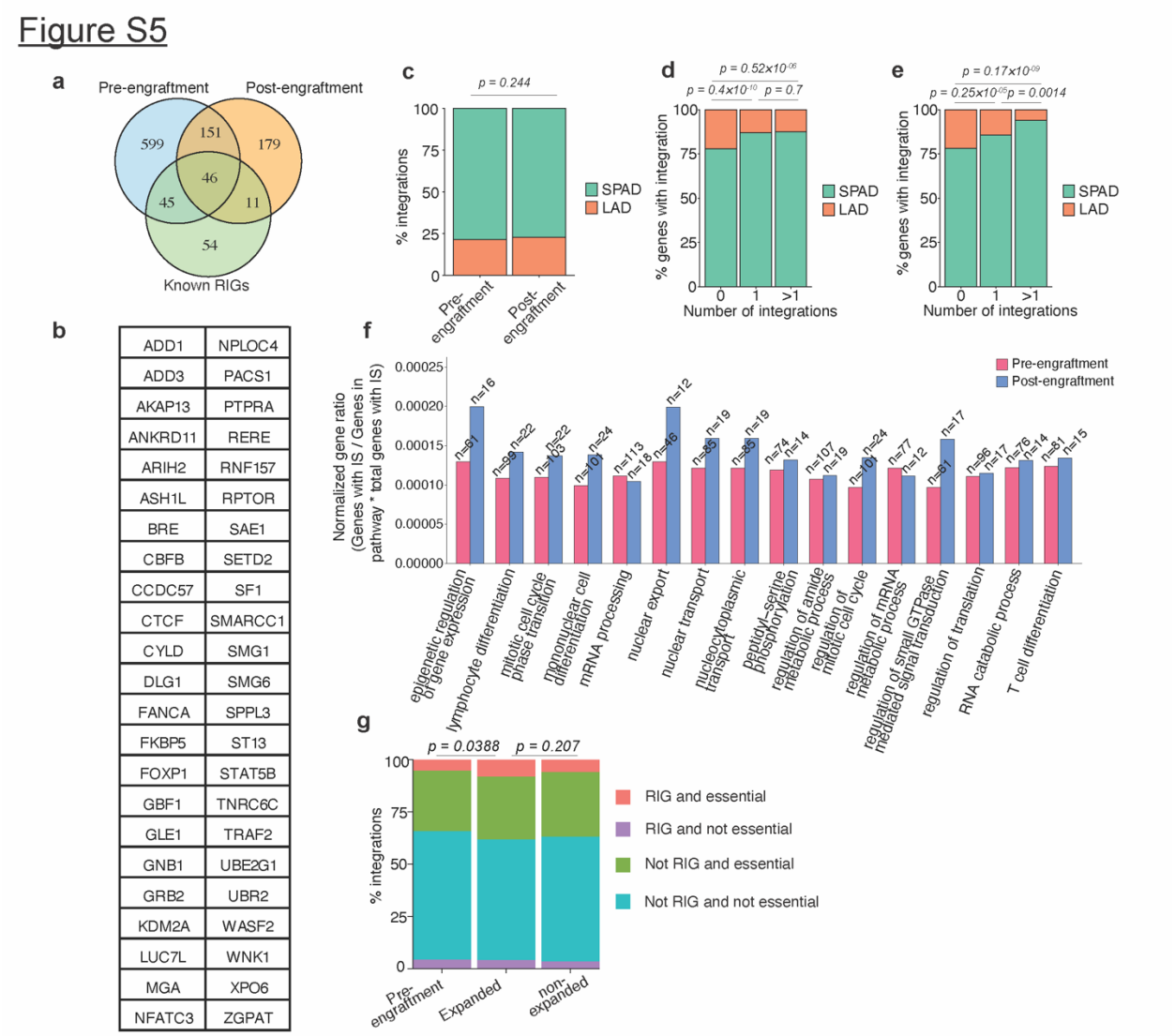
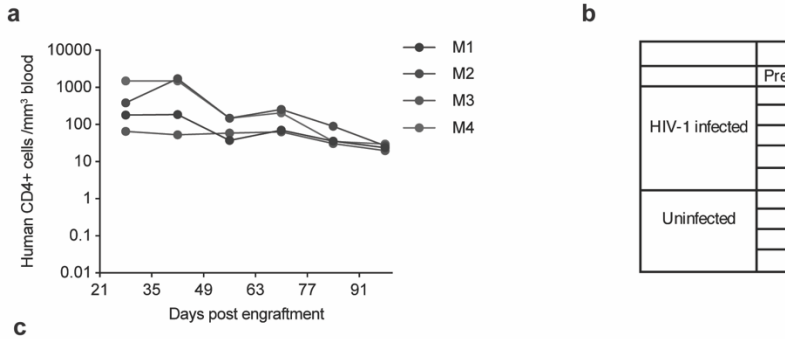


Figure S5. (a) Overlap of genes with >1 integration identified in the pre- and post- engraftment cell populations with genes designated RIGs in previous studies[51,55,56]. **(b)** List of 46 genes that overlap between the pre- /post-engraftment and the previously known RIGs. **(c)** Percentage of proviral integration in SPAD and LAD for integration sites obtained from pre- and post-engraftment populations. **(d, e)** Distribution in SPAD and LAD genomic regions for genes with 0, 1 or >1 integrations in pre- (d) and post-engraftment (e) samples. **(f)** Comparison of genes with HIV-1 integration sites in pre- and post-engraftment samples for association with various pathways as defined by gene ontology (GO) enrichment analysis. The top 15 enriched pathways based on the post-engraftment genic integration sites dataset were selected, and their respective enrichment in both pre- and post-engraftment samples compared. **(g)** Genic integration site distribution in genes that are RIG and essential, RIG and not essential, not RIG and essential, and not RIG and not essential, for pre-engraftment, as well as expanded, and non-expanded clones in post-engraftment datasets. P values determined by Pearson's Chi-squared test.

Figure S6



b

	Sample	Number of clones
	Pre-engraftment	38,621
HIV-1 infected	M1	190
	M2	126
	M3	111
	M4	121
	M5	142
Uninfected	UI 1	299
	UI 2	464
	UI 3	638
	UI 4	602

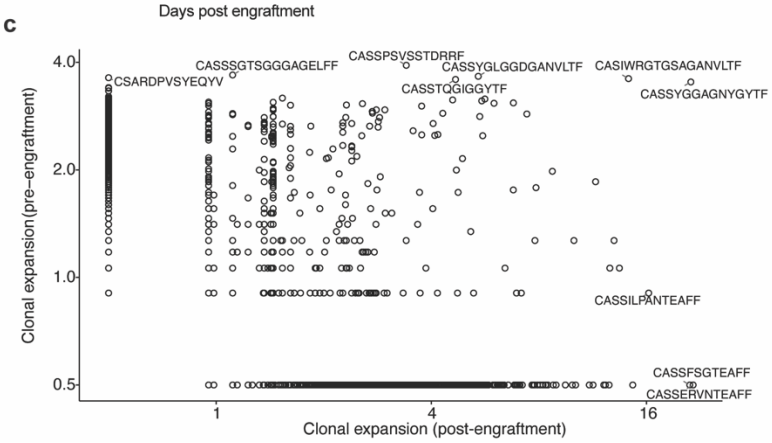


Figure S6. (a) Quantification of the number of human cells following engraftment of uninfected human memory $CD4^+$ T cells over time in multiple animals. Each mouse was sacrificed at 98 days to measure the clonal expansion of individual T cell clones. **(b)** Number of T cell clones retrieved from pre-engraftment, HIV-1-infected, and uninfected post engraftment cell populations from individual mouse spleen samples. **(c)** Comparison of TCR sequence abundance in pre- and post-engraftment cell populations. The dot plot shows the total expansion of all clones retrieved from the uninfected and infected grafts. Each dot represents a T cell clone marked by its CDR3 TCR sequence. TCR sequence not detected in one of the two populations are assigned an arbitrary value of 0.5 for representation on a logarithmic scale.

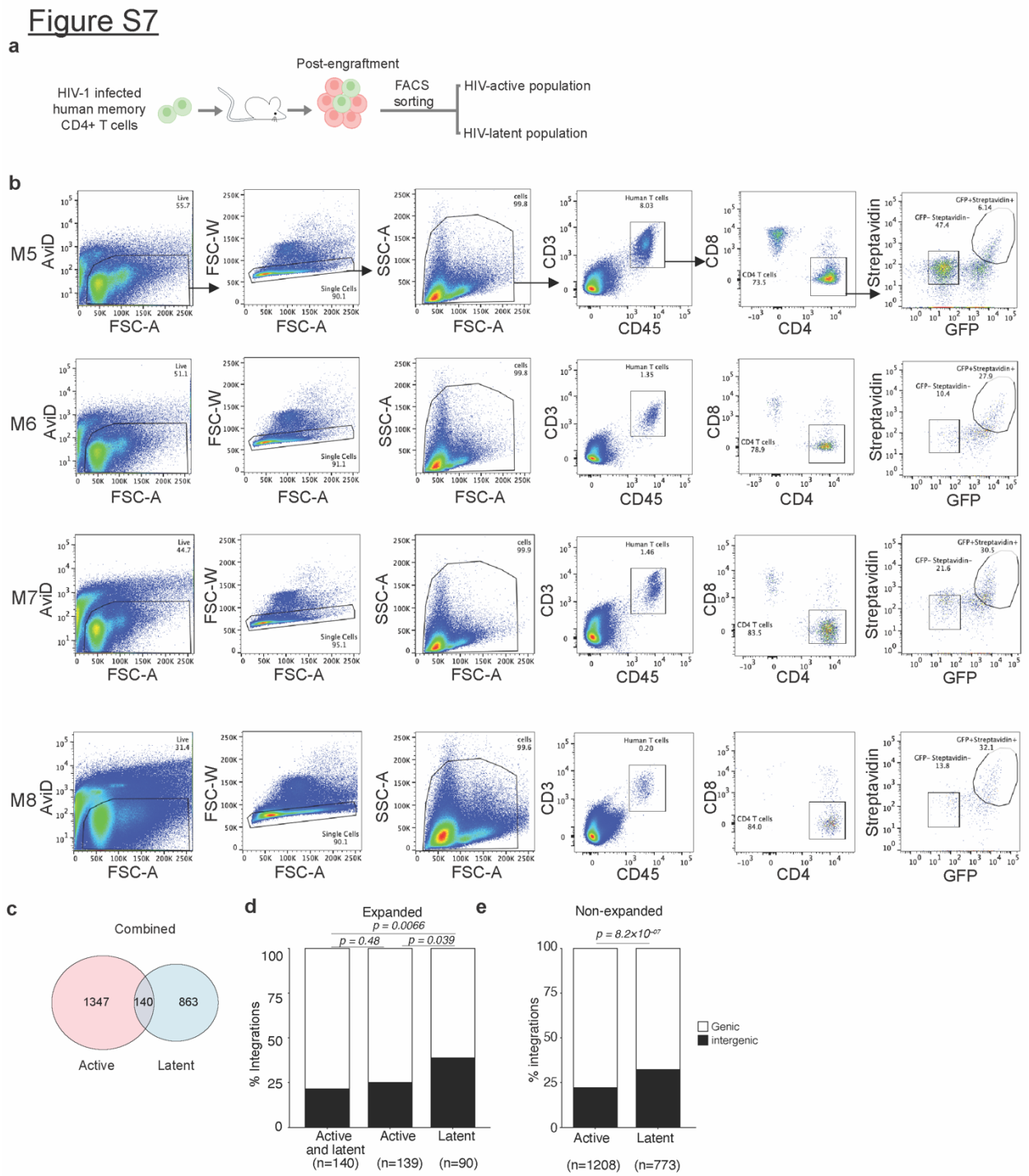


Figure S7. (a) Procedure for sorting of cell populations with active and latent HIV-1 proviruses post-engraftment. **(b)** Flow cytometry gating strategy to sort active ($GFP^{+}SBP+$) and latent ($GFP^{-}SBP^{-}$) HIV-1-infected cell subpopulations post-engraftment. **(c)** Overlap analysis of integration sites found in sorted active and latent cell populations post-engraftment, pooled data from four mice. **(d, e)** Genic and intergenic distribution of integration sites obtained in active, latent, or active and latent proviruses for expanded and non-expanded cell clones, respectively. P-value denotes Fisher's exact test.

Figure S8

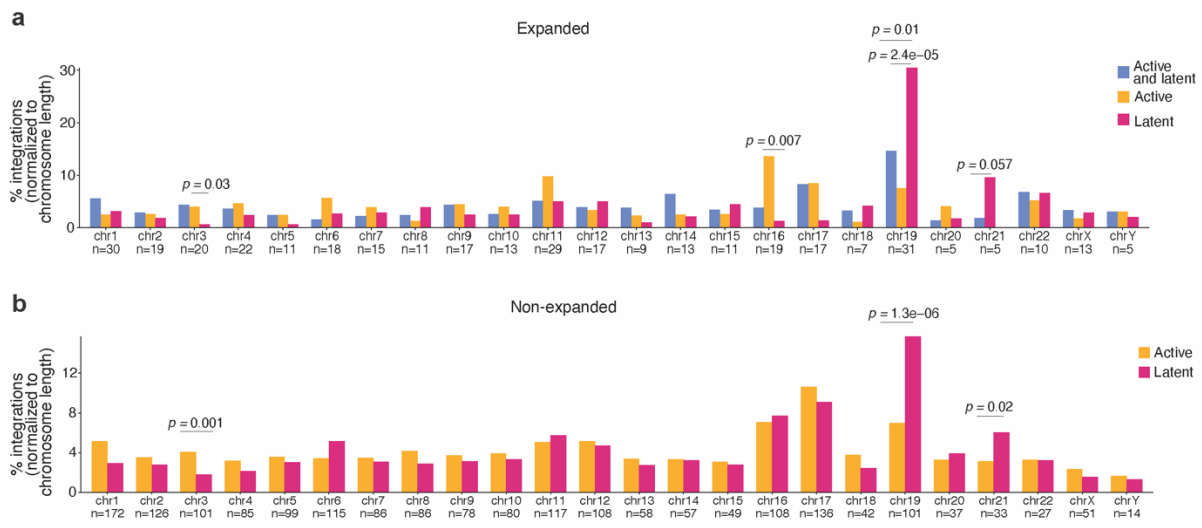


Figure S8. (a and **b)** Distribution of active, latent, or active and latent provirus integration sites found in expanded and non-expanded cell clones on human chromosomes. P-value denotes Fisher's exact test.

Figure S9

Figure S9. Karyoplot showing the chromosomal distribution of active, latent, or both active and latent provirus integration sites identified in expanded clones. Each ideogram represents a chromosome, and

integration sites are mapped to their genomic coordinates. Proviruses are colored according to transcriptional status. Cytogenetic bands are shown based on Giemsa staining patterns following UCSC Genome Browser and ISCN conventions. Cytoband color key shows white (gneg), gene-rich GC-rich euchromatin; light to dark gray (gpos25–gpos100), increasing Giemsa-positive intensity indicating progressively AT-rich, gene-poor heterochromatin; red, centromeric regions; blue, secondary constrictions containing rRNA gene clusters; gray (gvar), variable heterochromatin.

Figure S10. Karyoplot showing the chromosomal distribution of active and latent provirus integration sites identified in non-expanded clones. Each ideogram represents a chromosome, and integration sites are mapped to their genomic coordinates. Proviruses are colored according to transcriptional status, and

multiple integrations within non-expanded clones are displayed on the same chromosome. Cytogenetic bands are shown based on Giemsa staining patterns as described for Figure S9.

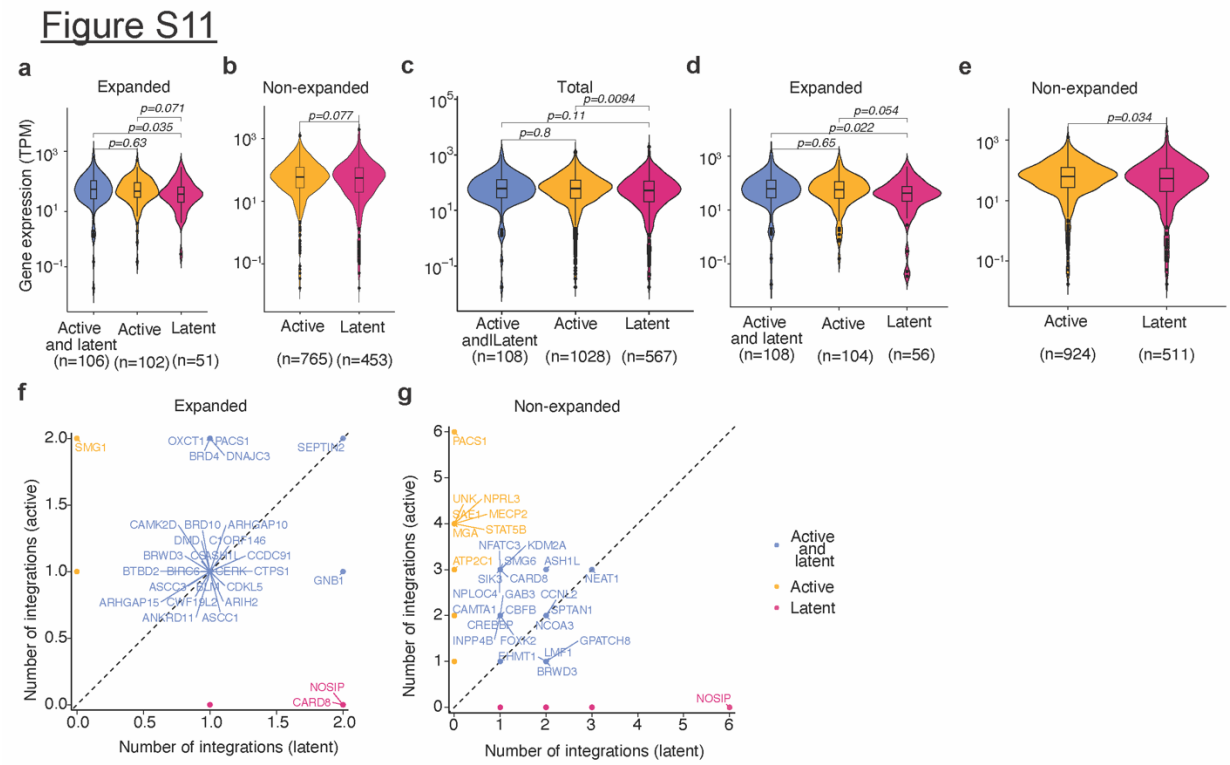


Figure S11. (a and **b)** Expression levels of genes with active, latent, or active and latent proviruses for expanded and non-expanded cell clones. **(c-e)** Gene expression associated with individual active, latent, or active and latent proviruses for expanded, non-expanded, and total cell clones. P value denotes the Wilcoxon rank-sum test with Bonferroni correction for multiple comparisons. **(f** and **g)** Number of distinct integrations in genes with active and latent proviruses found in expanded and non-expanded cell clones.

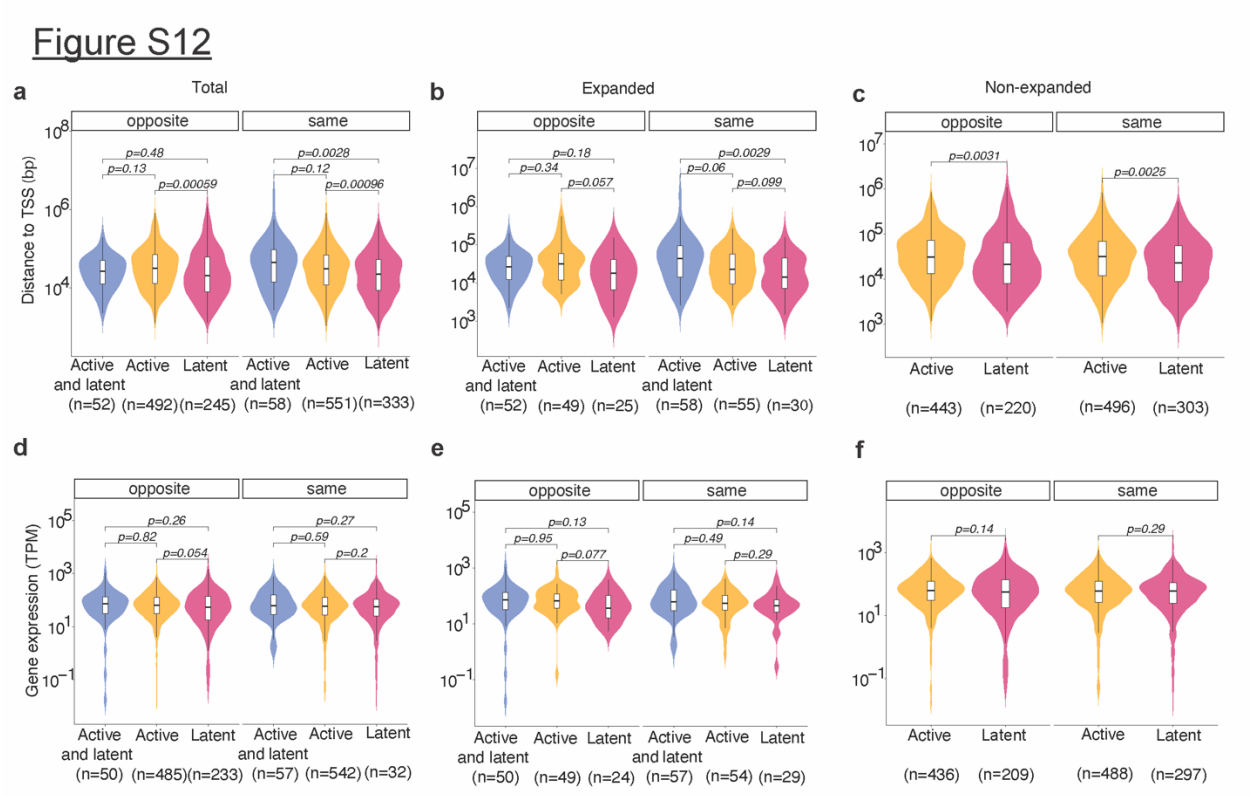


Figure S12. (a-c) Distance of integrated genic proviruses from the TSS for proviruses integrated in the same or opposite orientation of the TSS in the expanded, non-expanded, and total cell clones. **(d-f)** Expression of genes with integrated proviruses in the same or opposite orientation with respect to host transcription in expanded, non-expanded, and total cell clones. P value denotes the Wilcoxon rank-sum test with Bonferroni correction for multiple comparisons.

Figure S13

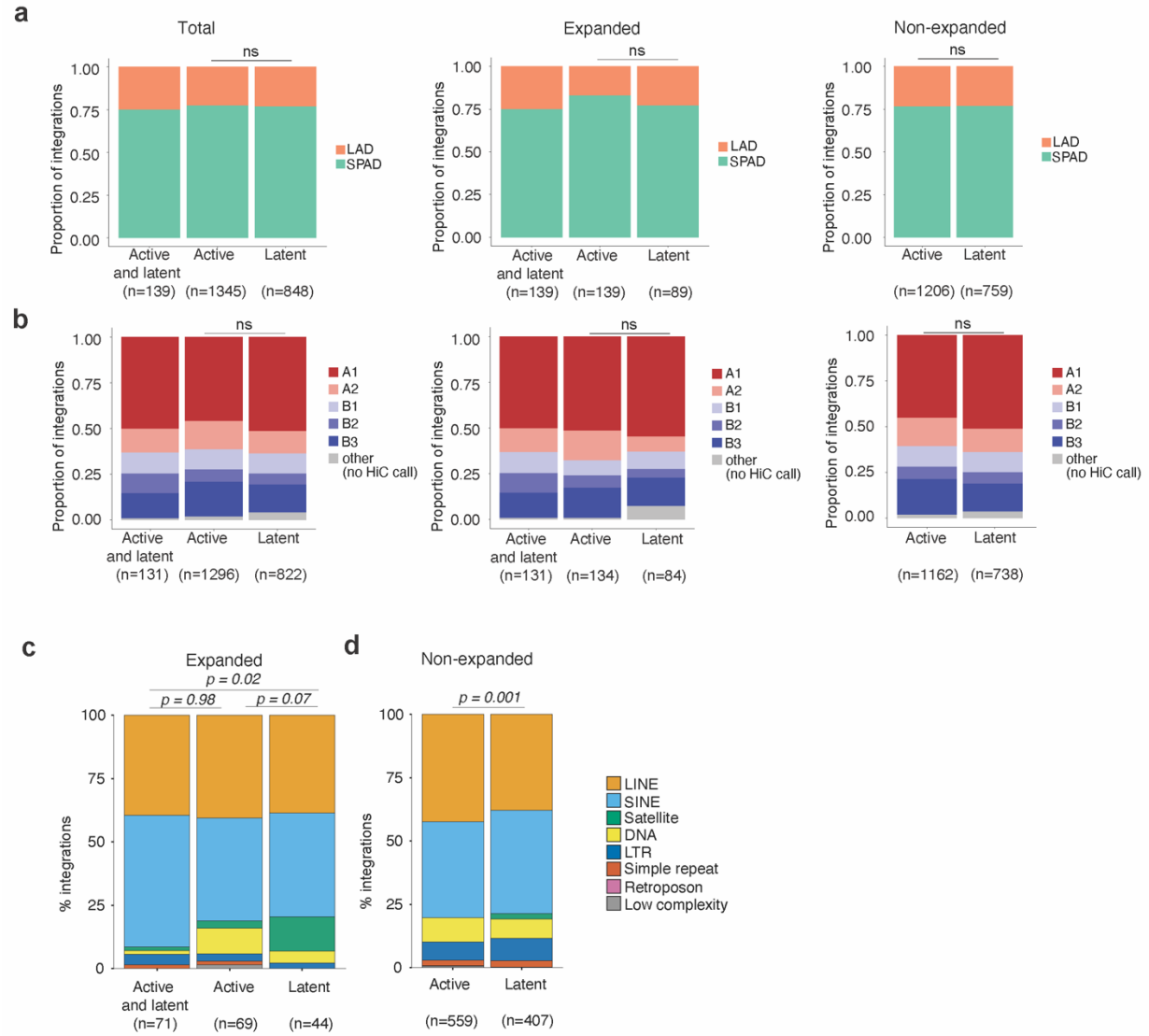


Figure S13. (a) Proportion of SPAD and LAD annotation for associated with active, latent, or active and latent proviruses in total, expanded, and non-expanded cell clones. **(b)** Proportion of active, latent, or active and latent proviruses found within chromatin structural compartments A and B and their respective sub-compartments as determined by Hi-C sequencing data for expanded, non-expanded, and total cell clones. P-value denotes the two-proportion z-test. **(c** and **d)** Percent of proviruses in distinct chromosomal repeat regions in the human genome for active, latent, or both active and latent proviruses in expanded and non-expanded cell clones. P-value denotes the two proportions z-test for the satellite repeat elements class.

Figure S14

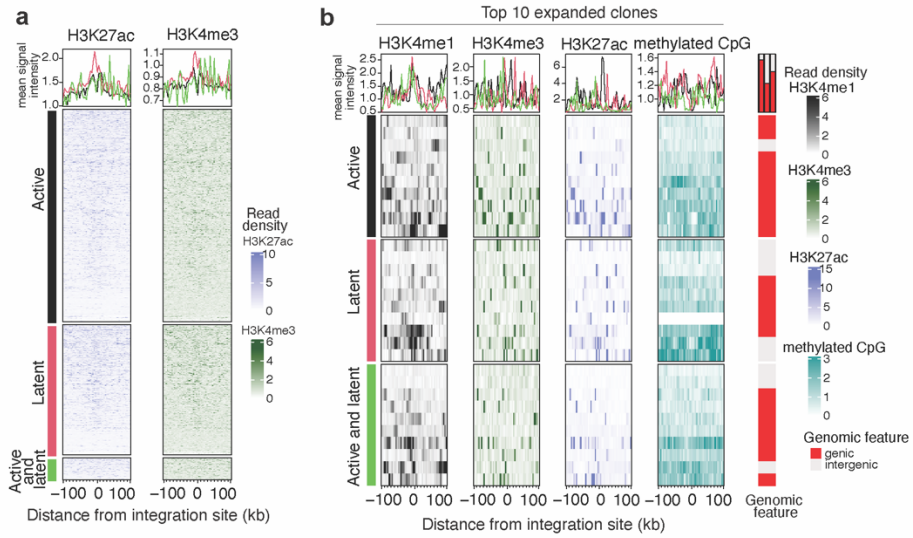


Figure S14. (a) Heat map showing levels of histone marks found up to 100kb upstream and downstream of the integration sites of active, latent, or both (active and latent) proviruses. The line plot above the heat map shows the mean signal across all integration sites found within the active (black), latent (pink), and both (green) categories. The heatmaps for each group are arranged from top to bottom according to the highest to lowest mean signal intensity for H3K36me3, as in Fig 5a. **(b)** Heat map showing levels of histone marks found up to 100kb upstream and downstream of the integration sites of active, latent, or both (active and latent) proviruses for the top 10 expanded clones in each category.

Figure S15

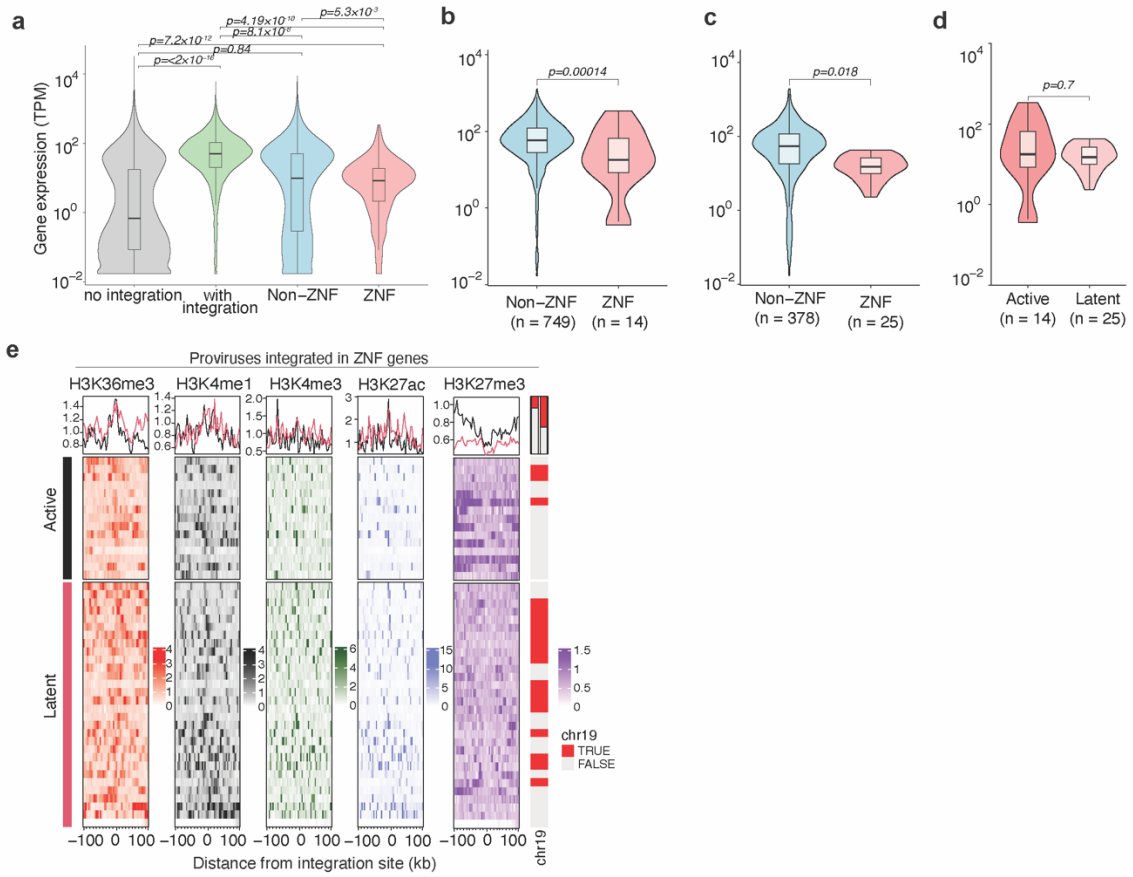


Figure S15. (a) Comparison of the level of expression of genes with or without HIV-1 integration, non-ZNF, and ZNF genes. **(b** and **c)** The expression of ZNF genes as compared to non-ZNF genes with either the active (**b**) or latent (**c**) proviruses. **(d)** Comparison of the expression of ZNF genes with active or latent proviruses. The box inside the violin plot shows median, interquartile ranges, and minimum/maximum values. P values were determined by the pairwise Wilcoxon rank-sum test. **(e)** Heat map showing levels of histone marks found up to 100kb upstream and downstream of the integration sites of active, latent, or both (active and latent) proviruses for proviruses integrated in ZNF genes.