Human immunodeficiency virus-1 induces and targets host genomic R-loops for viral genome integration

Kiwon Park[1,2¶], Dohoon Lee[3,4¶], Jiseok Jeong[1,2], Sungwon Lee[1,2], Sun Kim[5], and Kwangseog Ahn[1,2,6*]

[1]Center for RNA Research, Institute for Basic Science, Seoul 08826, Republic of Korea

[2]School of Biological Sciences, Seoul National University, Seoul 08826, Republic of Korea

[3]Bioinformatics Institute, Seoul National University, Seoul 08826, Republic of Korea

[4]BK21 FOUR Intelligence Computing, Seoul National University, Seoul 08826, Republic of Korea

[5]Department of Computer Science and Engineering, Seoul National University, Seoul 08826, Republic of Korea

[6]SNU Institute for Virus Research, Seoul National University, Seoul 08826, Republic of Korea

*Corresponding author

Email: ksahn@snu.ac.kr (KA)

¶These authors contributed equally to this work.

Abstract

Although HIV-1 integration sites are considered to favor active transcription units in the human genome, high-resolution analysis of individual HIV-1 integration sites have shown that the virus can integrate in a variety of host genomic locations, including non-genic regions. The invisible infection by HIV-1 integrating into non-genic regions challenging the traditional understanding of HIV-1 integration site selection are rather more problematic as they are selected to preserve in the host genome during prolonged antiretroviral therapies. Here, we showed that HIV-1 targets R-loops, a genomic structure made up of DNA–RNA hybrids, for integration. HIV-1 initiates the formation of R-loops in both genic and non-genic regions of the host genome and preferentially integrates into R-loop-rich regions. Using a cell model that can independently control transcriptional activity and R-loop formation, we demonstrated that the formation of R-loops directs HIV-1 integration targeting sites. We also found that HIV-1 integrase proteins physically bind to the host genomic R-loops. These findings provide fundamental insights into the mechanisms of retroviral integration and the new strategies of antiretroviral therapy against HIV-1 latent infection.

35 ## Introduction

36 Retroviruses cause permanent infection in the host by integrating their reverse-
37 transcribed viral genome into the host genome. Retroviral integration considerably impacts a
38 wide range of biological phenomena, including the persistence of fatal human diseases and
39 the shaping of metazoan evolution (*1*). Human immunodeficiency virus (HIV)-1 is a
40 representative retrovirus that underlies the global burden of acquired immune deficiency
41 syndrome (AIDS) (*2*). The chromosomal landscape of HIV-1 integration plays a critical role
42 in proviral gene expression, persistence of integrated proviruses, and prognosis of
43 antiretroviral therapy (*3-5*). Integration into the host genome is not random and displays
44 distinct preferences for gene-dense regions, where active transcription occurs (*6*), by
45 interacting host factors such as transcription activators, epigenetic marker binding proteins
46 and super enhancers (*7-13*). However, transcription activity is not the sole determinant of the
47 HIV-1 integration site landscape (*10*). For instance, the most favored region of HIV-1
48 integration is an intergenic locus, and despite the lower probability of integration, HIV-1
49 proviruses are observed in non-genic regions in the genomes of infected individuals (*4, 6*).
50 This indicates the possibility of there being an undiscovered mechanism or determinant that
51 composes the correct genomic environment for HIV-1 integration.

52 An R-loop is a three-stranded nucleic acid structure that comprises a DNA–RNA
53 hybrid and displaced strand of DNA, and has long been considered a transcription byproduct
54 (*14, 15*). R-loops in cellular genomes are enriched in actively transcribed genes as they occur
55 naturally during transcription (*14, 16*), but R-loop formation is not limited to gene body
56 regions and is widespread in the genome (*14*). As a result of *in trans* R-loop formation, R-
57 loops are also abundant in non-genic regions, such as intergenic regions, repetitive sequences,

including transposable elements, centromeres, or telomeres (*14, 17-19*), independently of transcription activity of the genes harboring the R-loops. Although R-loops are identified as critical intermediates and regulators in a number of biological processes (*14, 15, 20*), the dynamics and the role played by cellular R-loops in pathological contexts remain unrevealed.

R-loops are important contributors molding the genomic environment and spatial organization of the cellular genome, and can potentially take a novel role in host-pathogen interaction. In the cellular genome, R-loops relieve superhelical stresses and are often associated with open chromatin marks and active enhancers (*21, 22*), which are also distributed over HIV-1 integration sites (*6, 9, 10*). In the case of transcription-induced R-loop formation, a guanine-quadruplex (G4) structure can be generated in the non-template DNA strand of the R-loop (*23*). A recent study has shown that G4 DNA can influence both productive and latent HIV-1 integration (*24*). In addition, R-loops are prevalent non-canonical B-form DNA structures (*25*) and intermediates between B-form DNA and A-form RNA conformation (*26*), which have recently been disclosed to be the conformational characteristics of the target DNA during retroviral integration (*26, 27*). The accumulated evidence implicates that host genomic R-loops are undiscovered host factor in HIV-1 integration site selection mechanism, which dynamically interact with the host genomic environment.

80 Here, we showed a notable role of R-loops in the interaction between HIV-1 and its
81 host, specifically in HIV-1 integration. HIV-1-infection induces host cellular R-loop
82 formation and the R-loop rich regions of the host genome are preferred by HIV-1 integration.
83 HIV-1 integrase proteins showed considerable binding affinity to nucleic acid substrate
84 comprising R-loop structures. Our results suggest that R-loops are an important composer of
85 host genomic environment for HIV-1 integration site determination.

86

87 # Results

88 **Host genomic R-loops accumulate by HIV-infection**

89 To investigate the relationship between HIV-1 infection and host cellular R-loops, we
90 first analyzed R-loop dynamics in different types of cells infected with HIV-1 at early post-
91 infection time points using DNA–RNA immunoprecipitation followed by cDNA conversion
92 coupled to high-throughput sequencing (DRIPc-seq) using a DNA–RNA hybrid-specific
93 binding antibody, anti-S9.6 (*28*). HeLa cells, primary $CD4^+$ T cells isolated from two
94 individual donors and $CD4^+/CD8^-$ T cell lymphoma Jurkat cell line were infected with VSV-
95 G-pseudotyped HIV-1-EGFP and harvested at 0, 3, 6, and 12 h post infection (hpi) for
96 DRIPc-seq library construction (Fig. 1A and S1A-C Fig.). Our DRIPc-seq analysis yielded
97 loci specific R-loop signals at the referenced R-loop-positive loci (RPL13A and CALM3) and
98 an R-loop-negative locus (SNRPN) (*28*) that were both strand-specific and highly sensitive to
99 pre-immunoprecipitation in vitro RNase H treatment, in HeLa cells, $CD4^+$ and Jurkat T cells
100 (Table S1-3). Notably, the number of DRIPc-seq peaks mapped to the human reference
101 genome increased remarkably during early post infection of HIV-1 (3 and 6 hpi for HeLa
102 cells and 6 and 12 hpi for $CD4^+$ and Jurkat T cells; Fig. 1B). Most of the peaks mapped in

103 cells harvested at 0 hpi were commonly found in all other samples, but a significant numbers
104 of unique peaks were observed after infection (Fig. 1C).

105 In addition to our DRIPc-seq data analysis, we used different biochemical
106 approaches to examine R-loop accumulation after HIV-1 infection in HeLa cells. First, R-
107 loop accumulation in HIV-1-infected cells was observed using DNA–RNA hybrid dot blots
108 with the anti-S9.6 antibodies (Fig. 1D). The dot intensity increased significantly upon HIV-1
109 infection at 6 hpi and the enhanced R-loop signals on dot blots of HIV-1-infected cells were
110 highly sensitive to *in vitro* treatment with RNase H (Fig. 1D). This result was highly
111 consistent with our DRIPc-seq data analysis results in HIV-1-infected HeLa cells.
112 Subsequently, we observed HIV-1-induced R-loops using an immunofluorescence assay by
113 probing HIV-1-infected or non-infected control cells with S9.6 antibody at 6 hpi (Fig. 1E,
114 left). The nuclear fluorescence signal associated with the R-loops after subtracting the
115 nucleolar signal was significantly enhanced in cells infected with HIV-1 (Fig. 1E, right). We
116 validated and quantified HIV-1-infection induced R-loop formation on the host genome in a
117 genome-site specific manner by using DRIP followed by real-time polymerase chain reaction
118 (DRIP-qPCR). In this experiment, the S9.6 signal was determined for three and two HIV-1-
119 induced-R-loop-positive (P1, P2, and P3) and -negative regions (N1 and N2), respectively,
120 where were defined by DRIPc-seq data analysis (S2A-E Fig.). We detected significantly
121 increased R-loop signals that are highly sensitive to RNase H treatment of pre-
122 immunoprecipitates in the P1, P2, and P3 regions of HIV-1-infected cells at 6 hpi compared
123 to those in the cells harvested at 0 hpi (S3A Fig.). However, the HIV-1-induced R-loop-
124 negative regions, N1 and N2, did not show significant R-loop accumulations (S3A Fig.).

125

126 Importantly, the R-loop signal was enriched even in cells infected with HIV-1 when
127 the reverse transcription or integration of HIV-1 is blocked by enzyme inhibitors like
128 Nevirapine (NVP) or Raltegravir (RAL), respectively (S3B and S3C Fig.). This result
129 indicates that the enrichment of R-loop signals in cells are originated from the host genome
130 but not by DNA-RNA hybrid formation during the viral life cycle or transcriptional burst
131 from integrated HIV-1 proviruses. In addition, we confirmed that nearly 100% of DRIPc-seq
132 reads from HIV-1-infected HeLa, $CD4^+$ and Jurkat T cells were aligned to the host cellular
133 genome, but not on that of HIV-1 (S3D Fig.). Together, these data demonstrate that HIV-1
134 infection induced host genomic R-loop formation at early post-infection.

135 **R-loops accumulation after HIV-1 infection are widely distributed in both genic and**
136 **non-genic regions**

137 To investigate the distribution of cellular genomic R-loops during HIV-1 infection,
138 we conducted a genome-wide analysis of our DRIPc-seq data. The unique DRIPc-seq peaks
139 observed after HIV-1 infection were not only numerous but also relative longer (Fig. 2A).
140 This suggests that R-loops induced by HIV-1 infection occupy a genomic region larger than
141 that of the R-loops presents without HIV-1 infection. We observed a significant accumulation
142 of R-loops over diverse genomic compartments at the hpi of HIV-1-infection induced R-loop
143 formation (Fig. 2B). The presence of R-loops is often correlated with high transcriptional
144 activity, and we found significantly high proportion of DRIPc-seq peaks enrichment upon
145 HIV-1 infection in the gene body regions (Fig. 2B). However, we also observed enrichment
146 of HIV-1-infection induced DRIPc-seq peaks proportion mapped to intergenic or repeat
147 regions, including short interspersed nuclear elements (SINEs), long interspersed nuclear
148 elements (LINEs), and long terminal repeat (LTR) retrotransposons, where transcription is

typically repressed (Fig. 2B). Although the expression of repetitive elements is mostly repressed during normal cellular activities, HIV-1 infection could activate endogenous retroviral promoters (*29, 30*). To investigate the possibility that R-loop induction in gene-silent regions is associated with transcriptome changes during HIV-1 infection, we performed RNA sequencing (RNA-seq) for HIV-1-infected HeLa cells at 0, 3, 6, and 12 hpi. Consistent with previous reports, we observed an increase in the expression levels of repetitive elements at later time points post-infection (S4A Fig.; 12 hpi). In contrast, we found that the expression levels of SINEs, LINEs, and LTRs were even lower at both 3 and 6 hpi compared to 0 hpi while HIV-1-induced R-loops were significantly accumulated, compared to 0 hpi (S4A Fig.). We further examined the expression profile of genes containing R-loop in HeLa cells. The expression profile of genes harboring HIV-1-induced R-loops in their gene bodies showed very weak correlations with the signals of DRIPc-seq peaks at 3 hpi (Pearson's $r = 0.21$, $P = 1.08 \times 10^{-84}$; Fig. 2C) and at 6 hpi samples (Pearson's $r = -0.34$, $P = 2.40 \times 10^{-228}$; Fig. 2C), which implies that the unique R-loop peaks upon HIV-1 infection do not engage with transcriptional burst. In agreement with our DRIPc-seq and global RNA-seq data analysis, the expression level of the genes harboring HIV-1-infection induced R-loops, which were quantified by DRIP-qPCR (S3A Fig.), were not significantly affected by HIV-1 infection (S4B Fig. and Table S4). Together, our data demonstrate that host cellular R-loop accumulation upon HIV-1 infection are widely distributed in both genic and non-genic regions and are not necessarily correlate with the expression levels of the genes harboring the R-loops.

172 **HIV-1 integration sites are enriched at systemically induced sequence-specific R-loop**
173 **regions in cell model**

174 HIV-1 completes its infection by integrating its viral genome into the host's through
175 dynamic interaction with the host genome (*31*). Besides, as HIV-1 infection induced R-loop
176 accumulation at early post infection hours when HIV-1 genome are imported into nucleus and
177 integration may initiate (*32-34*), we hypothesized that host genomic R-loops play a role in
178 HIV-1 integration, and possibly in integration site selection. To systemically and directly
179 assess the relationship between host genomic R-loops and HIV-1 integration in a genome-
180 site-specific manner, we adapted and modified an elegantly designed episomal system that
181 induces sequence specific R-loops through DOX-inducible promoters (*16*). To most closely
182 mimic the presence R-loop in host cellular genome, we subcloned the R-loop-forming portion
183 of the mouse gene encoding AIRN (mAIRN) (*17*) or non-R-loop-forming ECFP sequence
184 with a DOX-inducible promoter into the piggyBac transposon vector and co-expressed the
185 piggyBac transposase in HeLa cells. These R-loop forming (mAIRN) or non-R-loop forming
186 sequence (ECFP) are non-human sequences. Therefore, our cell model allows us to induce
187 and quantify R-loop formation at designated genomic region and distinguish the R-loop
188 formation from the endogenous R-loops on the cellular genome, which are not sequence-
189 specific and impossible to control for induction. Moreover, by using this system we can
190 quantify R-loop-dependent site-specific HIV-1 integration events at the designated regions,
191 which can also be distinguished from HIV-1 integration event at endogenous host genomic
192 loci. We designated the pool of cells with the R-loop forming sequence (mAIRN) inserted
193 into its genome as "pgR-rich (piggyBac R-loop rich)" cell line and the pool of cells with the
194 non-R-loop forming sequence (ECFP) inserted into its genome as "pgR-poor (piggyBac R-
195 loop poor)" cell line (Fig. 3A).

196 A similar number of the copies of piggyBac transposon was successfully delivered to
197 the genome of each cell line (S5A Fig.), and DOX treatment strongly induced the
198 transcriptional activity of mAIRN or ECFP without affecting the transcription of endogenous
199 loci in both cell lines (S5B and S5C Fig.). Although the transcription of mAIRN or ECFP was
200 strongly induced upon DOX treatment, the activity did not exceed that of endogenous loci in
201 both cell lines (S5D and S5E Fig.). While two cell lines showed comparable level of DOX-
202 inducible transcription activity at the designated sequences (Fig. 3B), only pgR-rich cells
203 exhibited robust RNase H-sensitive stable R-loop formation upon DOX treatment (Fig. 3C,
204 mAIRN). By contrast, R-loops were weakly formed in the pgR-poor cells where non-R-loop
205 forming sequence (ECFP) inserted into its genome (Fig. 3C, ECFP).

206 To examine whether the formation of 'extra' R-loops in the host genome influence
207 HIV-1-infection to the host cells, we infected both cell lines with VSV-G-pseudotyped HIV-
208 1-luciferase viruses and examined the luciferase activity. Interestingly, we found that pgR-
209 rich cells showed significantly high luciferase activity only when R-loops were induced by
210 DOX treatment, whereas pgR-poor cells showed comparable luciferase activity regardless of
211 transcription activation by DOX treatment (Fig. 3D). We conducted HIV-1 integration site
212 sequencing in HIV-1-infected pgR-poor and pgR-rich cells to directly quantify site-specific
213 integration events at sequence-specific R-loop regions. Remarkably, integration events were
214 significantly higher in pgR-rich cells only when R-loops were induced by DOX treatment
215 (Fig. 3E). However, HIV-1 integration frequency within non-R-loop forming sequence in
216 pgR-poor cells remained very low, even with transcription activation by DOX treatment (Fig.
217 3E). HIV-1 integration frequency was enriched at the vicinity of R-loop forming regions in
218 pgR-rich cell line upon DOX treatment, but the enrichment was not observed in pgR-poor
219 cells that does not form stable R-loops even after transcription activation by DOX treatment

220 (Figs. 3F and 3G). This cell-based R-loop inducing system with independent control over
221 transcription and R-loop formation enabled the direct measurement of HIV-1 integration
222 events at the defined R-loop regions, and the results indicate that host genomic R-loops are
223 targeted by HIV-1 integration. Moreover, our data suggest that transcription activity itself is
224 not sufficient for HIV-1 integration site determination, but the formation of R-loops accounts
225 for HIV-1 integration site selection.

226 **Host genomic R-loops are targeted by HIV-1 integration**

227 We attempted to further validate the relationship between R-loops and the HIV-1
228 integration site selection by global analysis of HIV-1 integration sites on endogenous
229 genomic regions of HIV-1 infected host cells. We performed HIV-1 integration site
230 sequencing in HIV-1 infected HeLa cells, $CD4^+$ and Jurkat T cells and analyzed the
231 sequencing data combined with our DRIPc-seq data. We counted and compared the number
232 of successfully integrated proviruses in the R-loop regions (the combined genomic regions
233 within 30-kb windows centered on DRIPc-seq peaks from 0, 3, 6, and 12 hpi) to those in non-
234 R-loop forming regions (the total genomic regions outside of the 30-kb windows centered on
235 DRIPc-seq peaks). Notably, we found that approximately three to four times more integration
236 were detected in the R-loop regions as in other genomic regions without R-loops in HeLa
237 cells, $CD4^+$ and Jurkat T cells (Fig. 4A). Interestingly, the HIV-1 integration sites preferred
238 the center and nearby areas of the R-loops regions (Fig. 4B). We observed biases for HIV-1
239 integration in HIV-1-induced R-loop-positive regions, P2 and P3, where gave highly induced
240 R-loop signal upon HIV-1 infection in DRIPc-seq analysis and DRIP-qPCR (Fig. 4C). By
241 contrast, HIV-1 integration sites were not detected in R-loop-negative regions, N1 and N2
242 (Fig. 4D). Overall, our results from bioinformatics analysis using different types of naïve host

243 cells infected with HIV-1 are consistent with the idea that the virus has a preference for
244 targeting R-loops for integration (Fig. 3), and our data suggest R-loops as an important
245 composer of host genomic environment for HIV-1 integration site determination.

246 **HIV-1 integrase physically interacts with R-loops on the host genome**

247 HIV-1 intasome tether to the host genome for its viral cDNA integration. Intasomes
248 consist of HIV-1 viral cDNA and HIV-1 coding protein, integrases. We observed that HIV-1
249 preferentially integrated into R-loops on the host genome, thus we hypothesized that the HIV-
250 1 integrase protein could directly bind and be recruited to the genomic R-loops. To test this
251 hypothesis, we first investigated whether HIV-1 integrase proteins have physical binding
252 affinity to nucleic acid substrates possessing R-loop structure. Although HIV-1 integrases are
253 DNA and RNA binding proteins (*35, 36*), its binding ability towards such three-stranded
254 nucleic acid structure that is composed with a DNA-RNA hybrid like R-loop has not been
255 investigated. We carried *in vitro* protein-nucleic acid binding assay by electrophoretic
256 mobility shift assay (EMSA) with Sso7d-tagged HIV-1 integrase recombinant proteins and
257 diverse structures of nucleic acid substrates including R-loop and simple dsDNA duplex.
258 Interestingly, nucleic acid substrate consisted with R-loop structure bound to HIV-1 integrase
259 proteins with greater binding affinity than simple dsDNA duplex (Fig. 5A). Additionally, R-
260 loop composing forms of nucleic acid structures such as RNA-DNA hybrid with exposed
261 ssDNA (R:D+ssDNA) and RNA-DNA hybrid (hybrid) also hosed high binding affinity to
262 integrases (S6A Fig. and Fig. 5A).

263 We validated the interaction between cellular genomic R-loops and HIV-1 integrase
264 proteins by DNA–RNA hybrid immunoprecipitation using S9.6 antibodies in FLAG-tagged
265 HIV-1 integrase-expressing HeLa cells (Fig. 5B). Under our experimental conditions, R-loops

266 were reproducibly immunoprecipitated (S6B Fig.) and HIV-1 integrase proteins co-
267 immunoprecipitated with R-loops (Fig. 5C). DNA–RNA hybrids also co-immunoprecipitated
268 with the positive control H3 (*37*) but not with the negative control LaminA/C and Actin (*37*)
269 (Fig. 5C). To verify the specificity of our co-immunoprecipitation results for R-loops and
270 HIV-1 integrases, we performed DNA–RNA hybrid immunoprecipitation with RNase H
271 treatment (S6C Fig.). The S9.6 signal of immunoprecipitated nucleic acids was highly
272 sensitive to RNase H treatment of pre-immunoprecipitates (Fig. 5D). Accordingly, the
273 blotting signal of the co-immunoprecipitated HIV-1 integrase and H3 proteins was
274 significantly reduced upon RNase H treatment (Fig. 5E). We performed reciprocal
275 immunoprecipitation using an anti-FLAG monoclonal antibody and detected
276 immunoprecipitated R-loops using dot blot analysis with anti-S9.6. R-loops were
277 immunoprecipitated by HIV-1 integrase, and the S9.6 signal of immunoprecipitated nucleic
278 acids was highly sensitive to RNase H treatment (Fig. 5F and S6D Fig.). Subsequently, we
279 attempted to observe the interaction between the R-loops and HIV-1 integrase using a
280 proximity-ligation assay (PLA), in HIV-1-infected cells. We used two antibodies: one that
281 binds to R-loops (anti-S9.6) and another one that binds to GFP-tagged HIV-1 integrase. We
282 detected PLA signals in cells infected with HIV-IN-EGFP virions and in non-infected control
283 cells. PLA signals in non-infected cells were comparable to those in S9.6-alone and GFP-
284 alone single antibody-negative controls; however, PLA signals significantly increased upon
285 HIV-1 infection (Fig. 5G and S6E Fig.). Our data suggest that the HIV-1 frequently targets R-
286 loop-rich regions for viral genome integration by physical binding of HIV-1 integrase
287 proteins to R-loop structures on the host genome.

288

289

290 ## Discussion

291 In this study, we found that HIV-1 preferentially integrates into regions rich in R-
292 loops, suggesting that R-loops are a novel host factor governing HIV-1 integration site
293 selection. In our bioinformatics analysis, host cellular R-loops were induced by HIV-1
294 infection and widespread over host genomic regions. Using our R-loop-inducible cell models,
295 R-loop formation, not necessarily transcription activity itself, was found to be important for
296 HIV-1 integration site determination. In addition, HIV-1 integrase proteins favored physical
297 binding with R-loops *in vitro*, and they interacted with host genomic R-loops in HIV-1-
298 infected cells. These results demonstrated that HIV-1 exploits and frequently targets the host
299 genomic R-loops for successful integration and infection.

300 Our data show that HIV-1 targets host genomic R-loops for viral genome integration
301 and its integrase proteins physically interact with genomic R-loops *in vitro* and in cells. This
302 may because the R-loops own an unique nucleic acid conformation of B-form DNA and A-
303 form RNA intermediates, which possess intrinsic preferential binding ability to retroviral
304 intasome (*25-27*). Another possible explanation for why HIV-1 integration shows a
305 preference towards host genomic R-loops is that R-loops perhaps take a collaborative role
306 with host factors governing the HIV-1 integration site selection. Cellular R-loops are
307 recognized and regulated by numerous cellular proteins (*37, 38*). Besides, the correct
308 genomic environment for HIV-1 integration site selection are composed by host proteins (*9*).
309 LEDGF/p75 (*9, 13, 39*) and CPSF6 (*7, 9*) are two decisive host factors that direct HIV-1
310 integration by interacting with integrase or trafficking viral preintegration complex towards
311 nuclear interior (*7, 9*). In fact, these host factors have recently been identified as potential R-
312 loop binding proteins in DNA–RNA interactome analysis (*37*) and R-loop proximity
313 proteomics (*38*), respectively. R-loops are tightly regulated by DNA damage response

314 proteins (*40*) and DNA repair machineries take important roles in HIV-1 integration process
315 (*31*). For example, the Fanconi anemia pathway (*41, 42*), a well-known R-loop regulatory
316 pathway, has been recently proposed as an HIV-1 integration regulatory factor exploited by
317 HIV-1 (*43*). Taking into account theses previous studies alongside our current findings, we
318 propose R-loops as another pivotal host factor driving HIV-1 integration site determination
319 and as a possible intermediate regulator of HIV-1 integration site selection by such host
320 proteins.

321 Viruses often take advantage of various host factors, and targeting viral components
322 that manipulate the host cellular environment can be an effective strategy for antiviral
323 therapy. Our study has shown that host genomic R-loops accumulate significantly shortly
324 after HIV-1 infection. Thus, it is possible that virion-associated HIV-1 proteins are
325 responsible for inducing these R-loops. For instance, the HIV-1 accessory protein Vpr causes
326 genomic damage (*44*) and transcriptomic changes during the early stages post infection(*45*),
327 both of which can lead to *in cis* and *in trans* R-loop formation (*15*). Another HIV-1 accessory
328 protein, Vif, counteracts the host antiviral factor, APOBEC3 (*46, 47*), which were recently
329 found to regulates cellular R-loop levels (*48*). Identifying the HIV-1 components responsible
330 for inducing host cellular R-loops, and elucidating the mechanism by which they induce
331 genome-wide R-loop formation and contribute to successful viral integration into selective
332 genomic regions, represents an area for further research.

333 Although most HIV-1 integration occurs in genic regions (*4, 6*), HIV-1 proviruses are
334 also found in non-genic regions (*49*) and understanding these "transcriptionally silent"
335 proviruses is critical for developing strategies to completely eliminate HIV-1. In HIV-1 elite
336 controllers, who suppress viral gene expression to undetectable levels, HIV-1 proviruses

337 accumulate in heterochromatic regions (*5*). Moreover, proviruses with lower expression level
338 can persist in the host genome even during antiretroviral therapy (*4*). However, the
339 mechanism by which HIV-1 targets gene-silent regions for "invisible" integration remains
340 unclear. Our study has revealed that R-loops are enriched in both genic and non-genic regions
341 during HIV-1 infection, and that the virus preferentially targets these R-loops for integration.
342 We propose that R-loops, particularly those enriched in non-genic regions, may represent the
343 mechanism by which the virus achieves "invisible" and permanent infection.

344

345 Materials and methods

346 Cell culture

347 HeLa and HEK293T cells were cultured in Dulbecco's modified Eagle's medium (Gibco)
348 supplemented with 10% (v/v) fetal bovine serum (FBS, Cytiva), antibiotic mixture (100
349 units/ml penicillin–streptomycin, Gibco), and 1% (v/v) GlutaMAX-I (Gibco). Jurkat cells
350 were cultured in Roswell Park Memorial Institute (RPMI) 1640 medium (ATCC)
351 supplemented with 10% (v/v) FBS (Cytiva). Cells were incubated at 37°C and 5% CO_2.

352 Virus production and infection

353 VSV-G-pseudotyped HIV-1 virus stocks were prepared by performing standard
354 polyethylenimine-mediated transfection of HEK293T monolayers with pNL4-3 ΔEnv EGFP
355 (NIH AIDS Reagent Program 11100) or pNL4-3. Luc.R-E (NIH AIDS Reagent Program,
356 3418) along with pVSV-G at a ratio of 5:1. HIV-IN-EGFP virions were produced by
357 performing polyethylenimine-mediated transfection of HEK293T cells with 6 μg of pVpr-IN-
358 EGFP, 6 μg of HIV-1 NL4-3 non-infectious molecular clone (pD64E; NIH AIDS Reagent

359 Program 10180), and 1 μg of pVSV-G. The cells were incubated for 4 h before the medium
360 was replaced with fresh complete medium. Virion-containing supernatants were collected
361 after 48 h, filtered through a 0.45-μm syringe filter, and pelleted using the Lenti-X
362 Concentrator (631232; Clontech) according to the manufacturer's instructions. The
363 multiplicity of infection (MOI) of virus stocks was determined by transducing a known
364 number of HeLa cells with a known amount of virus particles and then counting GFP-positive
365 cells using flow cytometry. For luciferase reporter HIV-1 virus, the HIV-1 p24 antigen
366 content in viral stock were quantified using the HIV1 p24 ELISA kit (Abcam, ab218268),
367 according to the manufacturer's instruction. For virus infection, HeLa cells were seeded at a
368 density of 0.5–4 × 10^5 cells/mL on the day before infection. The culture medium was
369 replaced with fresh complete culture medium 2 hpi. The infected cells were washed twice
370 with PBS and harvested at the indicated time points. Jurkat cells were seeded at a density of
371 1× 10^6 cells/mL and inoculated with 300ng/p24 capsid antigen. The plates were centrifuged
372 at 1000 *g* at 30°C for 1 h. The medium was replaced with fresh RPMI 2 h after infection.

373 **Primary cell isolation, culture, T cell activation, and infection**

374 For $CD4^+$ T cells isolation, human PBMC (ST70025, STEMCELL Technologies) was mixed
375 and incubated with MACS CD4 MicroBeads (130-045-101, Miltenyi Biotec) and FITC-
376 conjugated mouse anti-CD4 (561005, BD Bioscience) according to the manufacturer's
377 instructions. Then the $CD4^+$ T cells were enriched by using LS Columns (130-042-401,
378 Miltenyi Biotec) and MidiMACS Separator (130-042-302, Miltenyi Biotec). The efficiency
379 of magnetic separation was analyzed by using Flow-Activated Cell Sorter Canto II (BD
380 Bioscience) and Flowjo software (Flowjo).

381

382 $CD4^+$ T cells were cultured in Roswell Park Memorial Institute (RPMI) 1640 medium
383 (Gibco), supplemented with 10% (v/v) fetal bovine serum (FBS, Cytiva), antibiotic mixture
384 (100 units/ml penicillin–streptomycin, Gibco), 1% (v/v) GlutaMAX-I (Gibco), and 20 ng/ml
385 of IL-2 (PHC0026, Gibco), left in resting state or activated with Dynabeads Human T-
386 Activator CD3/CD28 (1161D, Thermo Fisher Scientific) for 72 h. $CD4^+$ T cells activation
387 efficiency was assessed by staining cells with FITC-conjugated mouse anti-CD25 (340694,
388 BD Bioscience) and APC-conjugated mouse anti-CD69 (130-114-046, Miltenyi Biotec) and
389 using Flow-Activated Cell Sorter Canto II (BD Bioscience) and Flowjo software (Flowjo).

390 Purified and activated $CD4^+$ T cells were seeded at a density of 1×10^6 cells/mL and
391 inoculated with 600ng/p24 capsid antigen in presence of polybrene. The plates were
392 centrifuged at 1000 *g* at 30°C for 1 h. The medium was replaced with fresh RPMI 2 h after
393 infection.

394 **DRIP-qPCR**

395 DRIP was performed as described for the construction of the DRIPc-seq library. After the
396 elution of isolated complexes, nucleic acids were purified using the standard phenol-
397 chloroform extract method and used for qPCR. S6 Table presents details of the primer
398 sequences used for DRIP-qPCR analysis.

399 **RNA-seq library construction**

400 For RNA-seq, HeLa cells were infected with VSV-G-pseudotyped HIV-1 NL4-3 ΔEnv EGFP
401 virus at a MOI of 0.6 and harvested at 0, 3, 6, and 12 hpi. Sequencing was performed with
402 biological replicates. Total RNA was extracted using TRIzol reagent (Invitrogen), according
403 to the manufacturer's instructions. An mRNA sequencing library was constructed using

404 Illumina adaptors harboring p5 and p7 sequences and Rd1 SP and Rd2 SP sequences.
405 Sequencing was performed using the HiSeq2500 system (Illumina).

406 **Luciferase assay**

407 HeLa cells infected with VSV-G-pseudotyped pNL4-3.Luc.R-E HIV-1 viruses were harvested
408 at 48 hpi, and luminescence was measured using the Dual-Luciferase Reporter Assay System
409 (Promega) according to the manufacturer's instructions. Briefly, 250 μl of passive lysis buffer
410 was used to lyse cells for each sample, 20 μl of the lysate was mixed with 100 μl of the
411 Luciferase Assay Reagent II, and the luminescence of firefly luciferase was measured using a
412 microplate luminometer (Berthold). The luminescence signal were normalized with total
413 protein content, measured by BCA assay.

414 **Quantitative real-time PCR (qPCR)**

415 For RT (reverse transcription)-qPCR, 1 μg of RNA was reverse-transcribed using the
416 ReverTra Ace qPCR RT Kit (TOYOBO) following the manufacturer's instructions. For
417 qPCR, DNA extracts were prepared using a DNA purification kit (Qiagen, 51106) according
418 to the manufacturer's instructions. Equivalent amounts of purified gDNA from each sample
419 were analyzed using qPCR. qPCR was performed using TOPreal qPCR PreMIX
420 (Enzynomics, RT500M). The reactions were performed in duplicate or triplicate for technical
421 replicates. PCR was performed using the iCycler iQ real-time PCR detection system (Bio-
422 Rad). All the primers used for qPCR are listed in S6 Table.

423 **DRIPc-seq library construction**

424 DRIP followed by library preparation, next-generation sequencing, and peak calling were
425 performed as described earlier (*28*). Briefly, the corresponding cells were harvested and their

426 gDNA was extracted. The extracted nucleic acids were fragmented using a restriction enzyme
427 cocktail with BsrB I (NEB, R0102S), HindIII (NEB, R0136L), Xba I (NEB, R0145L), and
428 EcoRI (NEB, R3101L) overnight at 37°C. Half of the fragmented nucleic acids were digested
429 with RNase H (New England Biolabs) overnight at 37°C to serve as a negative control. The
430 digested nucleic acids were cleaned using standard phenol-chloroform extraction and
431 resuspended in DNase/RNase-free water. DNA–RNA hybrids were immunoprecipitated from
432 total nucleic acids using mouse anti-DNA–RNA hybrid S9.6 (Kerafast, ENH001) DRIP
433 binding buffer and incubated overnight at 4°C. Dynabeads Protein A (Invitrogen, 10001D)
434 was used to pull down the DNA-antibody complexes by incubation for 4 h at 4°C. The
435 isolated complexes were washed twice with DRIP binding buffer before elution. For elution,
436 the isolated complexes were incubated in an elution buffer containing proteinase K for 45
437 min at 55 °C. Subsequently, DNA was purified using the standard phenol-chloroform extract
438 method and subjected to DNase I (Takara, 2270 B) treatment and reverse transcription for
439 DRIPc-seq library construction. DRIPc-seq was performed in biological replicates. S5 Table
440 shows details of the oligonucleotides used for DRIPc-seq library construction. DRIPc-seq
441 libraries were analyzed using 150 bp paired-end sequencing on a HiSeqX Illumina
442 instrument.

443 **Immunofluorescence microscopy**

444 For immunofluorescence assays of S9.6 nuclear signals, when indicated, the cells were pre-
445 extracted with cold 0.5% NP-40 for 3 min on ice. Cells were fixed with 100% ice-cold
446 methanol for 10 min on ice and then incubated with 100% ice-cold acetone for 1 min. The
447 slides were washed three times with 1× PBS and incubated with or without 60 U/mL RNase
448 H (M0297S, NEB) at 37°C for 36 h or left untreated. The slides were subsequently briefly

449 rinsed thrice with 2% BSA/0.05% Tween (in PBS) and incubated with mouse anti-DNA–
450 RNA hybrid S9.6 (Kerafast, ENH001; 1:100) and rabbit anti-nucleolin (Abcam, ab22758;
451 1:300) in 2% BSA/0.05% Tween (in PBS) for 4 h at 4°C. The slides were then washed three
452 times with 2% BSA/0.05% Tween (in PBS) and incubated with goat anti-rabbit AlexaFluor-
453 488-conjugated (Invitrogen, A-11008) and goat anti-mouse AlexaFluor-568-conjugated
454 (Molecular Probes, A11004) secondary antibodies (1:200) for 2 h at room temperature. The
455 slides were then washed three times with 2% BSA/0.05% Tween (in PBS) and mounted using
456 the ProLong Gold AntiFade reagent (Invitrogen). Images were obtained using an inverted
457 microscope Nikon Eclipse Ti2, equipped with a 1.45 numerical aperture, plan apochromat
458 lambda 100× oil objective, and an scientific complementary metal–oxide–semiconductor
459 camera (Photometrics prime 95 B 25 mm). For each field of view, images were obtained with
460 DAPI395, GFP488, and Alexa594 channels using the NIS-Elements software. For
461 quantification analysis, binary masks of nuclei and nucleoli were generated using the ROI
462 manager and auto local thresholding using the ImageJ software. The intensity of nuclear
463 signals for DNA–RNA hybrids and nucleolin was then quantified. The final DNA–RNA
464 hybrid signals in the nucleus were calculated by subtracting the nucleolin signals from the
465 DNA–RNA hybrid signals.

466 **pgR-rich and -poor cell line generation with piggyBac transposition**

467 We adapted and modified an elegantly designed episomal system that induces defined R-
468 loops with controlled transcription levels (*16*) for R-loop-forming or non-R-loop-forming
469 sequence subcloning into the piggyBac transposon vector. HeLa cells were seeded at a
470 density of 5×10^4 cells/ml in a 6-well plate. The next day, cells were transfected with 0.2 μg
471 of Super PiggyBac Transposase Expression Vector (System Biosciences, PB210PA-1) and

472 0.2, 1, or 2 μg of transposon vectors with appropriate "cargo" sub cloned using
473 Lipofectamine 3000 (Invitrogen) according to the manufacturer's instructions. After 3 days,
474 the cells were treated with 10 μg/ml blasticidin S (Gibco, A1113903) for selection. Cells with
475 positive integrants for more than 7 days were validated using immunoblotting or RT-qPCR
476 following treatment with DOX. Jurkat cells were seeded at a density of 8×10^5 cells/ml in a
477 6-well plate and transfected with 0.2 μg of transposase and 1 μg of corresponding transposon
478 vectors with Lipofectamine 3000, like HeLa cells. After 3 days, the cells were treated with 10
479 μg/ml blasticidin S (Gibco, A1113903) for selection. For each passage, cells were cushioned
480 onto Ficoll-Pacque (Cytiva, 17144002) to separate live cells from dead cell debris. The cells
481 over the cushion were washed with PBS and incubated in cell culture medium with 10 μg/ml
482 of blasticidin for further selection for at least 14 days. Cells with positive integrants were
483 validated by immunoblotting after treatment with DOX. Quantification of successfully
484 integrated piggyBac transposons was performed using a piggyBac qPCR copy number kit
485 (System Biosciences, PBC100A-1) according to the manufacturer's instructions.

486 **HIV-1 integration site sequencing library construction**

487 HIV-1 integration site sequencing library construction was performed as described earlier (*7,*
488 *9*). Summarily, HeLa cells were infected with VSV-G-pseudotyped HIV-1 NL4-3 ΔEnv
489 EGFP virus at a MOI of 0.6 and harvested 5 days post infection. gDNA was isolated using a
490 DNA purification kit (Qiagen, 51106), according to the manufacturer's instructions. gDNA
491 (10 μg) was digested overnight at 37°C with 100 U each of the restriction endonucleases
492 MseI (NEB, R0525L) and BglII (NEB, R0144L). Linker oligonucleotides, which were
493 compatible for ligation with the MseI-generated DNA ends, were ligated with gDNA
494 overnight at 12°C in reactions containing 1.5 μM ligated linker, 1 μg fragmented DNA, and

495 800 U T4 DNA ligase (NEB, M0202S). Viral LTR–host DNA junctions were amplified using
496 semi-nested PCR with a unique linker-specific primer and LTR primers. The second round of
497 PCR was carried out with primers binding to the LTR and the linkers for next-generation
498 sequencing. Two PCRs were performed in parallel for the first round of PCR and five PCRs
499 were performed in parallel for the second round of PCR to enhance library diversity. S7 Table
500 presents details of the oligonucleotides used for HIV-1 integration site sequencing library
501 construction. HIV-1 integration site sequencing was performed in biological replicates.
502 Integration site libraries were analyzed using 150 bp paired-end sequencing on a HiSeqX
503 Illumina instrument.

504 **Recombinant Sso7d-IN protein purification**

505 Sso7d-integrase active site mutant E152Q was expressed in Escherichia coli BL21-AI and
506 purified essentially as previously described (*50*). Briefly, Sso7d-IN (E152Q) expressed BL21-
507 AI cells were lysis in lysis buffer (20 mM HEPES pH 7.5, 2 mM 2-mercaptoethanol, 1 M
508 NaCl, 10% (w/v) glycerol, 20 mM imidazole, 1 mg RNase A, and 1000 U DNase I) and
509 purified by nickel affinity chromatography (Qiagen, 30210). Protein were first loaded on
510 HeparinHP column (GE Healthcare) equilibrated with equilibrated with 20 mM Tris, pH 8.0,
511 0.5 mM TCEP, 200 mM NaCl, 10% glycerol for anion exchange chromatography prior to the
512 size exclusion chromatography. Proteins were eluted with a linear gradient of NaCl from 200
513 mM to 1 M. Eluted fractions were pooled and then separated on a Superdex-200 PC 10/300
514 GL column (GE Healthcare) equilibrated with 20 mM Tris pH 8.0, 0.5 mM TCEP, 500 mM
515 NaCl and 6% (w/v) glycerol. The purified protein was concentrated to 0.6 mg/ml using an
516 Amicon centrifugal contentrator (EMD Millipore), flash-frozen in liquid nitrogen and stored
517 at -80°C.

518 **Electrophoretic mobility shift assay for R-loop binding of Sso7d-IN**

519 To test the binding affinity of Sso7d-tagged HIV-1 IN to different types of nucleic acid
520 substrates, we prepared R-loop, dsDNA, RNA-DNA hybrid with exposed ssDNA
521 (R:D+ssDNA), RNA-DNA hybrid (Hybrid), ssDNA, and ssRNA by annealing different
522 combinations of Cy3, Cy5 or non-labeled oligonucleotides following the previous protocol
523 (*51*). 10 nM of DNA substrate was incubated with Sso7d-IN at different concentrations in
524 assembly buffer (20 mM HEPES pH 7.5, 5 mM $CaCl_2$, 8 mM 2-mercaptoethanol, 4 uM
525 $ZnCl_2$, 100 mM NaCl, 25% (w/v) glycerol and 50 mM 3-(Benzyldimethylammonio)
526 propanesulfonate (NDSB-256)), for 1 h at 30°C then incubated for 15 min on ice. All the
527 reactants were run on 4.5% non-denaturing PAGE in 1× TBE and then Cy3 or Cy5
528 fluorescence signal was imaged by ChemiDoc MP imaging system (Bio-Rad). S8 Table
529 presents details of the oligonucleotide sequence used for EMSA.

530 **Co-immunoprecipitation of DNA–RNA hybrid**

531 DNA–RNA hybrid immunoprecipitation was performed as described earlier (*37*). Summarily,
532 non-crosslinked HeLa cells transfected with the pFlag-IN codon-optimized plasmid were
533 lysed in 85 mM KCl, 5 mM PIPES (pH 8.0), and 0.5% NP-40 for 10 min on ice, and then, the
534 lysates were centrifuged at 750 *g* for 5 min to pellet the nuclei. The pelleted nuclei were
535 resuspended in sodium deoxycholate, SDS, and sodium lauroyl sarcosinate in RSB buffer and
536 were sonicated for 10 min (Diagenode Bioruptor). Extracts were then diluted (1:4 in RSB + T
537 buffer) and subjected to immunoprecipitation with the S9.6 antibody overnight at 4°C.
538 Antibody-bound complexes were incubated with Protein A Dynabeads (Invitrogen) for 4 h at
539 4°C for immunoprecipitation. Normal mouse IgG antibodies (Santa Cruz, sc-2025) were used
540 as negative controls. RNase A (Thermo Scientific, EN0531) was added during

541 immunoprecipitation at 0.1 ng RNase A per µg gDNA. Beads were washed four times with
542 RSB + T; twice with RSB, and eluted either in 2× LDS (Novex, NP0007), 100 mM DTT for
543 10 min at 70°C (for western blot), or 1% SDS and 0.1 M $NaHCO_3$ for 30 min at room
544 temperature (for DNA–RNA hybrid dot blot).

545 For co-immunoprecipitation of DNA–RNA hybrids with RNase H treatment, gDNA
546 containing RNA–DNA hybrids was isolated from HeLa cells transfected with a pFlag-IN
547 codon-optimized plasmid using a QIAmp DNA Mini Kit (Qiagen, 51304). gDNA was
548 sonicated for 10 min (Diagenode Bioruptor) and then treated with 5.5 U RNase H (NEB,
549 M0297) per µg of DNA overnight at 37 °C. A fraction of gDNA was stored as "nucleic acid
550 input" for dot blot analysis. gDNA was cleaned using standard phenol-chloroform extraction,
551 resuspended in DNase/RNase-free water, enriched for DNA–RNA hybrids using
552 immunoprecipitation with the S9.6 antibody (overnight at 4°C), isolated with Protein A
553 Dynabeads (Invitrogen; 4 h at 4°C), washed thrice with RSB+T. The immunoprecipitated
554 complexes were incubated with nuclear extracts of HeLa cells transfected with the pFlag-IN
555 codon-optimized plasmid for 2 h at 4°C with diluted HeLa nuclear extracts. The cell lysate
556 containing proteins were pre-treated with 0.1 mg/ml RNase A (Thermo Scientific, EN0531)
557 for 1 h at 37°C to degrade all RNA–DNA hybrids, and the excess of RNase A was blocked by
558 adding 200 U of SUPERase in RNase inhibitor (Invitrogen, AM2694) for
559 immunoprecipitation. In addition, 100 µL fraction of diluted and RNase A pre-treated extracts
560 prior to immunoprecipitation was stored as "protein input" for western blotting. Beads were
561 washed four times with RSB + T; twice with RSB, and eluted either in 2× LDS (Novex,
562 NP0007), 100 mM DTT for 10 min at 70°C (for western blot), or 1% SDS, and 0.1 M
563 $NaHCO_3$ for 30 min at room temperature (for DNA–RNA hybrid dot blot).

564 **PLA**

565 For PLA, HeLa cells were grown on coverslips and infected with HIV-IN-EGFP virions. At 6
566 hpi, cells were pre-extracted with cold 0.5% NP-40 for 3 min on ice. The cells were fixed
567 with 4% paraformaldehyde in PBS for 15 min at 4 °C. The cells were then blocked with 1×
568 blocking solution (Merck, DUO92102) for 1 h at 37°C in a humidity chamber. After
569 blocking, cells were incubated with the following primary antibodies overnight at 4°C for
570 S9.6-HIV-1-IN_PLA: mouse anti-DNA–RNA hybrid S9.6 (1:250; Kerafast, ENH001) and
571 rabbit anti-GFP (1:500; Abcam, ab6556). The following day, after washing with once with
572 buffer A twice (Merck, DUO92102), cells were incubated with pre-mixed Duolink PLA plus
573 (anti-mouse) and PLA minus probes (anti-rabbit) antibodies for 1 h at 37°C. The subsequent
574 steps in the proximal ligation assay were performed using the Duolink PLA Fluorescence kit
575 (Sigma) according to the manufacturer's instructions. To obtain images, the mounted
576 specimens were visually scanned and representative images were acquired using a Zeiss LSM
577 710 laser scanning confocal microscope (Carl Zeiss). The number of intranuclear PLA puncta
578 was quantified using the ImageJ software. For each biological replicate and experiment, a
579 PLA with a single antibody was performed as a negative control under the same conditions.

580 **DRIPc-Seq data processing and peak calling**

581 DRIPc-seq reads were quality-controlled using FastQC v0.11.9 (*52*), and sequencing adapters
582 were trimmed using Trim Galore! v0.6.6 (*53*) based on Cutadapt v2.8 (*54*). Trimmed reads
583 were aligned to the hg38 reference genome using bwa v0.7.17-r1188 (*55*). Read
584 deduplication and peak calling were performed using MACS v2.2.7.1 (*56*). Because R-loops
585 appear as both narrow and broad peaks in DRIPc-seq read alignment owing to its variable
586 length, two independent "MACS2 callpeak" runs were performed for narrow and broad peak

587 calling. The narrow and broad peaks were merged using Bedtools v2.26.0 (*57*). To increase
588 the sensitivity of DRIPc-seq peak identification, peaks were called after pooling the two
589 biological replicates of the DRIPc-seq sequencing data for each condition.

590 **Consensus R-loop peak calling**

591 The R-loop peaks at 0, 3, 6 and 12 hpi were first merged using "bedtools merge" to create a
592 universal set of R-loop peaks across time points (n = 46542). Then, each of the universal R-
593 loop peaks was tested for overlap with the R-loop peaks for 0, 3, 6 and 12 hpi using "bedtools
594 intersect". In all, 9,190, 21,403, 33,544, and 9,941 peaks overlapped with 0, 3, 6, and 12 hpi
595 R-loop peaks, respectively. For CD4 cells, we identified a universal R-loop set consisting of
596 3,928 R-loops, and among them, 737, 722, 1,796 and 2,766 peaks overlapped with 0, 3, 6 and
597 12hpi R-loop peaks.

598 **HIV-1 integration site sequencing data processing**

599 Quality control of HIV-1 integration site-sequencing reads was performed using FastQC
600 v0.11.9. To discard primers and linkers specific for integration site-sequencing from reads,
601 we used Cutadapt v2.8 with the following option: "-u 49 -U 38 --minimum-length 36 --pair-
602 filter any --action trim -q0,0 –a linker -A
603 TGCTAGAGATTTTCCACACTGACTGGGTCTGAGGG -A GGGTCTGAGGG --no-indels
604 --overlap 12". This allowed the first position of the read alignment to directly represent the
605 genomic position of HIV-1 integration. Processed reads were aligned to the hg38 reference
606 genome using bwa v0.7.17-r1188, and integration sites were identified using an in-house
607 Python script. Genomic positions supported by more than five read alignments were regarded
608 as HIV-1 integration sites. For Jurkat cells, we adopted integration site sequencing data of
609 HIV-1 infected wild type Jurkat cells from SRR12322252 (*58*).

610 **Co-localization analysis of R-loops and integration sites**

611 Enrichment of integration sites near the R-loop peaks was tested using a randomized
612 permutation test. Randomized R-loop peaks were generated using "bedtools shuffle"
613 command, thus preserving the number and the length distribution of the R-loop peaks during
614 the randomization process. Similarly, integration sites were randomized using the "bedtools
615 shuffle" command. Randomization was performed 100 times. ENCODE blacklist regions
616 (*59*) were excluded while shuffling the R-loops and integration sites to exclude inaccessible
617 genomic regions from the analysis. For each of the observed (or randomized) integration
618 sites, the closest observed (or randomized) R-loop peak and the corresponding genomic
619 distance were identified using the "bedtools closest" command. The distribution of the
620 genomic distances was displayed to show the local enrichment of integration sites in terms of
621 the increased proportion of integration sites within the 30-kb window centered on R-loops
622 compared to their randomized counterparts.

623 **DNA plasmid construction and transfection**

624 R-loop-forming mAIRN and non-R-loop forming ECPF sequences were subcloned from
625 pSH26 and pSH36 plasmids, which were generously provided by Prof. Karlene A. Cimprich,
626 into the piggyBac transposon vector, where the tet operator sequences were located upstream
627 of the minimal CMV promoter. The pFlag-IN codon-optimized plasmid and pVpr-IN-EGFP
628 were kindly provided by Prof. A. Engelman and Prof. Anna Cereseto, respectively.
629 Lipofectamine 3000 (Invitrogen) transfection reagent was used for the transfection of all
630 plasmids into cells, according to the manufacturer's protocol.

631

632 **DNA–RNA hybrid dot blotting**

633 Total gDNA was extracted using the QIAmp DNA Mini Kit (Qiagen, 51304) according to the
634 manufacturer's instructions. gDNA (1.2 μg) was treated with 2 U RNase H (NEB, M2097)
635 per μg of gDNA for 4 h at 37°C, with half of the sample left untreated but denatured. Half of
636 the DNA sample was probed with S9.6 antibody (1:1000), and the other half was probed with
637 an anti-ssDNA antibody (MAB3034, Millipore, 1:10000).

638 **Immunoblotting**

639 Cells were lysed using RIPA buffer (50 mM Tris, 150 mM sodium chloride, 0.5% sodium
640 deoxycholate, 0.1% SDS, and 1.0% NP-40) supplemented with 10 μM leupeptin (Sigma-
641 Aldrich) and 1 mM phenylmethanesulfonyl fluoride (Sigma-Aldrich) and boiled at 98°C for
642 10 min with SDS sample buffer prior to SDS-PAGE. The primary antibodies used were
643 mouse monoclonal anti-FLAG M2 (Sigma, F3165), monoclonal mouse anti-HSC70 (Abcam,
644 ab2788), polyclonal rabbit anti-histone H3 (tri methyl K4) antibody (Abcam, ab8580),
645 monoclonal mouse anti- HIV-1 Integrase (Santa Cruz, sc-69721), rabbit anti-LaminA/C
646 antibody (Cell Signaling, 2032), and monoclonal mouse anti-Actin (Invitrogen, MA1-744).
647 All primary antibodies were used at a dilution of 1:1000 for western blotting. Peroxidase-
648 conjugated anti-mouse IgG (115-035-062) and anti-rabbit IgG (111-035-003; both Jackson
649 Laboratories) were used as secondary antibodies at 1:5000 dilution. Signals were detected
650 using the SuperSignal West Pico chemiluminescence kit (Thermo Fisher Scientific).

651 **RNA-seq data processing**

652 RNA-seq reads were quality-controlled and adapter-trimmed as in DRIPc-seq processing. To
653 quantify the expression levels of protein-coding genes, processed reads were aligned to the

654 hg38 reference genome with GENCODE v37 gene annotation (*60*) using STAR v2.7.3a (*61*).
655 Gene expression quantification was performed using RSEM v1.3.1. To quantify the
656 expression levels of transposable elements (TEs), we used TEtranscripts v2.2.1 (*62*).
657 Processed reads were first aligned to the hg38 reference genome using GENCODE v37 and
658 RepeatMasker TE annotation using STAR v2.7.3a. In this case, STAR options were modified
659 as follows to utilize multimapping reads in downstream analyses: "--outFilterMultimapNmax
660 100 --winAnchorMultimapNmax 100 --outMultimapperOrder random --runRNGseed 77 --
661 outSAMmultNmax 1 --outFilterType BySJout --alignSJoverhangMin 8 --
662 alignSJDBoverhangMin 1 --alignIntronMin 20 --alignIntronMax 1000000 --
663 alignMatesGapMax 1000000". Expression levels of TEs were quantified as read counts with
664 the "TEcount" command.

665 **Genome annotations**

666 All bioinformatic analyses were performed using the hg38 reference genome and GENCODE
667 v37 gene annotation. Promoters were defined as a 2-kb region centered at the transcription
668 start sites of the APPRIS principal isoform of protein-coding genes. TTS regions were
669 defined as the 2-kb region centered at the 3′ terminals of protein-coding transcripts. CpG
670 island annotations were downloaded from the UCSC table browser. CpG shores were defined
671 as 2-kb regions flanking CpG islands, excluding the regions overlapping with CpG islands.
672 Similarly, CpG shelves were defined as 2-kb regions flanking the stretch of CpG islands and
673 shores while excluding the regions overlapping with CpG islands and shores. Annotations for
674 LINE, SINE, and LTR were extracted from the RepeatMasker track in the UCSC table
675 browser.

676

677 **Identification of viral sequencing reads in DRIPc-seq**

678 To identify sequencing reads originating from the viral genome, we aligned DRIPc-seq reads
679 to a composite reference genome consisting of the human and HIV1 genome (Genbank
680 accession number: AF324493.2) and computed the proportion of the reads mapped to HIV1
681 genome.

682 **Code availability**

683 Bioinformatics pipelines and scripts used in this study are accessible from
684 https://github.com/dohlee/hiv1-rloop.

685

Acknowledgements

We are grateful to Prof. Karlene A. Cimprich (Standford University) for providing the pSH26 and pSH36 plasmids, Prof. A. Engelman (Harvard Medical School) for providing pFlag-IN codon optimized plasmid and Prof. Anna Cereseto (University of Trento) for providing pVpr-IN-EGFP. The NL4-3 ΔEnv EGFP and pNL4-3.Luc.R-E- viral plasmids were obtained through the NIH HIV Reagent Program, Division of AIDS, NIAID, NIH. We thank Dr. Sungchul Kim (IBS center for RNA Research) and Seongjin An (Korea University) for their technical support in recombinant protein purification.

Author contributions

K.P. and K.A. designed experiments. K.P., J.J and S.L. performed experiments. D.L. performed the bioinformatical and statistical analyses. K.P., D.L., K.A. and S.K. analyzed the data. K.P., D.L., and K.A. wrote the manuscript.

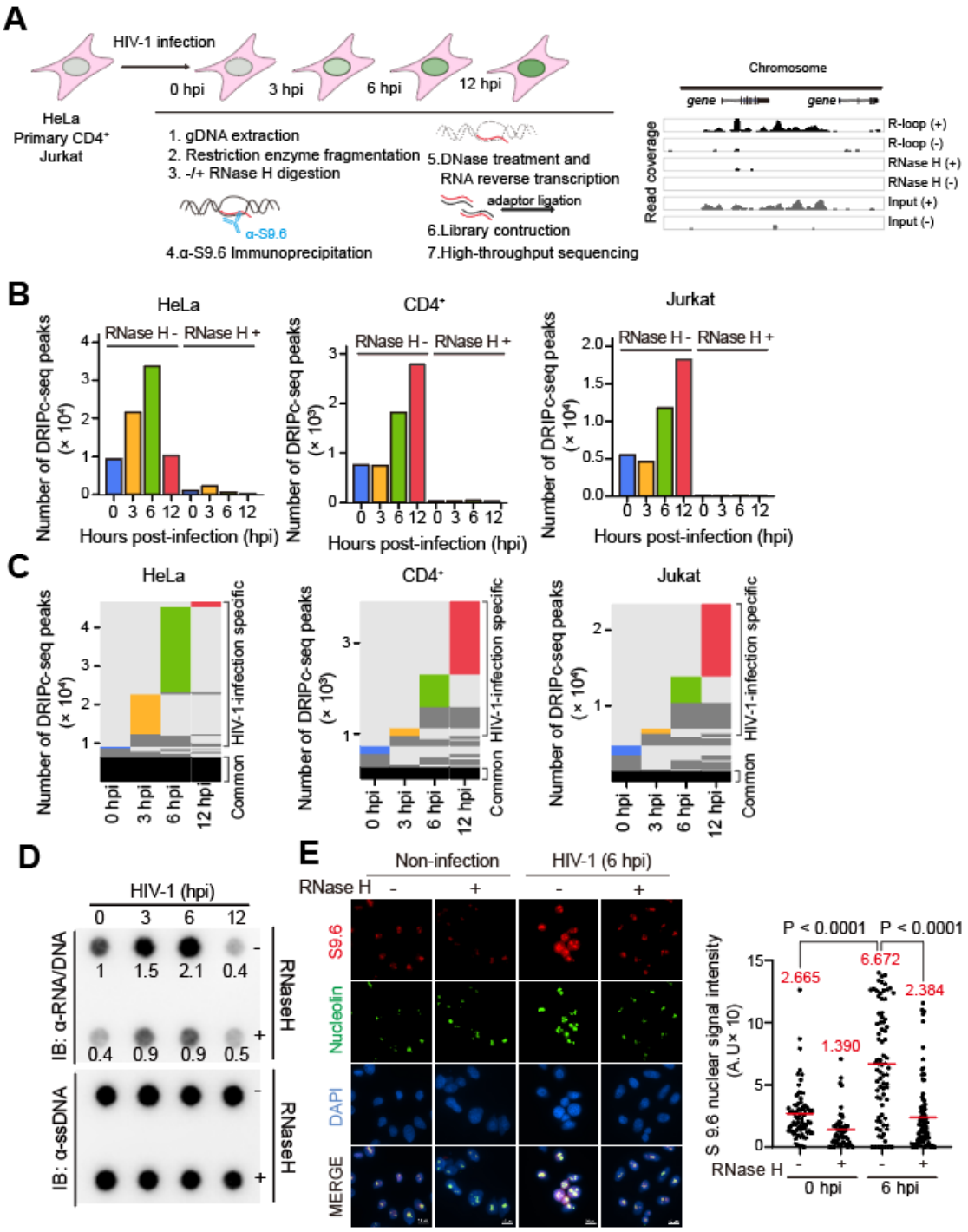
Funding

This work was supported by the Institute for Basic Science of the Ministry of Science Grant (IBS-R008-D1) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (NRF-2020R1A2C3011298) (to K. A.) and (NRF-2020R1A5A1018081) (to K.A.). The funders had no role in the study design, data collection, analysis, decision to publish, or preparation of the manuscript.

Competing interests

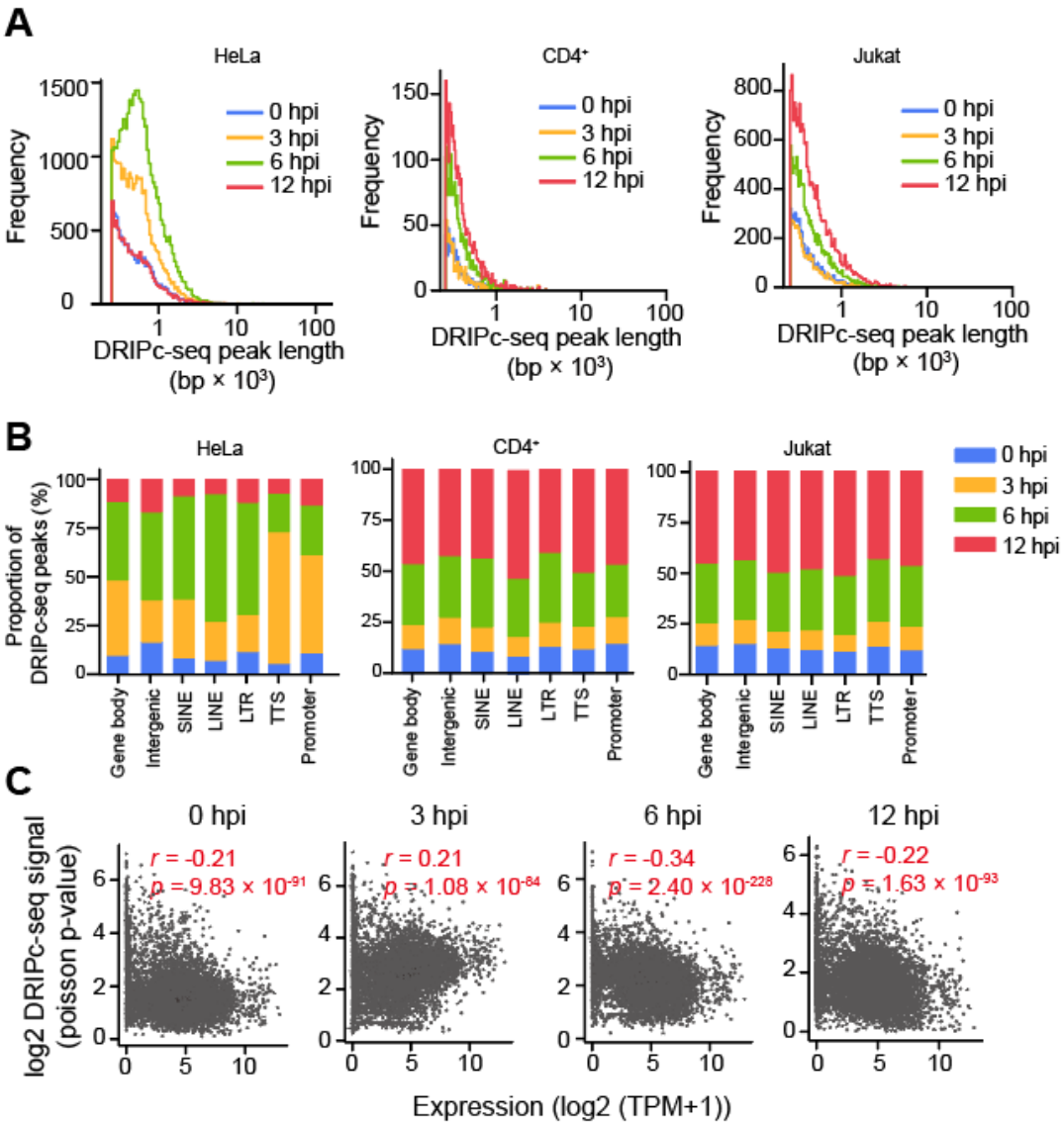
The authors have declared that no competing interests exist.

Figures


A
HIV-1 infection
0 hpi
3 hpi
6 hpi
12 hpi
HeLa
Primary CD4+
Jurkat
1. gDNA extraction
2. Restriction enzyme fragmentation
3. -/+ RNase H digestion
4.α-S9.6 Immunoprecipitation
5.DNase treatment and RNA reverse transcription
adaptor ligation
6.Library contruction
7.High-throughput sequencing
α-S9.6
Chromosome
gene
Read coverage
R-loop (+)
R-loop (-)
RNase H (+)
RNase H (-)
Input (+)
Input (-)
B
HeLa
CD4+
Jurkat
RNase H -
RNase H +
Number of DRIPc-seq peaks (× 10⁴)
Number of DRIPc-seq peaks (× 10³)
Hours post-infection (hpi)
C
HIV-1-infection specific
Common
D
HIV-1 (hpi)
IB: α-RNA/DNA
IB: α-ssDNA
RNaseH
E
Non-infection
HIV-1 (6 hpi)
RNase H
S9.6
Nucleolin
DAPI
MERGE
S 9.6 nuclear signal intensity (A.U× 10)
P < 0.0001
P < 0.0001
2.665
1.390
6.672
2.384
0 hpi
6 hpi

710 **Fig 1. HIV-1 infection induces genomic R-loop accumulation in cells at early post-infection.** (**A**) Summary
711 of experimental design for DRIPc-seq in HeLa cells, primary $CD4^+$ T cells and Jurkat cells infected with HIV-1.
712 (**B**) Bar graphs indicating DRIPc-seq peak counts for HIV-1-infected HeLa cells, primary $CD4^+$ T cells and
713 Jurkat cells harvested at the indicated hours post infection (hpi). Pre-immunoprecipitated samples were
714 untreated (−) or treated (+) with RNase H, as indicated. Each bar corresponds to pooled datasets from two
715 biologically independent experiments. (**C**) All genomic loci overlapping a DRIPc-seq peak from HIV-1 infected
716 HeLa cells, primary $CD4^+$ T cells and Jurkat cells in at least one sample are stacked vertically; the position of
717 each peak in a stack is constant horizontally across samples. Each hpi occupies a vertical bar, as indicated. Each
718 bar corresponds to pooled datasets from two biologically independent experiments. Common peaks for all
719 samples are represented in black, and in dark gray for those common for at least two samples. The lack of a
720 DRIP signal over a given peak in any sample is shown in light gray. The sample-unique peaks are colored blue,
721 yellow, green, and red at 0, 3, 6, and 12 hpi, respectively. (**D**) Dot blot analysis of the R-loop in gDNA extracts
722 from HIV-1 infected HeLa cells with MOI of 0.6 harvested at the indicated hpi. gDNAs were probed with anti-
723 S9.6. gDNA extracts were incubated with or without RNase H in vitro before membrane loading (anti-
724 RNA/DNA signal). Fold-induction was normalized to the value of harvested cells at 0 hpi by quantifying the dot
725 intensity of the blots and calculating the ratios of the S9.6 signal to the total amount of gDNA (anti-ssDNA
726 signal). (**E**) Representative images of the immunofluorescence assay of S9.6 nuclear signals in HIV-1 infected
727 HeLa cells with MOI of 0.6 harvested at 6 hpi. The cells were pre-extracted of cytoplasm and co-stained with
728 anti-S9.6 (red), anti-nucleolin antibodies (green), and DAPI (blue). The cells were incubated with or without
729 RNase H in vitro before staining with anti-S9.6 antibodies, as indicated. Quantification of S9.6 signal intensity
730 per nucleus after nucleolar signal subtraction for the immunofluorescence assay. The mean value for each data
731 point is indicated by the red line. Statistical significance was assessed using one-way ANOVA ($n > 53$).

732



733

734 **Fig. 2. HIV-1-induced R-loops are enriched at both transcriptionally active and silent regions.** (**A**)
735 Distribution of DRIPc-seq peak lengths for HIV-1-infected HeLa cells, primary CD4$^+$ T cells and Jurkat cells
736 harvested at the indicated time points (blue, 0 hpi; yellow, 3 hpi; green, 6 hpi; red, 12 hpi). (**B**) Stacked bar graphs
737 indicating the proportion of DRIPc-seq peaks mapped for HIV-1-infected HeLa cells, primary CD4$^+$ T cells and
738 Jurkat cells harvested at the indicated hpi over different genomic features. (**C**) Correlation between gene
739 expression and DRIPc-seq signals of HIV-1-infected HeLa cells with MOI of 0.6 harvested at the indicated hpi.
740 Statistical significance was assessed using Pearson's r and p-values.

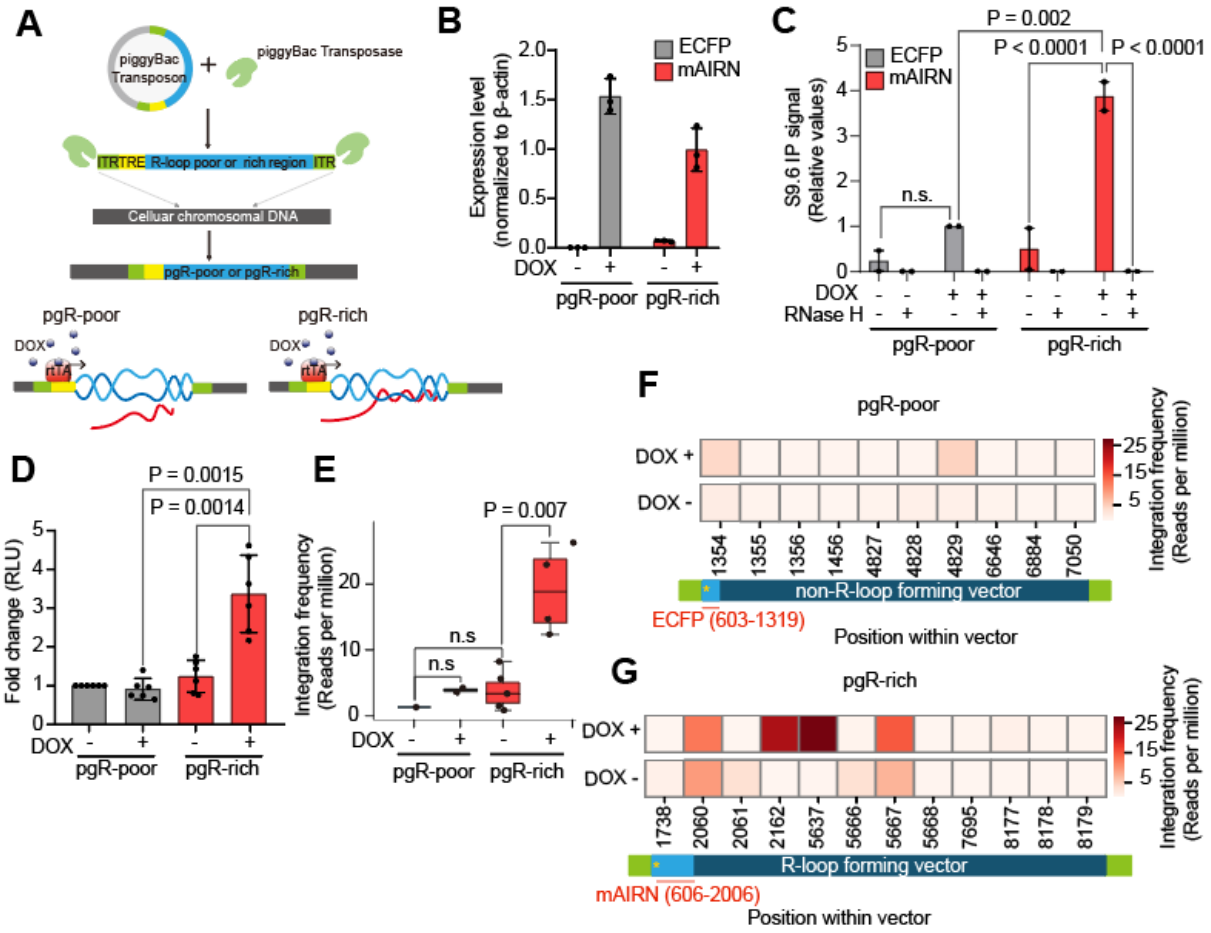


Fig. 3. R-loop inducible cell line model directly addresses R-loop-mediated HIV-1 integration site selection. (**A**) Summary of the experimental design for R-loop inducible cell lines, pgR-poor and pgR-rich. (**B**) Gene expression of ECFP (gray) and mAIRN (red), as measured using RT-qPCR in pgR-poor or pgR-rich cells. Where indicated, the cells were incubated with 1 µg/ml DOX for 24 h. Gene expression was normalized relative to *β-actin*. Data are presented as the mean ± SEM, n = 3. (**C**) DRIP-qPCR using the anti-S9.6 antibody against ECFP and mAIRN in pgR-poor or pgR-rich cells. Where indicated, the cells were incubated with 1 µg/ml DOX for 24 h. Pre-immunoprecipitated samples were untreated or treated with RNase H as indicated. Values are relative to those of DOX-treated (+) RNase H-untreated (−) pgR-poor cells. Data are presented as the mean ± SEM; statistical significance was assessed using two-way ANOVA (n = 2). (**D**) Bar graphs indicate luciferase activity at 48 hpi in pgR-poor or prR-rich cells infected with 100ng/p24 capsid antigen of luciferase reporter HIV-1 virus per 1×10^5 cells/mL. Data are presented as the mean ± SEM; P values were calculated using one-way ANOVA (n = 6). (**E**) Box graph indicating the quantified HIV-1 integration site sequencing read count across pgR-poor and pgR-rich transposon sequences in untreated (–) or DOX-treated (+) pgR-poor or pgR-rich cell line infected with 100ng/p24

755 capsid antigen of luciferase reporter HIV-1 virus per 1×10^5 cells/mL. Each bar corresponds to pooled datasets
756 from three biologically independent experiments (n =3). In each boxplot, the centerline denotes the median, the
757 upper and lower box limits denote the upper and lower quartiles, and the whiskers denote the 1.5 × interquartile
758 range. Statistical significance was assessed using a two-sided Mann–Whitney U test. (**F** and **G**) Heat maps
759 representing HIV-1 integration frequency across pgR-poor (**F**) or pgR-rich (**G**) transposon sequence in untreated
760 (-) or DOX-treated (+) pgR-poor (**F**) or pgR-rich (**G**) cell line. Each rectangular box corresponds to the pooled
761 integration frequency from three biologically independent experiments (n =3) at the indicated position within
762 pgR-poor (**F**) or pgR-rich (**G**) transposon vector. Each light blue box represents actual position of R-loop forming
763 or non-R-loop forming sequence (ECFP or mAIRN) and the yellows stars indicate TRE promoter position within
764 vector.

765

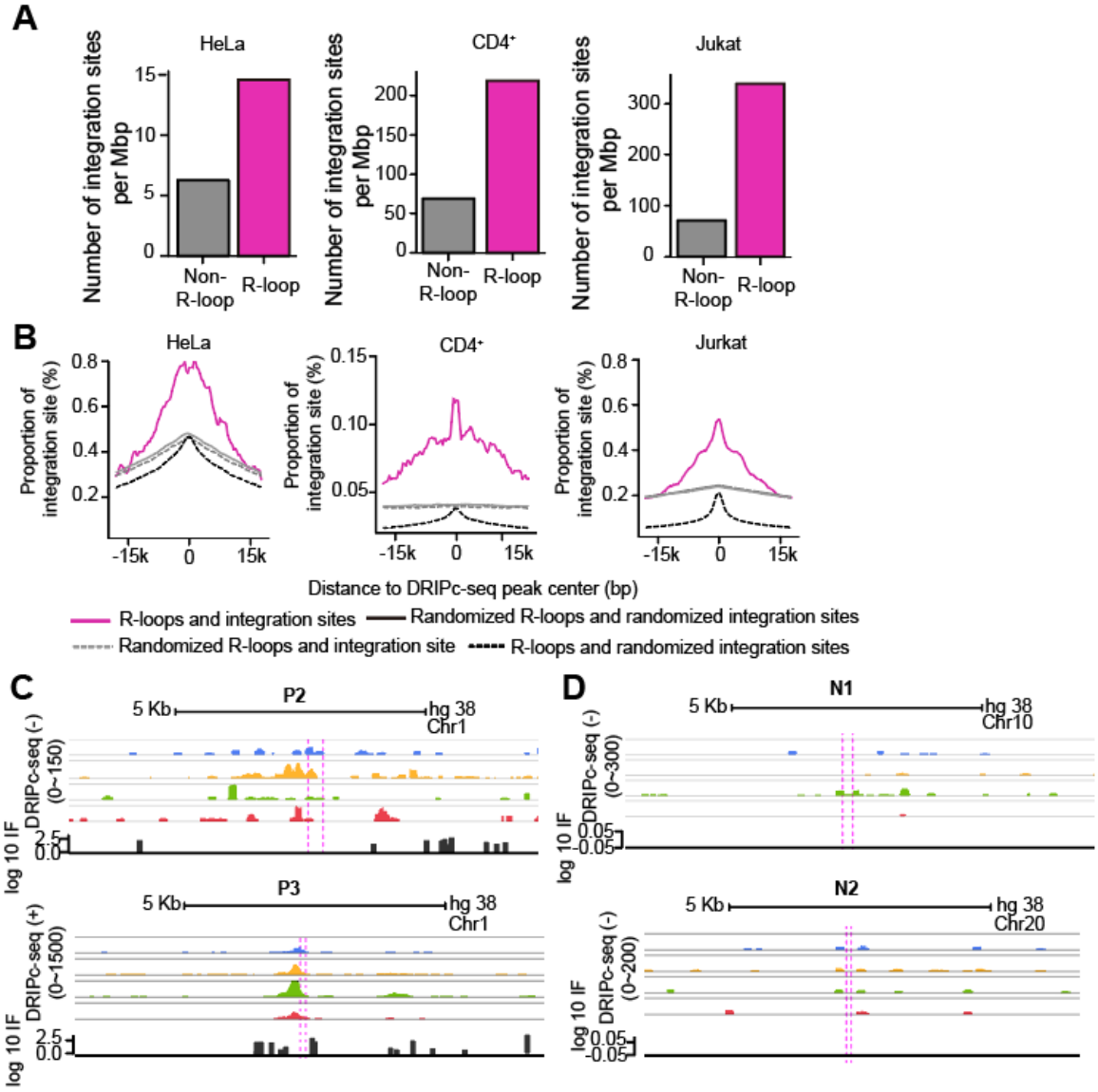


Fig. 4. HIV-1 targets host genomic R-loop for its viral cDNA integration. (**A**) Bar graphs showing quantified number of HIV-1 integration sites per Mb pairs in total regions of 30-kb windows centered on DRIPc-seq peaks from HIV-1 infected HeLa cells, primary CD4$^+$ T cells and Jurkat cells (magenta) or non-R-loop region in the cellular genome (gray). (**B**) Proportion of integration sites within the 30-kb windows centered on R-loops (magenta solid lines) or randomized R-loops (gray dotted lines). Control comparisons between randomized integration sites with R-loops and randomized R-loops are indicated by black dotted lines and gray solid lines, respectively. (**C** and **D**) Superimpositions of HIV-1-induced R-loop positive chromatin regions, P2 and P3 (**C**), and HIV-1-induced R-loop negative chromatin regions, N1 and N2 (**D**), on DRIPc-seq (blue, 0 hpi; yellow, 3 hpi; green, 6 hpi; red, 12 hpi) and HIV-1 integration frequency (IF, black).

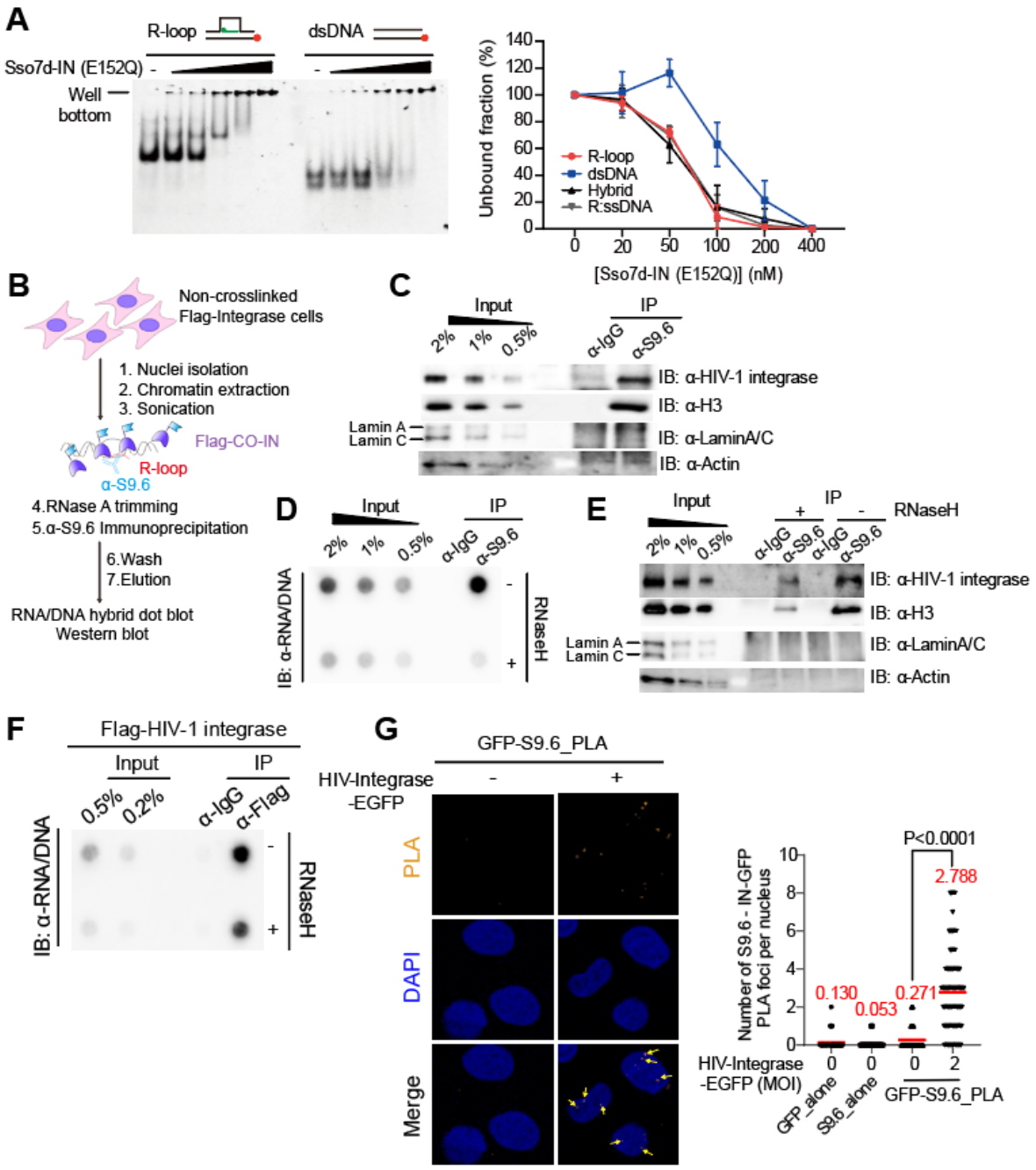


Fig. 5. HIV-1 integrase proteins directly bind to host genomic R-loops. (**A**) Representative gel images for EMSA of Sso7d-tagged HIV-1-integrase (E152Q) with R-loop and dsDNA, 10 nM nucleic acid substrate was incubated with Sso7d-tagged HIV-1-integrase (E152Q) at 0 nM, 20 nM, 50 nM, 100 nM, 200 nM, and 400 nM (left). Unbound fraction were quantified for EMSA of Sso7d-tagged HIV-1-integrase (E152Q) with different types substrates (R-loop, dsDNA, R-loop, R:D+ssDNA and Hybrid). Data are presented as the mean ± SEM, n = 3 (right). (**B**) Summary of the experimental design for R-loop immunoprecipitation using S9.6 antibody in FLAG-

784 tagged HIV-1 integrase protein-expressing HeLa cells. (**C**) Western blotting for HIV-1 integrase protein, H3, and
785 LaminA/C of DNA–RNA hybrid immunoprecipitation using the S9.6 antibody. (**D**) **and** (**E**) HeLa gDNA input
786 was either untreated (–) or treated (+) with RNase H before enrichment for DNA–RNA hybrids using the S9.6
787 antibody. gDNA–RNA hybrids were incubated with nuclear extracts depleted of DNA–RNA hybrids with RNase
788 A followed by S9.6 immunoprecipitation. DNA–RNA hybrid dot blot (**D**) and western blot of DNA–RNA hybrid
789 immunoprecipitation, probed with the indicated antibodies (**E**). (**F**) DNA–RNA hybrid dot blot of FLAG antibody-
790 immunoprecipitated nucleic acid extracts. Where indicated, nucleic acid extracts were untreated (–) or treated (+)
791 with RNase H before probing with the S9.6 antibodies. (**G**) Representative images of the proximity-ligation assay
792 (PLA) between GFP and S9.6 antibodies in HIV-IN-EGFP virion-infected HeLa cells at 6 hpi. Cells were
793 subjected to PLA (orange) and co-stained with DAPI (blue). PLA puncta in the nucleus are indicated by the yellow
794 arrows. Quantification analysis of number of PLA foci per nucleus (left). GFP_alone and S9.6_alone were used
795 as single-antibody controls from HIV-IN-EGFP virion-infected HeLa cells (right). The mean value for each data
796 point is indicated by the red line. P value was calculated using a two-tailed unpaired *t*-test ($n > 50$).

797

798 **Supplemental Information**

799 Materials and Methods

800 S1 Fig. Primary $CD4^+$ T cells sorting strategies and GFP-HIV-1 infection

801 S2 Fig. Genome browser screenshot over the HIV-1-induced R-loop forming positive or
802 negative genomic regions

803 S3 Fig. Host cellular R-loop induction by HIV-1 infection is host-genome specific.

804 S4 Fig. R-loop induction by HIV-1 infection does not follow transcriptome changes in HeLa
805 cells

806 S5 Fig. PiggyBac transposon-transposase insertion of R-loop forming and non-R-loop
807 forming sequences in HeLa cells

808 S6 Fig. HIV-1 integrase proteins directly binds to host genomic R-loops

809 S1 Table. Chromosomal position and DRIPc-seq signal for referenced R-loop-positive and –
810 negative regions in HIV-1 infected HeLa cells

811 S2 Table. Chromosomal position and DRIPc-seq signal for referenced R-loop-positive and –
812 negative regions in HIV-1 infected primary $CD4^+$ T cells

813 S3 Table. Chromosomal position and DRIPc-seq signal for referenced R-loop-positive and –
814 negative regions in HIV-1 infected Jurkat cells

815 S4 Table. RNA-seq analysis of relative gene expression levels of P1-3 and N1,2 R-loop
816 regions

817 S5 Table. Oligonucleotides used for DRIPc-seq library construction

818 S6 Table. Primers used for qPCR

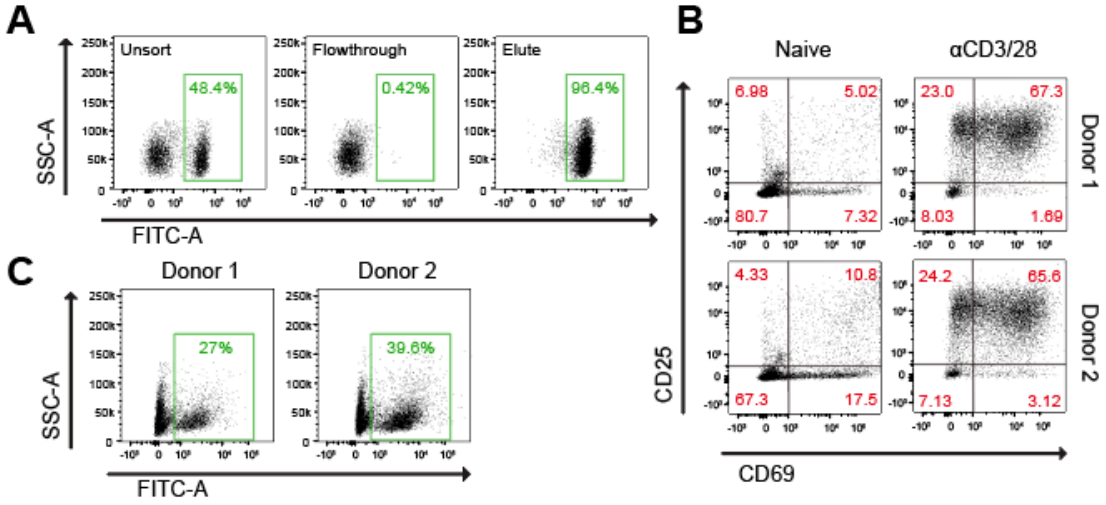
819 S7 Table. Oligonucleotides used for HIV-1 integration site sequencing library construct

820 S8 Table. Oligonucleotides used for electrophoretic mobility shift assay substrate preparation

821 S9 Table. Accession numbers and data sources.

822

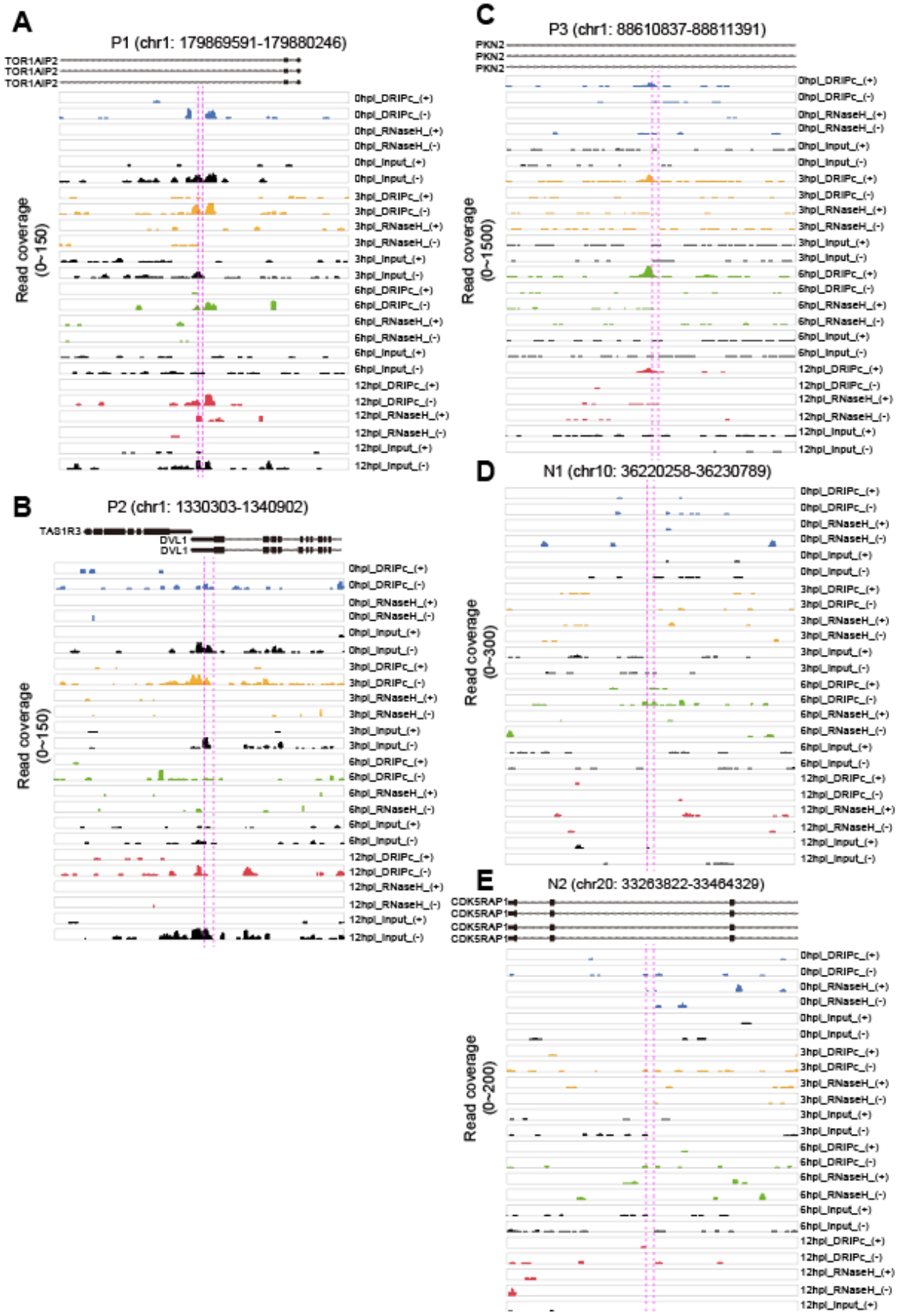
823 # Supplementary figures



824

825 **S1 Fig. Primary CD4$^+$ T cells sorting strategies and GFP-HIV-1 infection.** (**A**) Gating strategy used to
826 determine the efficiency of CD4$^+$ T cells sorting from human PBMC. Pre-sorted PBMCs were staining with
827 FITC-conjugated anti-CD4 and subjected for positive CD4+ T cell sorting. The percentages of FITC stained cell
828 population at each step of cell sorting are as indicated. (**B**) Gating strategy used to determine non-activated
829 (Naïve) and activated cells (αCD3/28) with two markers, CD25 (FITC) and CD69 (APC), for each donor (upper
830 panels, Donor 1; lower panels, Donor 2). (**C**) Gating strategy used to determine HIV-1-infectivity of CD4+ T
831 cells from each donor infected with GFP reporter HIV-1 virus at 48 hpi. The percentages of GFP positive cell
832 population at are as indicated.

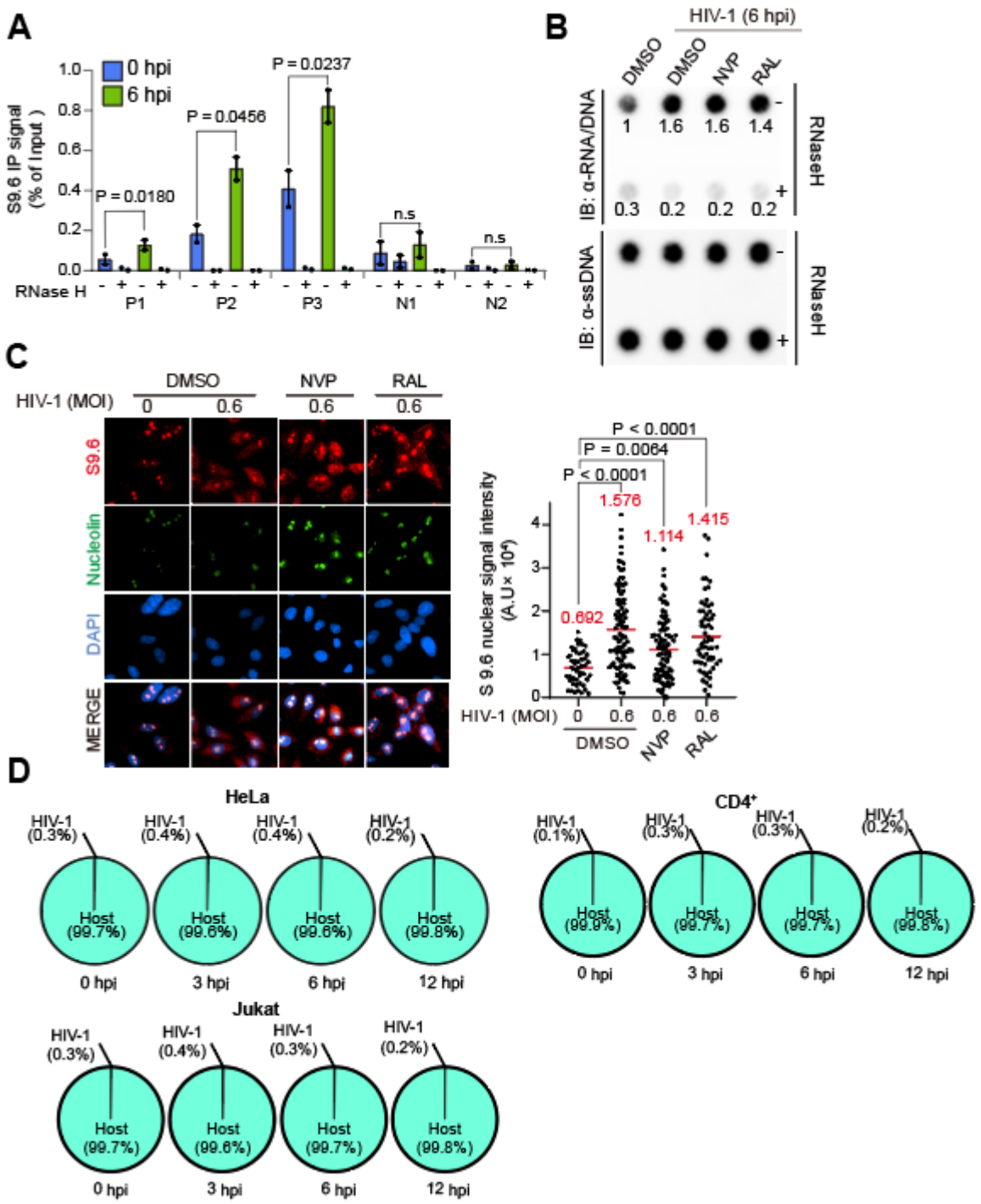
833


A
P1 (chr1: 179869591-179880246)
TOR1AIP2
TOR1AIP2
TOR1AIP2
Read coverage (0~150)
0hpI_DRIPc_(+)
0hpI_DRIPc_(-)
0hpI_RNaseH_(+)
0hpI_RNaseH_(-)
0hpI_Input_(+)
0hpI_Input_(-)
3hpI_DRIPc_(+)
3hpI_DRIPc_(-)
3hpI_RNaseH_(+)
3hpI_RNaseH_(-)
3hpI_Input_(+)
3hpI_Input_(-)
6hpI_DRIPc_(+)
6hpI_DRIPc_(-)
6hpI_RNaseH_(+)
6hpI_RNaseH_(-)
6hpI_Input_(+)
6hpI_Input_(-)
12hpI_DRIPc_(+)
12hpI_DRIPc_(-)
12hpI_RNaseH_(+)
12hpI_RNaseH_(-)
12hpI_Input_(+)
12hpI_Input_(-)
B
P2 (chr1: 1330303-1340902)
TAS1R3
DVL1
DVL1
Read coverage (0~150)
C
P3 (chr1: 88610837-88811391)
PKN2
PKN2
PKN2
Read coverage (0~1500)
D
N1 (chr10: 36220258-36230789)
Read coverage (0~300)
E
N2 (chr20: 33263822-33464329)
CDK5RAP1
CDK5RAP1
CDK5RAP1
CDK5RAP1
Read coverage (0~200)

834

835 **S2 Fig. Genome browser screenshot over the HIV-1-induced R-loop forming positive or negative genomic**
836 **regions. (A-C),** Genome browser screenshot over the P1 (**A**), P2 (**B**), and P3 (**C**) HIV-1 induced R-loop-positive
837 chromosomal regions showing result from DRIPc-seq in HIV-1-infected HeLa cells (blue, 0 hpi; yellow, 3 hpi;
838 green, 6 hpi; red, 12 hpi; black, input signals for each indicated sample) on plus (+) or minus (-) DNA strand.
839 Magenta dotted lines represent primer binding sites in qPCR following DRIP. **(D** and **E),** Genome browser
840 screenshot over the N1 (**D**), and N2 (**E**) HIV-1 induced R-loop-negative chromosomal regions showing result
841 from DRIPc-seq in HIV-1-infected HeLa cells (blue, 0 hpi; yellow, 3 hpi; green, 6 hpi; red, 12 hpi; black, input
842 signals for each indicated sample) on plus (+) or minus (-) DNA strand. Magenta dotted lines represent primer
843 binding sites in qPCR following DRIP.

844

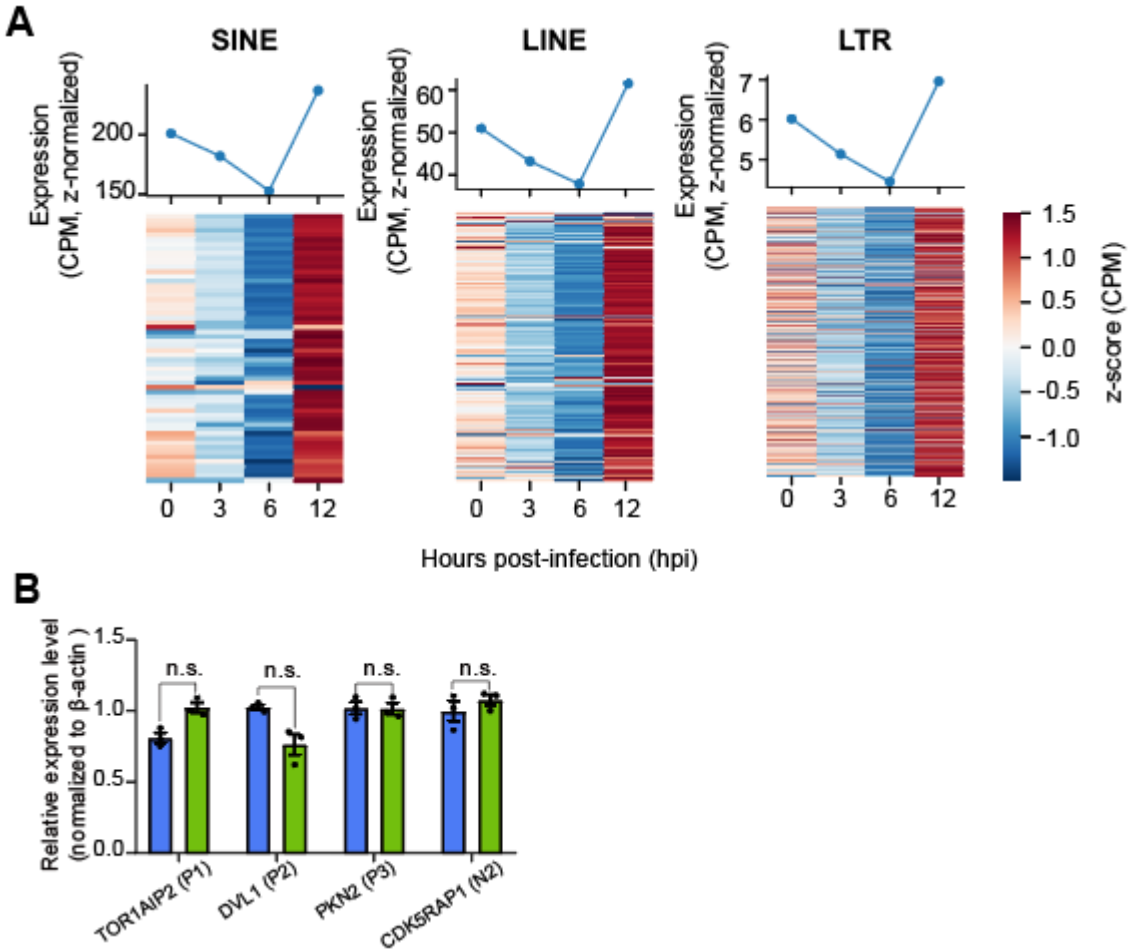


845

846 **S3 Fig. Host cellular R-loop induction by HIV-1 infection is host-genome specific. (A)** DRIP-qPCR using
847 the anti-S9.6 antibody at P1, P2, P3, N1, and N2 in HIV-1-infected cells with MOI of 0.6 harvested at the
848 indicated hpi (blue, 0 hpi; green, 6 hpi). Pre-immunoprecipitated materials were untreated (−) or treated (+) with
849 RNase H, as indicated. Data are presented as the mean ± SEM; P-values were calculated using one-way ANOVA

850 (n = 2). (**B**) Dot blot analysis of the R-loop in gDNA extracts from HIV-1 infected HeLa cells with MOI of 0.6
851 harvested at 6hpi. The cells were treated with DMSO, 10uM of Nevirapine (NVP), or 10uM of Raltegravir
852 (RAL) for 24 h before infection, as indicated. gDNAs were probed with anti-S9.6. gDNA extracts were
853 incubated with or without RNase H in vitro before membrane loading (anti-RNA/DNA signal). Fold-induction
854 was normalized to the value of harvested cells at 0 hpi by quantifying the dot intensity of the blots and
855 calculating the ratios of the S9.6 signal to the total amount of gDNA (anti-ssDNA signal). (**C**) Representative
856 images of the immunofluorescence assay of S9.6 nuclear signals in HIV-1 infected HeLa cells with MOI of 0.6
857 at 6 hpi. The cells were pre-extracted of cytoplasm and co-stained with anti-S9.6 (red), anti-nucleolin antibodies
858 (green), and DAPI (blue). The cells were treated with DMSO, 10uM of Nevirapine (NVP), or 10uM of
859 Raltegravir (RAL) for 24 h before infection, as indicated. Quantification of S9.6 signal intensity per nucleus
860 after nucleolar signal subtraction for the immunofluorescence assay. The mean value for each data point is
861 indicated by the red line. Statistical significance was assessed using one-way ANOVA (n >51). (**D**) Pie graphs
862 indicating the percentage of DRIPc-seq reads aligned to host cellular genome (aquamarine) or to HIV-1 viral
863 genome (gray), out of the total consensus DRIPc-seq peaks from HIV-infected HeLa cells, primary $CD4^+$ T cells
864 and Jurkat cells.

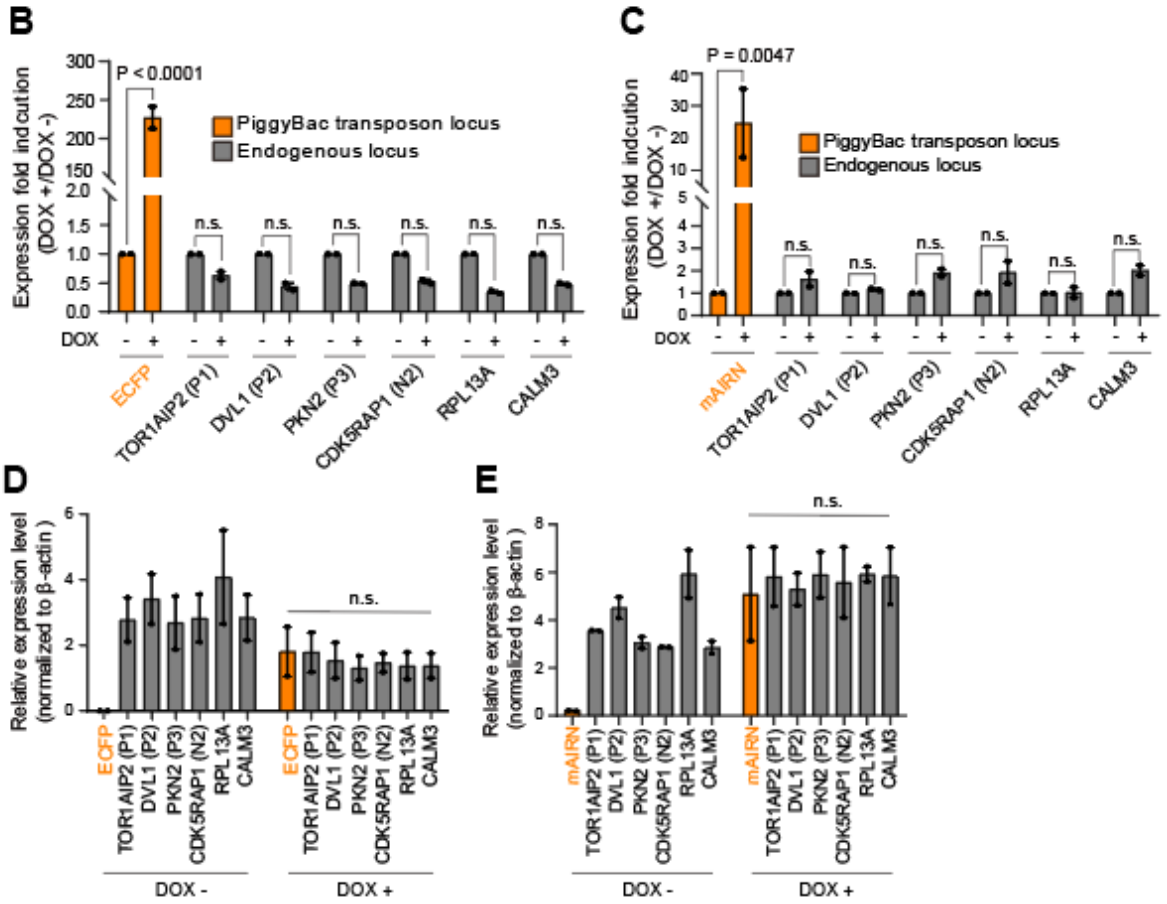
865



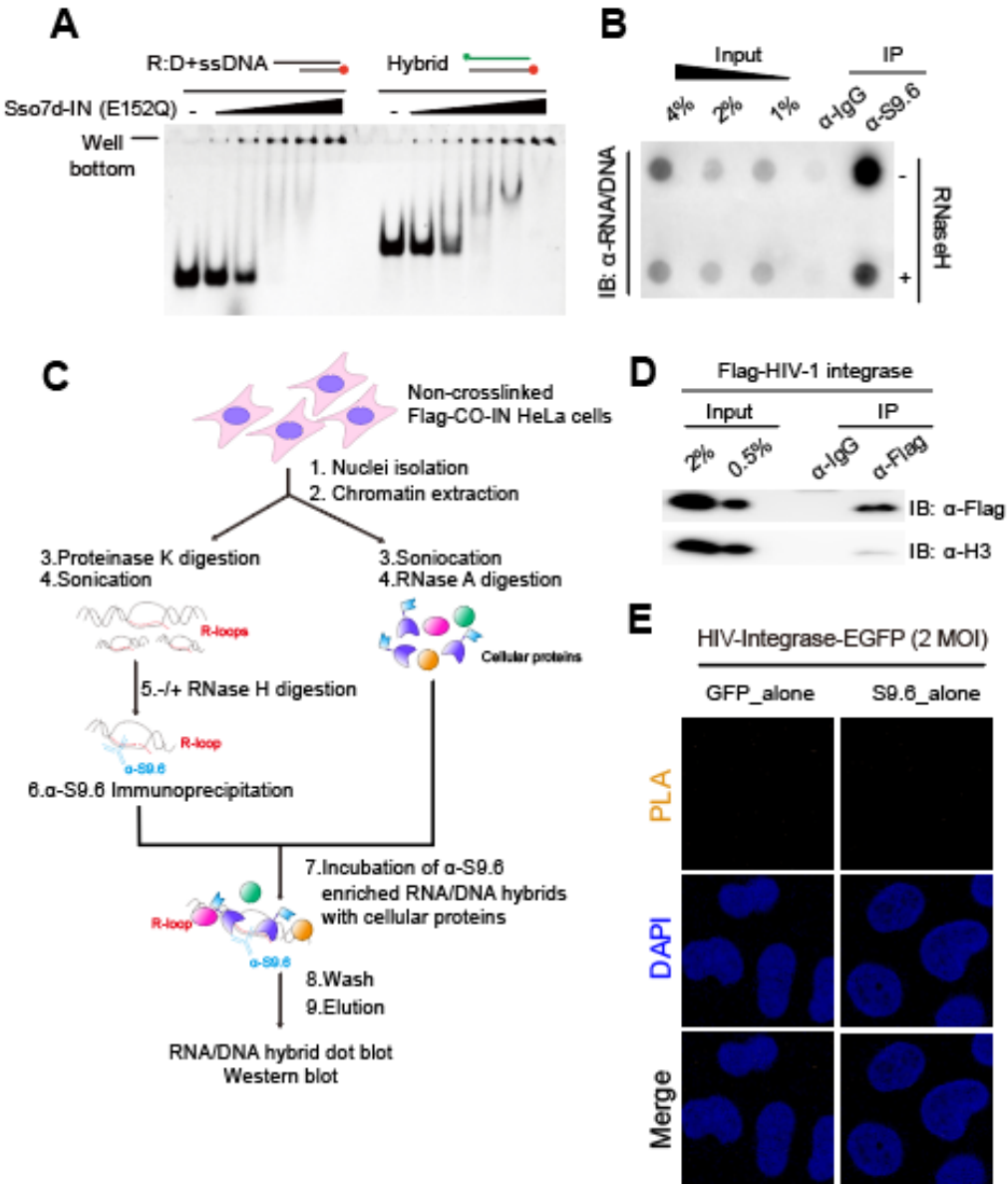
S4 Fig. R-loop induction by HIV-1 infection does not follow transcriptome changes in HeLa cells. (**A**) Line graphs and heat maps representing expression levels of indicated repetitive elements (SINE, right; LINE, middle; LTR, left) at the indicated hpi of HIV-1 in HeLa cells. Data are presented as the mean expression levels of two biologically independent experiments. (**B**) Indicated gene expression as measured by RT-qPCR in 0 or 6 hpi harvested HIV-1-infected HeLa cells. Data represent mean ± SEM, n = 3, P values were calculated according to two-tailed Student's t-test. P > 0.05; n.s, not significant.

A

Cell line	Pbcopy avg Ct	UCR1 avg Ct	ΔΔCt: 2^-(Pbcopy-UCR1)	Copy #per Genome (ΔΔCt/2)
pgR-poor	20.15	25.91	54.19	27
pgR-rich	20.12	25.63	45.57	23



S5 Fig. Regulation of cellular R-loops by RNase H1 expression, or by transposon-transposase insertion of R-loop forming and non-R-loop forming sequences in HeLa cells. (**A**) Copy number of piggyBac transposon inserts in each cell line constructed by transfecting the transposon vector and transposase-expressing vector. Cell lines used for further experiments are shaded gray (pgR-poor) or red (pgR-rich). (**B** and **C**) Fold induction of gene expression for the indicated genes as measured by RT-qPCR. Fold induction were calculated by dividing the gene expression level of DOX-treated (+) by that of DOX-untreated (-) in pgR-poor cells (**B**) or pgR-rich cells (**C**). Data represent mean ± SEM, n = 2, P values were calculated according to two-way ANOVA. P > 0.05; n.s, not significant. (**D** and **E**) Relative gene expression of the indicated genes as measured by RT-qPCR in DOX-treated (+) or DOX-untreated (-) pgR-poor cells (**D**) or pgR-rich cells (**E**). Data represent mean ± SEM, n = 2, P values were calculated according to two-way ANOVA. P > 0.05; n.s, not significant.



886

887 **S6 Fig. HIV-1 integrase proteins directly binds to host genomic R-loops.** (**A**) Representative gel images for
888 EMSA of Sso7d-tagged HIV-1-integrase (E152Q) with different types of nucleic acids substrates (R:D+ssDNA
889 and Hybrid). 100 nM nucleic acid substrate was incubated with Sso7d-tagged HIV-1-integrase (E152Q) at 0 nM,
890 20 nM, 50 nM, 100 nM, 200 nM, and 400 nM (n = 3). (**B**) Nucleic acid extracts from FLAG-HIV-1-integrase-
891 expressing cells were immunoprecipitated using S9.6 antibody. gDNA was precipitated from the elutes of
892 immunoprecipitation and subjected to DNA–RNA hybrid dot blotting. Where indicated, the gDNA extracts were
893 either untreated (–) or treated (+) with RNase H after elution of immunoprecipitated materials. (**C**) Summary of
894 the experimental design for R-loop immunoprecipitation using S9.6 antibody in FLAG-tagged HIV-1 integrase
895 protein-expressing HeLa cells with pre-immunoprecipitation in vitro RNase H treatment. (**D**) Protein extracts

896 from FLAG-HIV-1-integrase-expressing cells were immunoprecipitated using anti-FLAG antibody. Western blot
897 of FLAG immunoprecipitation was probed with anti-FLAG or anti-H3 antibodies. (**E**) Representative images of
898 the proximity-ligation assay (PLA) using single antibody (anti-GFP or anti-S9.6) in HIV-IN-EGFP virion-infected
899 HeLa cells at 6 hpi, as PLA signal negative controls. Cells were subjected to PLA (orange) and co-stained with
900 DAPI (blue) ($n > 50$).

901

HeLa				
Gene	**Chromosom**	**Position (hg38)**	**Description**	**Average DRIPc-seq signal**
RPL13A	chr19	49487608-49493057	Input (-)_0hpi	3.59
			Input (-)_3hpi	0.24
			Input (-)_6hpi	2.39
			Input (-)_12hpi	3.51
			Input (+)_0hpi	82.29
			Input (+)_3hpi	51.76
			Input (+)_6hpi	39.14
			Input (+)_12hpi	176.73
			IP_RNase H- (-)_0hpi	2.21
			IP_RNase H- (-)_3hpi	2.73
			IP_RNase H- (-)_6hpi	2.39
			IP_RNase H- (-)_12hpi	4.25
			IP_RNase H- (+)_0hpi	110.32
			IP_RNase H- (+)_3hpi	140.22
			IP_RNase H- (+)_6hpi	58.36
			IP_RNase H- (+)_12hpi	137.37
			IP_RNase H+ (-)_0hpi	0.00
			IP_RNase H+ (-)_3hpi	4.48
			IP_RNase H+ (-)_6hpi	3.74
			IP_RNase H+ (-)_12hpi	0.00
			IP_RNase H+ (+)_0hpi	1.98
			IP_RNase H+ (+)_3hpi	3.36
			IP_RNase H+ (+)_6hpi	1.60
			IP_RNase H+ (+)_12hpi	6.81
Gene	**Chromosom**	**Position (hg38)**	**Description**	**Average DRIPc-seq signal**
CALM3	chr19	46601330-46610782	Input (-)_0hpi	1.47
			Input (-)_3hpi	1.02
			Input (-)_6hpi	2.46
			Input (-)_12hpi	0.74
			Input (+)_0hpi	26.50
			Input (+)_3hpi	19.95
			Input (+)_6hpi	11.61
			Input (+)_12hpi	56.92
			IP_RNase H- (-)_0hpi	0.90
			IP_RNase H- (-)_3hpi	1.54
			IP_RNase H- (-)_6hpi	1.23
			IP_RNase H- (-)_12hpi	1.73
			IP_RNase H- (+)_0hpi	13.97
			IP_RNase H- (+)_3hpi	28.68
			IP_RNase H- (+)_6hpi	10.58
			IP_RNase H- (+)_12hpi	24.70
			IP_RNase H+ (-)_0hpi	0.71
			IP_RNase H+ (-)_3hpi	1.83
			IP_RNase H+ (-)_6hpi	2.78
			IP_RNase H+ (-)_12hpi	1.04
			IP_RNase H+ (+)_0hpi	2.12
			IP_RNase H+ (+)_3hpi	1.64
			IP_RNase H+ (+)_6hpi	2.26
			IP_RNase H+ (+)_12hpi	1.65
Gene	**Chromosom**	**Position (hg38)**	**Description**	**Average DRIPc-seq signal**
SNRPN	chr15	24823647-24978582	Input (-)_0hpi	1.46
			Input (-)_3hpi	1.27
			Input (-)_6hpi	1.34
			Input (-)_12hpi	1.76
			Input (+)_0hpi	1.21
			Input (+)_3hpi	0.81
			Input (+)_6hpi	1.25
			Input (+)_12hpi	0.41
			IP_RNase H- (-)_0hpi	0.45
			IP_RNase H- (-)_3hpi	0.47
			IP_RNase H- (-)_6hpi	0.37
			IP_RNase H- (-)_12hpi	0.05
			IP_RNase H- (+)_0hpi	0.37
			IP_RNase H- (+)_3hpi	0.24
			IP_RNase H- (+)_6hpi	0.54
			IP_RNase H- (+)_12hpi	0.07
			IP_RNase H+ (-)_0hpi	1.40
			IP_RNase H+ (-)_3hpi	0.93
			IP_RNase H+ (-)_6hpi	1.10
			IP_RNase H+ (-)_12hpi	1.31
			IP_RNase H+ (+)_0hpi	1.18
			IP_RNase H+ (+)_3hpi	1.12
			IP_RNase H+ (+)_6hpi	1.26
			IP_RNase H+ (+)_12hpi	1.10

S1 Table. Chromosomal position and DRIPc-seq signal for referenced R-loop-positive and –negative regions in HIV-1 infected HeLa cells

CD4+				
Gene	**Chromosom**	**Position (hg38)**	**Description**	**Average DRIPc-seq signal**
RPL13A	chr19	49487608-49493057	Input (-)_0hpi	2.33
			Input (-)_3hpi	1.51
			Input (-)_6hpi	2.56
			Input (-)_12hpi	0.77
			Input (+)_0hpi	2.91
			Input (+)_3hpi	1.94
			Input (+)_6hpi	2.36
			Input (+)_12hpi	2.19
			IP_RNase H- (-)_0hpi	0.00
			IP_RNase H- (-)_3hpi	3.63
			IP_RNase H- (-)_6hpi	0.00
			IP_RNase H- (-)_12hpi	0.00
			IP_RNase H- (+)_0hpi	144.19
			IP_RNase H- (+)_3hpi	77.26
			IP_RNase H- (+)_6hpi	130.86
			IP_RNase H- (+)_12hpi	190.08
			IP_RNase H+ (-)_0hpi	1.42
			IP_RNase H+ (-)_3hpi	0.00
			IP_RNase H+ (-)_6hpi	0.00
			IP_RNase H+ (-)_12hpi	0.00
			IP_RNase H+ (+)_0hpi	0.93
			IP_RNase H+ (+)_3hpi	0.00
			IP_RNase H+ (+)_6hpi	0.00
			IP_RNase H+ (+)_12hpi	2.28
Gene	**Chromosom**	**Position (hg38)**	**Description**	**Average DRIPc-seq signal**
CALM3	chr19	46601330-46610782	Input (-)_0hpi	4.58
			Input (-)_3hpi	4.64
			Input (-)_6hpi	2.96
			Input (-)_12hpi	4.04
			Input (+)_0hpi	3.62
			Input (+)_3hpi	3.65
			Input (+)_6hpi	3.40
			Input (+)_12hpi	4.11
			IP_RNase H- (-)_0hpi	0.00
			IP_RNase H- (-)_3hpi	0.00
			IP_RNase H- (-)_6hpi	0.00
			IP_RNase H- (-)_12hpi	2.70
			IP_RNase H- (+)_0hpi	108.23
			IP_RNase H- (+)_3hpi	183.80
			IP_RNase H- (+)_6hpi	87.73
			IP_RNase H- (+)_12hpi	181.80
			IP_RNase H+ (-)_0hpi	2.80
			IP_RNase H+ (-)_3hpi	0.00
			IP_RNase H+ (-)_6hpi	0.00
			IP_RNase H+ (-)_12hpi	1.94
			IP_RNase H+ (+)_0hpi	4.11
			IP_RNase H+ (+)_3hpi	1.19
			IP_RNase H+ (+)_6hpi	9.88
			IP_RNase H+ (+)_12hpi	6.17
Gene	**Chromosom**	**Position (hg38)**	**Description**	**Average DRIPc-seq signal**
SNRPN	chr15	24823647-24978582	Input (-)_0hpi	1.65
			Input (-)_3hpi	1.41
			Input (-)_6hpi	1.74
			Input (-)_12hpi	1.15
			Input (+)_0hpi	1.72
			Input (+)_3hpi	1.46
			Input (+)_6hpi	1.97
			Input (+)_12hpi	1.29
			IP_RNase H- (-)_0hpi	0.31
			IP_RNase H- (-)_3hpi	0.27
			IP_RNase H- (-)_6hpi	0.10
			IP_RNase H- (-)_12hpi	0.27
			IP_RNase H- (+)_0hpi	0.98
			IP_RNase H- (+)_3hpi	1.00
			IP_RNase H- (+)_6hpi	0.53
			IP_RNase H- (+)_12hpi	0.56
			IP_RNase H+ (-)_0hpi	0.94
			IP_RNase H+ (-)_3hpi	1.57
			IP_RNase H+ (-)_6hpi	0.00
			IP_RNase H+ (-)_12hpi	2.17
			IP_RNase H+ (+)_0hpi	1.37
			IP_RNase H+ (+)_3hpi	1.14
			IP_RNase H+ (+)_6hpi	1.42
			IP_RNase H+ (+)_12hpi	1.19

S2 Table. Chromosomal position and DRIPc-seq signal for referenced R-loop-positive and –negative regions in HIV-1 infected primary CD4^{+} T cells

Jurkat				
Gene	**Chromosom**	**Position (hg38)**	**Description**	**Average DRIPc-seq signal**
RPL13A	chr19	49487608-49493057	Input (-)_0hpi	1.46
			Input (-)_3hpi	1.92
			Input (-)_6hpi	1.92
			Input (-)_12hpi	1.58
			Input (+)_0hpi	1.40
			Input (+)_3hpi	2.02
			Input (+)_6hpi	1.15
			Input (+)_12hpi	1.54
			IP_RNase H- (-)_0hpi	0.00
			IP_RNase H- (-)_3hpi	10.17
			IP_RNase H- (-)_6hpi	9.60
			IP_RNase H- (-)_12hpi	2.64
			IP_RNase H- (+)_0hpi	404.40
			IP_RNase H- (+)_3hpi	183.88
			IP_RNase H- (+)_6hpi	486.50
			IP_RNase H- (+)_12hpi	526.25
			IP_RNase H+ (-)_0hpi	0.00
			IP_RNase H+ (-)_3hpi	3.53
			IP_RNase H+ (-)_6hpi	0.00
			IP_RNase H+ (-)_12hpi	0.00
			IP_RNase H+ (+)_0hpi	6.13
			IP_RNase H+ (+)_3hpi	0.00
			IP_RNase H+ (+)_6hpi	0.00
			IP_RNase H+ (+)_12hpi	0.00
Gene	**Chromosom**	**Position (hg38)**	**Description**	**Average DRIPc-seq signal**
CALM3	chr19	46601330-46610782	Input (-)_0hpi	2.40
			Input (-)_3hpi	2.18
			Input (-)_6hpi	2.26
			Input (-)_12hpi	2.78
			Input (+)_0hpi	2.08
			Input (+)_3hpi	2.78
			Input (+)_6hpi	1.99
			Input (+)_12hpi	2.38
			IP_RNase H- (-)_0hpi	0.00
			IP_RNase H- (-)_3hpi	11.73
			IP_RNase H- (-)_6hpi	5.58
			IP_RNase H- (-)_12hpi	5.22
			IP_RNase H- (+)_0hpi	208.25
			IP_RNase H- (+)_3hpi	182.67
			IP_RNase H- (+)_6hpi	167.98
			IP_RNase H- (+)_12hpi	220.30
			IP_RNase H+ (-)_0hpi	0.00
			IP_RNase H+ (-)_3hpi	2.04
			IP_RNase H+ (-)_6hpi	0.00
			IP_RNase H+ (-)_12hpi	4.84
			IP_RNase H+ (+)_0hpi	13.84
			IP_RNase H+ (+)_3hpi	1.62
			IP_RNase H+ (+)_6hpi	4.37
			IP_RNase H+ (+)_12hpi	3.29
Gene	**Chromosom**	**Position (hg38)**	**Description**	**Average DRIPc-seq signal**
SNRPN	chr15	24823647-24978582	Input (-)_0hpi	1.75
			Input (-)_3hpi	1.94
			Input (-)_6hpi	1.87
			Input (-)_12hpi	1.84
			Input (+)_0hpi	1.86
			Input (+)_3hpi	1.89
			Input (+)_6hpi	1.81
			Input (+)_12hpi	1.73
			IP_RNase H- (-)_0hpi	0.12
			IP_RNase H- (-)_3hpi	0.00
			IP_RNase H- (-)_6hpi	0.17
			IP_RNase H- (-)_12hpi	0.00
			IP_RNase H- (+)_0hpi	2.43
			IP_RNase H- (+)_3hpi	2.19
			IP_RNase H- (+)_6hpi	2.23
			IP_RNase H- (+)_12hpi	2.36
			IP_RNase H+ (-)_0hpi	2.58
			IP_RNase H+ (-)_3hpi	3.46
			IP_RNase H+ (-)_6hpi	1.62
			IP_RNase H+ (-)_12hpi	1.87
			IP_RNase H+ (+)_0hpi	1.78
			IP_RNase H+ (+)_3hpi	2.38
			IP_RNase H+ (+)_6hpi	1.06
			IP_RNase H+ (+)_12hpi	1.43

S3 Table. Chromosomal position and DRIPc-seq signal for referenced R-loop-positive and –negative regions in HIV-1 infected Jurkat cells

	gene_symbol	0hpi	3hpi	6hpi	12hpi	48hpi
P1	TOR1AIP2	1	1.163832	1.247899	1.024926	0.619497
P2	DVL1	1	0.781593	0.571348	0.901502	0.270459
P3	PKN2	1	1.280974	1.891552	1.31842	1.515107
N1	N/A	N/A	N/A	N/A	N/A	N/A
N2	CDK5RAP1	1	0.73977	0.775	1.143662	0.472377

S4 Table. RNA-seq analysis of relative gene expression levels of P1-3 and N1,2 R-loop regions

915

916 **S5 Table. Oligonucleotides used for DRIPc-seq library construction**

Oligonucleotides	Sequence 5' to 3'	Remark
PCR primer 1.0 P5	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGA	amplification primer
PCR primer 2.0 P7	CAAGCAGAAGACGGCATACGAGAT	amplification primer
Index Adapter 1	GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTG	HeLa 0hpi Input replicate 1
Index Adapter 2	GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCCGTCTTCTGCTTG	HeLa 3hpi Input replicate 1
Index Adapter 3	GATCGGAAGAGCACACGTCTGAACTCCAGTCACTTAGGCATCTCGTATGCCGTCTTCTGCTTG	HeLa 6hpi Input replicate 1
Index Adapter 4	GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATGCCGTCTTCTGCTTG	HeLa 12hpi Input replicate 1
Index Adapter 5	GATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGCCGTCTTCTGCTTG	HeLa 0hpi RNH-IP replicate 1
Index Adapter 6	GATCGGAAGAGCACACGTCTGAACTCCAGTCACGCCAATATCTCGTATGCCGTCTTCTGCTTG	HeLa 3hpi RNH-IP replicate 1
Index Adapter 7	GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGATCATCTCGTATGCCGTCTTCTGCTTG	HeLa 6hpi RNH-IP replicate 1
Index Adapter 8	GATCGGAAGAGCACACGTCTGAACTCCAGTCACACTTGAATCTCGTATGCCGTCTTCTGCTTG	HeLa 12hpi RNH-IP replicate 1
Index Adapter 9	GATCGGAAGAGCACACGTCTGAACTCCAGTCACGATCAGATCTCGTATGCCGTCTTCTGCTTG	HeLa 0hpi RNH+IP replicate 1
Index Adapter 10	GATCGGAAGAGCACACGTCTGAACTCCAGTCACTAGCTTATCTCGTATGCCGTCTTCTGCTTG	HeLa 3hpi RNH+IP replicate 1
Index Adapter 11	GATCGGAAGAGCACACGTCTGAACTCCAGTCACGGCTACATCTCGTATGCCGTCTTCTGCTTG	HeLa 6hpi RNH+IP replicate 1
Index Adapter 12	GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTTGTAATCTCGTATGCCGTCTTCTGCTTG	HeLa 12hpi RNH+IP replicate 1
Index Adapter 28	GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAAAAGATCTCGTATGCCGTCTTCTGCTTG	HeLa 0hpi Input replicate 2
Index Adapter 29	GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAACTAATCTCGTATGCCGTCTTCTGCTTG	HeLa 3hpi Input replicate 2
Index Adapter 30	GATCGGAAGAGCACACGTCTGAACTCCAGTCACCACCGGATCTCGTATGCCGTCTTCTGCTTG	HeLa 6hpi Input replicate 2
Index Adapter 31	GATCGGAAGAGCACACGTCTGAACTCCAGTCACCACGATATCTCGTATGCCGTCTTCTGCTTG	HeLa 12hpi Input replicate 2
Index Adapter 32	GATCGGAAGAGCACACGTCTGAACTCCAGTCACCACTCAATCTCGTATGCCGTCTTCTGCTTG	HeLa 0hpi RNH-IP replicate 2
Index Adapter 33	GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGGCGATCTCGTATGCCGTCTTCTGCTTG	HeLa 3hpi RNH-IP replicate 2
Index Adapter 34	GATCGGAAGAGCACACGTCTGAACTCCAGTCACCATGGCATCTCGTATGCCGTCTTCTGCTTG	HeLa 6hpi RNH-IP replicate 2
Index Adapter 35	GATCGGAAGAGCACACGTCTGAACTCCAGTCACCATTTTATCTCGTATGCCGTCTTCTGCTTG	HeLa 12hpi RNH-IP replicate 2
Index Adapter 36	GATCGGAAGAGCACACGTCTGAACTCCAGTCACCCAACAATCTCGTATGCCGTCTTCTGCTTG	HeLa 0hpi RNH+IP replicate 2
Index Adapter 37	GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGGAATATCTCGTATGCCGTCTTCTGCTTG	HeLa 3hpi RNH+IP replicate 2
Index Adapter 38	GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTAGCTATCTCGTATGCCGTCTTCTGCTTG	HeLa 6hpi RNH+IP replicate 2
Index Adapter 39	GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTATACATCTCGTATGCCGTCTTCTGCTTG	HeLa 12hpi RNH+IP replicate 2

Oligonucleotides	Sequence 5' to 3'
P1 Fwd	TTATAAGTCAGCCTCCAGGATCAA
P1 Rev	TTCAGGTCTAGGCAGTCTGA
P2 Fwd	GGA CAG ATG ACA GGG TCG C
P2 Rev	ATG AGG AAG ACC CCC TCG G
P3 Fwd	CTCTGTGTAACGCTGGTGCT
P3 Rev	ACACGCTTCTGACCACTAAGG
N1 Fwd	TTG GCC CTA CTG AAT GAT TGG T
N1 Rev	TTA AGG CAT GCT CAG GCG A
N2 Fwd	TGA GAT TTC AGG TTC CAT GAT TTG
N2 Rev	TGC TCA GTG TTC TAA TTT CCC TGT
β-actin Fwd	AGAGCTACGAGCTGCCTGAC
β-actin Rev	AGCACTGTGTTGGCGTACAG
SH49 (ECFP Fwd)	TGGTTTGTCCAAACTCATCAA
SH40 (mAlRN Fwd)	CGAGAGAGGCTAAGGGTGAA
SH21 (ECFP/mAlRN Rev)	ACATGGTCCTGCTGGAGTTC
RT-qPCR P1 (TOR1AIP2) Fwd	CCTTGGTCTTTCCCACTTGAGTG
RT-qPCR P1 (TOR1AIP2) Rev	GCAGGGTTAAAACCAGCTACTCG
RT-qPCR P2 (DVL1) Fwd	GCATAACCGACTCCACCATGTC
RT-qPCR P2 (DVL1) Rev	GATGGAGCCAATGTAGATGCCG
RT-qPCR P3 (PKN2) Fwd	GCATCACCAACACTAAGTCCACG
RT-qPCR P3 (PKN2) Rev	GCTTTTGACCGTCCAGGGACAT
RT-qPCR N2 (CDK5RAP1) Fwd	AGAGTGGAAGCAGCCGTGTGTT
RT-qPCR N2 (CDK5RAP1) Rev	GATCTTCCTCCGTCTCACCACA

S6 Table. Primers used for qPCR

Oligonucleotides	Sequence 5' to 3'	Remark
AE5316	TGTGACTCTGGTAACTAGAGATCCCTC	First round LTR primer
AE6380	TAGTCCCTTAAGCGGAG-NH2	replicate 1 5dpi Linker short / replicate 1 pgR-poor DOX- Linker short / CD4+ donor 1 Linker short
AE6381	GTAATACGACTCACTATAGGGCCTCCGCTTAAGGGAC	replicate 1 5dpi Linker long / replicate 1 pgR-poor DOX- Linker long / CD4+ donor 1 Linker long
AE6382	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTGTAATACGACTCACTATAGGGC	replicate 1 5dpi Linker primer / replicate 1 pgR-poor DOX- Linker primer / CD4+ donor 1 Linker primer
AE6404	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCGATGTGAGATCCCTCAGACCCTTTTAGTCAG	replicate 1 5dpi Second round LTR primer / replicate 1 pgR-poor DOX- Second round LTR primer / CD4+ donor 1 Second round LTR primer
AE6380	TAGTCCCTTAAGCGGAG-NH2	replicate 2 5dpi Linker short / replicate 2 pgR-poor DOX+ Linker short
AE6381	GTAATACGACTCACTATAGGGCCTCCGCTTAAGGGAC	replicate 2 5dpi Linker long / replicate 2 pgR-poor DOX+ Linker long
AE6382	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTGTAATACGACTCACTATAGGGC	replicate 2 5dpi Linker primer / replicate 2 pgR-poor DOX+ Linker prime
AE6404-1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTTAGGCGAGATCCCTCAGACCCTTTTAGTCAG	replicate 2 5dpi Second round LTR primer / replicate 2 pgR-poor DOX+ Second round LTR primer
AE6386	TACTATGACGGTGACGC-NH2	replicate 1 pgR-rich DOX- Linker short / CD4+ donor 2 Linker short
AE6387	GAGAATCCATGAGTATGCTCACGCGTCACCGTCATAG	replicate 1 pgR-rich DOX- Linker long / CD4+ donor 2 Linker long
AE6388	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTGAGAATCCATGAGTATGCTCAC	replicate 1 pgR-rich DOX- Linker primer / CD4+ donor 2 Linker primer
AE6406	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTACAGTGGAGATCCCTCAGACCCTTTTAGTCAG	replicate 1 pgR-rich DOX- Second round LTR primer / CD4+ donor 2 Second round LTR primer
AE6456	TAGACTGACGCAGTCTG-NH2	replicate 1 pgR-poor DOX+ Linker short
AE6457	GACGTACATACTGATCGCATAGCAGACTGCGTCAGTC	replicate 1 pgR-poor DOX+ Linker long
AE6458	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTGACGTACATACTGATCGCATAG	replicate 1 pgR-poor DOX+ Linker primer
AE6405	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTGACCAGAGATCCCTCAGACCCTTTTAGTCAG	replicate 1 pgR-poor DOX+ Second round LTR primer
AE6386	TACTATGACGGTGACGC-NH2	replicate 2 pgR-rich DOX+ Linker short
AE6387	GAGAATCCATGAGTATGCTCACGCGTCACCGTCATAG	replicate 2 pgR-rich DOX+ Linker long
AE6388	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTGAGAATCCATGAGTATGCTCAC	replicate 2 pgR-rich DOX+ Linker primer
AE6406-1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGCCAATGAGATCCCTCAGACCCTTTTAGTCAG	replicate 2 pgR-rich DOX+ Second round LTR primer
AE6456	TAGACTGACGCAGTCTG-NH2	replicate 3 pgR-rich DOX- Linker short
AE6457	GACGTACATACTGATCGCATAGCAGACTGCGTCAGTC	replicate 3 pgR-rich DOX- Linker long
AE6458	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTGACGTACATACTGATCGCATAG	replicate 3 pgR-rich DOX- Linker primer
AE6411	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTTCCGAGATCCCTCAGACCCTTTTAGTCAG	replicate 3 pgR-rich DOX- Second round LTR primer
AE6392	TACTGAGACGTCGATGC-NH2	replicate 1 RNH_mut 5dpi Linker short / replicate 2 RNH_mut 5dpi Linker short
AE6393	GATCATGCGAGATACATCTCAGGCATCGACGTCTCAG	replicate 1 RNH_mut 5dpi Linker long / replicate 2 RNH_mut 5dpi Linker long
AE6394	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTGATCATGCGAGATACATCTCAG	replicate 1 RNH_mut 5dpi Linker primer / replicate 2 RNH_mut 5dpi Linker primer
AE6493	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGGCTACGAGATCCCTCAGACCCTTTTAGTCAG	replicate 1 RNH_mut 5dpi Second round LTR primer
AE6493-1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTACTTGAGAGATCCCTCAGACCCTTTTAGTCAG	replicate 2 RNH_mut 5dpi Second round LTR primer
AE6462	TAGTAGTCACGAGCGTC-NH2	replicate 1 RNH_wt 5dpi Linker short / replicate 2 RNH_wt 5dpi Linker short
AE6463	CAGTTAGACTACACGTTAGACGGACGCTCGTGACTAC	replicate 1 RNH_wt 5dpi Linker long / replicate 2 RNH_wt 5dpi Linker long
AE6464	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTCAGTTAGACTACACGTTAGACG	replicate 1 RNH_wt 5dpi Linker primer / replicate 2 RNH_wt 5dpi Linker primer
AE6492	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTAGCTTGAGATCCCTCAGACCCTTTTAGTCAG	replicate 1 RNH_wt 5dpi Second round LTR primer
AE6497	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTATCACGGAGATCCCTCAGACCCTTTTAGTCAG	replicate 2 RNH_wt 5dpi Second round LTR primer

S7 Table. Oligonucleotides used for HIV-1 integration site sequencing library construct

Oligonucleotides	Sequence 5' to 3'	Remark
R-loop oligo1*	5'-[Cy3]-GCC AGG GAC GAG GTG AAC CTG CAG GTG GGC GGC TAC TAC TTA GAT GTC ATC CGA GGC TTA TTG GTA GAA TTC GGC AGC GTC ATG C GA CGG C-3'	R-loop, R:D+ssDNA, ssDNA
R-loop oligo2*	5'-GCC GTC GCA TGA CGC TGC CGA ATT CTA CCA CGC GAT TCA TAC CTG TCG TGC CAG CTG CTT TGC CCA CCT GCA GGT TCA CCT CGT CCC TGG C-3'	R-loop, dsDNA
R-loop RNA	5'-[Cy5]-GCA GCU GGC ACG ACA GGU AUG AAU C-3'	R-loop, R:D+ssDNA, ssRNA
Homoduplex	5'-[Cy3]-GCC AGG GAC GAG GTG AAC CTG CAG GTG GGC AAA GCA GCT GGC ACG ACA GGT ATG AAT CGC GTG GTA GAA TTC GGC AGC GTC ATG CGA CGG C-3'	dsDNA
Hybrid DNA	5'-CCC ATA CCG TAT AAC CAT TTG GCT GTC CAA GCT CCG GGT-3'	Hybrid
Hybrid RNA	5'-[Cy5]-ACC CGG AGC UUG GAC AGC CAA AUG GUU AUA CGG UAU GGG-3'	Hybrid
oligo 5	5′GCAGTAGCATGACGCTGCTGAATTCTACCACGCTATGCT CTCGTCTAGGTTCACTCCGT CCCTGCGATTCATACCTGTCGTGCCAGCTGC	R:D+ssDNA

S8 Table. Oligonucleotides used for electrophoretic mobility shift assay substrate preparation

Data	Accession Number/ Website
Jurkat integration site	SRR12322252
TSA-seq_SPAD	SRR3538917, SRR3538918, SRR3538919, SRR3538920
SPIN (Spatial Position Inference of the Nuclear genome) annotation of speckle	https://github.com/ma-compbio/SPIN
LADs	GSE22428

S9 Table. Accession numbers and data sources.

926 # References

927 1. W. E. Johnson, Origins and evolutionary consequences of ancient endogenous retroviruses.
928 *Nat Rev Microbiol* **17**, 355-370 (2019).
929 2. M. Lusic, R. F. Siliciano, Nuclear landscape of HIV-1 infection and integration. *Nat Rev*
930 *Microbiol* **15**, 69-82 (2017).
931 3. H. C. Chen, J. P. Martinez, E. Zorita, A. Meyerhans, G. J. Filion, Position effects influence HIV
932 latency reversal. *Nat Struct Mol Biol* **24**, 47-54 (2017).
933 4. K. B. Einkauf *et al.*, Parallel analysis of transcription, integration, and sequence of single HIV-
934 1 proviruses. *Cell* **185**, 266-282 e215 (2022).
935 5. C. Jiang *et al.*, Distinct viral reservoirs in individuals with spontaneous control of HIV-1.
936 *Nature* **585**, 261-267 (2020).
937 6. A. R. Schroder *et al.*, HIV-1 integration in the human genome favors active genes and local
938 hotspots. *Cell* **110**, 521-529 (2002).
939 7. V. Achuthan *et al.*, Capsid-CPSF6 Interaction Licenses Nuclear HIV-1 Trafficking to Sites of
940 Viral DNA Integration. *Cell Host Microbe* **24**, 392-404 e398 (2018).
941 8. A. Ciuffi *et al.*, A role for LEDGF/p75 in targeting HIV DNA integration. *Nat Med* **11**, 1287-
942 1289 (2005).
943 9. G. A. Sowd *et al.*, A critical role for alternative polyadenylation factor CPSF6 in targeting
944 HIV-1 integration to transcriptionally active chromatin. *Proc Natl Acad Sci U S A* **113**, E1054-
945 1063 (2016).
946 10. B. Lucic *et al.*, Spatially clustered loci with multiple enhancers are frequent targets of HIV-1
947 integration. *Nat Commun* **10**, 4059 (2019).
948 11. B. Marini *et al.*, Nuclear architecture dictates HIV-1 integration site selection. *Nature* **521**,
949 227-231 (2015).
950 12. M. Kvaratskhelia, A. Sharma, R. C. Larue, E. Serrao, A. Engelman, Molecular mechanisms of
951 retroviral integration site selection. *Nucleic Acids Res* **42**, 10209-10225 (2014).
952 13. P. Cherepanov *et al.*, HIV-1 integrase forms stable tetramers and associates with LEDGF/p75
953 protein in human cells. *J Biol Chem* **278**, 372-381 (2003).
954 14. C. Niehrs, B. Luke, Regulatory R-loops as facilitators of gene expression and genome stability.
955 *Nat Rev Mol Cell Biol* **21**, 167-178 (2020).
956 15. E. Petermann, L. Lan, L. Zou, Sources, resolution and physiological relevance of R-loops and
957 RNA-DNA hybrids. *Nat Rev Mol Cell Biol* **23**, 521-540 (2022).
958 16. S. Hamperl, M. J. Bocek, J. C. Saldivar, T. Swigut, K. A. Cimprich, Transcription-Replication
959 Conflict Orientation Modulates R-Loop Levels and Activates Distinct DNA Damage
960 Responses. *Cell* **170**, 774-786 e719 (2017).
961 17. P. A. Ginno, P. L. Lott, H. C. Christensen, I. Korf, F. Chedin, R-loop formation is a distinctive

962 characteristic of unmethylated human CpG island promoters. *Mol Cell* **45**, 814-825 (2012).
963 18. Y. W. Lim, L. A. Sanz, X. Xu, S. R. Hartono, F. Chedin, Genome-wide DNA hypomethylation
964 and RNA:DNA hybrid accumulation in Aicardi-Goutieres syndrome. *Elife* **4**, (2015).
965 19. R. Arora *et al.*, RNaseH1 regulates TERRA-telomeric DNA hybrids and telomere maintenance
966 in ALT tumour cells. *Nat Commun* **5**, 5220 (2014).
967 20. T. Garcia-Muse, A. Aguilera, R Loops: From Physiological to Pathological Roles. *Cell* **179**,
968 604-618 (2019).
969 21. L. A. Sanz *et al.*, Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific
970 Epigenomic Signatures in Mammals. *Mol Cell* **63**, 167-178 (2016).
971 22. F. Chedin, Nascent Connections: R-Loops and Chromatin Patterning. *Trends Genet* **32**, 828-
972 838 (2016).
973 23. C. Y. Lee *et al.*, R-loop induced G-quadruplex in non-template promotes transcription by
974 successive R-loop formation. *Nat Commun* **11**, 3392 (2020).
975 24. H. O. Ajoge *et al.*, G-Quadruplex DNA and Other Non-Canonical B-Form DNA Motifs
976 Influence Productive and Latent HIV-1 Integration and Reactivation Potential. *Viruses* **14**,
977 (2022).
978 25. F. Chedin, C. J. Benham, Emerging roles for R-loop structures in the management of
979 topological stress. *J Biol Chem* **295**, 4684-4695 (2020).
980 26. I. K. Jozwik *et al.*, B-to-A transition in target DNA during retroviral integration. *Nucleic Acids*
981 *Res* **50**, 8898-8918 (2022).
982 27. A. Ballandras-Colas *et al.*, Multivalent interactions essential for lentiviral integrase function.
983 *Nat Commun* **13**, 2416 (2022).
984 28. L. A. Sanz, F. Chedin, High-resolution, strand-specific R-loop mapping via S9.6-based DNA-
985 RNA immunoprecipitation and high-throughput sequencing. *Nat Protoc* **14**, 1734-1755
986 (2019).
987 29. R. B. Jones *et al.*, LINE-1 retrotransposable element DNA accumulates in HIV-1-infected cells.
988 *J Virol* **87**, 13307-13320 (2013).
989 30. S. Srinivasachar Badarinarayan *et al.*, HIV-1 infection activates endogenous retroviral
990 promoters regulating antiviral gene expression. *Nucleic Acids Res* **48**, 10890-10908 (2020).
991 31. P. Lesbats, A. N. Engelman, P. Cherepanov, Retroviral DNA Integration. *Chem Rev* **116**, 12730-
992 12757 (2016).
993 32. A. Brussel, P. Sonigo, Analysis of early human immunodeficiency virus type 1 DNA synthesis
994 by use of a new sensitive assay for quantifying integrated provirus. *J Virol* **77**, 10119-10124
995 (2003).
996 33. A. Albanese, D. Arosio, M. Terreni, A. Cereseto, HIV-1 pre-integration complexes selectively
997 target decondensed chromatin in the nuclear periphery. *PLoS One* **3**, e2413 (2008).
998 34. A. Dharan, N. Bachmann, S. Talley, V. Zwikelmaier, E. M. Campbell, Nuclear pore blockade
999 reveals that HIV-1 completes reverse transcription and uncoating in the nucleus. *Nat*

1000 *Microbiol* **5**, 1088-1095 (2020).
1001 35. J. J. Kessl *et al.*, HIV-1 Integrase Binds the Viral RNA Genome and Is Essential during Virion
1002 Morphogenesis. *Cell* **166**, 1257-1268 e1212 (2016).
1003 36. D. C. van Gent, Y. Elgersma, M. W. Bolk, C. Vink, R. H. Plasterk, DNA binding properties of
1004 the integrase proteins of human immunodeficiency viruses types 1 and 2. *Nucleic Acids Res*
1005 **19**, 3821-3827 (1991).
1006 37. A. Cristini, M. Groh, M. S. Kristiansen, N. Gromak, RNA/DNA Hybrid Interactome Identifies
1007 DXH9 as a Molecular Player in Transcriptional Termination and R-Loop-Associated DNA
1008 Damage. *Cell Rep* **23**, 1891-1905 (2018).
1009 38. T. Mosler *et al.*, R-loop proximity proteomics identifies a role of DDX41 in transcription-
1010 associated genomic instability. *Nat Commun* **12**, 7314 (2021).
1011 39. R. Schrijvers *et al.*, LEDGF/p75-independent HIV-1 replication demonstrates a role for HRP-
1012 2 and remains sensitive to inhibition by LEDGINs. *PLoS Pathog* **8**, e1002558 (2012).
1013 40. P. C. Stirling, P. Hieter, Canonical DNA Repair Pathways Influence R-Loop-Driven Genome
1014 Instability. *J Mol Biol* **429**, 3132-3138 (2017).
1015 41. M. L. Garcia-Rubio *et al.*, The Fanconi Anemia Pathway Protects Genome Integrity from R-
1016 loops. *PLoS Genet* **11**, e1005674 (2015).
1017 42. M. Giannini *et al.*, TDP-43 mutations link Amyotrophic Lateral Sclerosis with R-loop
1018 homeostasis and R loop-mediated DNA damage. *PLoS Genet* **16**, e1009260 (2020).
1019 43. S. Fu *et al.*, HIV-1 exploits the Fanconi anemia pathway for viral DNA integration. *Cell Rep*
1020 **39**, 110840 (2022).
1021 44. D. Li, A. Lopez, C. Sandoval, R. Nichols Doyle, O. I. Fregoso, HIV Vpr Modulates the Host
1022 DNA Damage Response at Two Independent Steps to Damage DNA and Repress Double-
1023 Strand DNA Break Repair. *mBio* **11**, (2020).
1024 45. H. Bauby *et al.*, HIV-1 Vpr Induces Widespread Transcriptomic Changes in CD4(+) T Cells
1025 Early Postinfection. *mBio* **12**, e0136921 (2021).
1026 46. K. Stopak, C. de Noronha, W. Yonemoto, W. C. Greene, HIV-1 Vif blocks the antiviral activity
1027 of APOBEC3G by impairing both its translation and intracellular stability. *Mol Cell* **12**, 591-
1028 601 (2003).
1029 47. D. Kmiec, F. Kirchhoff, Antiviral factors and their counteraction by HIV-1: many uncovered
1030 and more to be discovered. *J Mol Cell Biol*, (2024).
1031 48. J. L. McCann *et al.*, R-loop homeostasis and cancer mutagenesis promoted by the DNA
1032 cytosine deaminase APOBEC3B. 2021.2008.2030.458235 (2021).
1033 49. S. A. Yukl *et al.*, HIV latency in isolated patient CD4(+) T cells may be due to blocks in HIV
1034 transcriptional elongation, completion, and splicing. *Sci Transl Med* **10**, (2018).
1035 50. D. O. Passos *et al.*, Cryo-EM structures and atomic model of the HIV-1 strand transfer
1036 complex intasome. *Science* **355**, 89-92 (2017).
1037 51. H. D. Nguyen *et al.*, Functions of Replication Protein A as a Sensor of R Loops and a

Regulator of RNaseH1. *Mol Cell* **65**, 832-847 e834 (2017).
52. S. Andrews. (2010).
53. F. J. Felix Krueger, Phil Ewels, Ebrahim Afyounian, & Benjamin Schuster-Boeckler, FelixKrueger/TrimGalore: v0.6.7 - DOI via Zenodo (0.6.7). Zenodo. (2021).
54. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011* **17**, 3 %J EMBnet.journal (2011).
55. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
56. Y. Zhang *et al.*, Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
57. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
58. W. Li *et al.*, CPSF6-Dependent Targeting of Speckle-Associated Domains Distinguishes Primate from Nonprimate Lentiviral Integration. *mBio* **11**, (2020).
59. H. M. Amemiya, A. Kundaje, A. P. Boyle, The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9**, 9354 (2019).
60. A. Frankish *et al.*, Gencode 2021. *Nucleic Acids Res* **49**, D916-D923 (2021).
61. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
62. Y. Jin, O. H. Tam, E. Paniagua, M. Hammell, TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* **31**, 3593-3599 (2015).