이학박사 학위논문

인간면역결핍 바이러스 감염에 의한 R-루프 구조 생성 및 바이러스 게놈 삽입 위치 조절 연구

Human immunodeficiency virus-1 induces and exploits host genomic R-loops for its proviral integration site selection

2024 년 8 월

서울대학교 대학원

생명과학부

박기원

Human immunodeficiency virus-1 induces and exploits host genomic R-loops for its proviral integration site selection

A Dissertation Submitted in Partial Fulfillment of the Requirement for the Degree of
DOCTOR OF PHILOSOPHY

To the Faculty of
School of Biological Sciences
at
Seoul National University
by
Kiwon Park

Data approved

_____

Chair _____(Seal)
Vice Chair _____(Seal)
Examiner _____(Seal)
Examiner _____(Seal)
Examiner _____(Seal)

ABSTRACT

Kiwon Park

School of Biological Sciences

The Graduate School

Seoul National University

Integration of viral DNA into the host genome is a crucial step in the Human immunodeficiency virus-1 (HIV-1) infection; however, the factors that influence the selection of integration sites are not comprehensively understood. 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, challenging the traditional understanding of HIV-1 integration site selection. Here, I 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 regions of HIV-1-induced R-loops. Using a novel cell model that can independently control transcriptional activity and R-loop formation, I demonstrated that the presence of R-loops, regardless of transcriptional activity, directs HIV-1 integration targeting sites. I also found that HIV-1 integrase proteins bind to the R-loops, in vitro and the host genomic R-loops

in cells infected with HIV-1. These findings provide fundamental insights into the mechanisms of retroviral integration and the prognosis of antiretroviral therapy.

Keyword : Human immunodeficiency virus-1, retrovirus, R-loop, DNA-RNA hybrid, integration, intasome

Student Number : 2017-26763

Table of Contents

Acknowledgement

One thing that I most frequently and deeply think during my PhD was the array of responsibilities.

It was a responsibility for the mission of working for the healthy and happy life of mankind as a PhD candidate in biological sciences, a responsibility for conducting an original research as a researcher, a responsibility of being a colleague in a genuine scientific society, and a responsibility for the faith of people who support me as a person. Sometimes, these responsibilities felt overwhelming, but as reminded me throughout the course of my degree, it is very clear that many people ease my burdens and rather these responsibilities gave me strong motivations above all else.

First of all, I would like to express my deepest gratitude to my supervisor and mentor, Professor Kwangseog Ahn. His passion in asking important questions in biology and finding the answers in a creative way have always been inspirational. As a scientist, Prof. Ahn strictly emphasizes professionalism, but I know that he cares for his students more than anyone else. I was often anxious and sometimes stubborn during my PhD, but he encouraged me to develop these weakness of mine into strengths by saying that they can be good qualifications for a scientist. Above all, there is no doubt that his respect and optimism in me as an independent researcher was the greatest motivation to me to grow as a biologist.

What was truly fortunate during my PhD was working with trustworthy and passionate collaborators. First, I would like to thank Dr. Jeongmin Ryoo who taught me how logically formulate and test scientific hypotheses as my daily supervisor in my first year of graduate school. I learned what an integrity in science is by being a co-first author with Dr. Ryoo on my first first-author paper. I look forward to continuing to be her dependable scientist colleague and a talkative friend. I am truly grateful to Dr. Dohoon Lee who is the smartest and most brilliant bioinformatician I know. I am grateful to him a lot for efficiently translating and powerfully 'dataficating' my biological questions. Also, I thank him for being a magical friend who always makes things much simpler through conversation over a cup of coffee. I am also grateful to Seongjin Ahn, a band vocalist who is the smartest and most passionate about science I know. Conversations about stupid but ideal sciences I had with him were always enjoyable. Although I sometimes felt a great sense of responsibility in carrying out my own research project, the presence of these collaborators alone helped me to cross the finish line and strongly motivated me to become a better researcher.

I am also very grateful to my closest friends and the best critics, the colleagues from the Ahn Lab. I thank the lab seniors who were ahead of me in both life and science, including the aforementioned Dr. Ryoo, Dr. Changhoon Oh who took me to the gym, and Dr. Sung-Yeon Hwang who I have always admired for his cleverness. I am especially

grateful to Seongwon Lee, Heena Jeong, and Hyewon Kim who stood by me throughout the darkest and the best time during my PhD. I wish good luck and thank to Jiseok Jeong, Dongjoon Jeong and Se Hong Park as they bring the history of the Ahn Lab to its conclusion. I would also like to thank Dr. Jun hyun Park, the best morning scientist I know, who opens the lab door before me every morning and helped me to start my days positively.

I would like to express my gratitude to my thesis committee members, Prof. V. Narry Kim, Prof. Boyoun Park, Prof. Chanhee Kang and Prof. Hyeshik Chang. They encouraged me a lot by showing the interest in my research and giving meaningful comments on the progress of the thesis, not only at the thesis review meetings but also in various scientific communication settings. In particular, I am grateful to Prof. V. Narry Kim and Prof. Hyeshik Chang who are the members of the Institute for Basic Science (IBS), Center for RNA research and have supported me through many collaborative research projects and academic events. During my coursework, I had the opportunity to take a lecture given by Prof. Chanhee Kang. He not only gave practical advices on research design and logical writing but also showed me how enjoyable it is to talk about everyone's science. I was deeply motivated by being in such genuine scientific society.

I would like to thank my friends who took me out of the lab and helped me to not be disconnected from the real world by immersing myself in science. My birthday party committee members, Sunmin, Yeojin, Seongwon Min, Seul, Juyoung, Yeonsu, Soyoung, Jiho and Haram, always remind me of different ways of surviving and diverse perspectives towards life. While I couldn't joke about the smell of TEMED ($C_6H_{16}N_2$) with them, having conversations with them about the joys and sorrows of life gave me the strength to continue doing science, despite my Western blot did not give me the secretes of life.

Lastly, but most importantly, I would like to thank my family. My mother always backs up my own decisions, but she imparts the wisest advices for important decisions at different stages of my life. She taught me the value of sincerity for progression. My father, who still worries his youngest daughter might get hurt, has shown and also taught me that the kindest person is the strongest person. I would like to take this opportunity to tell them that becoming a proud daughter has been the biggest motivation in my life. I am also grateful to my sister, who is not only the kindest but also the most sincere nurse I know, and I truly respect her. I also convey the news to my dear younger brother, Bori, in the dog heaven, that I am finally done with my doctorate.

List of Tables and Figures

List of Abbreviations

HIV-1: Human immunodeficiency virus-1

AIDS: Acquired immune deficiency syndrome

LEDGF/p75: Lens epithelium-derived growth factor/p75

CPSF6: Cleavage and polyadenylation specificity factor 6

DRIPc-seq: DNA-RNA immunoprecipitation followed by cDNA conversion coupled to high-throughput sequencing

hpi: hours post-infection

MOI: Multiplicity of infection

SINEs: Short interspersed nuclear elements

LINEs: Long interspersed nuclear elements

LTR: Long terminal repeat

RNH1: RNase H1

DRIP-qPCR: DNA-RNA immunoprecipitation followed by real-time polymerase chain reaction

DOX: Doxycycline

qPCR: Quantitative real-time PCR

EMSA: Electrophoretic mobility shift assay

PLA: Proximity-ligation assay

ChIP: Chromatin Immunoprecipitation

PIC: Preintegration complex

Vpr: Viral protein R

Vif: Viral infectivity factor

Summary

BACKGROUND: HIV-1 causes permanent infection by integrating its reverse transcribed viral genome into the host genome and the chromosomal landscape of HIV-1 integration plays a critical role in proviral gene expression, persistence of integrated proviruses and prognosis of antiretroviral therapy. HIV-1 integration has a distinct preference to actively transcribed gene regions. However, high-resolution analysis of individual HIV-1 integration sites have shown that the HIV-1 still integrates into a variety of host genomic regions, including 'gene desert' regions, challenging the traditional understanding of HIV-1 integration site selection mechanism and reflexing the possibility of there being an undiscovered determinant that composes the correct genomic environment for HIV-1 integration.

RATIONAL: R-loops are inherent nucleic acids structure that are enriched in transcribed genes during active transcription as well as widespread over the genome including non-genic regions as a result of *in trans* R-loop formation. R-loops relieve superhelical stresses and are often associated with open chromatin marks and active enhances, which are also distributed over HIV-1 integration sites. Besides, R-loops are prevalent non-canonical B-form DNA structure, which has recently been revealed as an intermediate conformation of target DNA bound by retroviral integrases. Therefore, I rationalized that investigating host genomic R-loop as a

pivotal determinant of HIV-1 integration site selection mechanism would provide fundamental insight into the unexplained mechanisms of retroviral integration.

RESULTS: Through genome-wide maps of host genomic R-loop in different cell types including primary $CD4^+$ T cells, which are natural target cell type of HIV-1 infection, I found that HIV-1 infection initiate the enrichment of host genomic R-loops over diverse genomic compartments during early post infection. I conducted global analysis of host genomic R-loops and HIV-1 integration site and showed HIV-1 preferentially integrate into the R-loop rich regions. I demonstrated that HIV-1 directly targets R-loop for integration by using a cell model that can induce site-specific R-loop formation at designated non-human sequence in the host genome together with an extra control of non-R-loop forming sequence, which shows comparable transcriptional activity but disable to form a stable R-loop. I also found that HIV-1 integrase proteins directly bind to R-loop structures in vitro and physically interact with the host genomic R-loops in cells.

CONCLUSION: In this study, I demonstrate that HIV-1 exploits host genomic R-loops for its successive integration and infection. How HIV-1 induces host genomic R-loop formation over diverse host genomic regions remains unknown. However, these results suggesting R-loop as a novel determinant in HIV-1 integration site selection bridge the under explored relationship between gene transcription and HIV-1 integration site determination

1. INTRODUCTION

1.1. Mechanism of HIV-1 integration site selection and HIV-1 infection

Retroviruses take a unique viral life cycle of integrating their reverse-transcribed viral genome into the host genome and thereby cause permanent infection in the host (Figure 1). Retroviral integration considerably impacts a wide range of biological phenomena, including the persistence of fatal human diseases and the shaping of metazoan evolution (Johnson, 2019). Human immunodeficiency virus-1 (HIV)-1 is a representative retrovirus that underlies the global burden of acquired immune deficiency syndrome (AIDS) (Lusic and Siliciano, 2017). The chromosomal landscape of HIV-1 integration plays a critical role in proviral gene expression, persistence of integrated proviruses, and prognosis of antiretroviral therapy (Chen et al., 2017; Einkauf et al., 2022; Jiang et al., 2020). Integration into the host genome is not random and displays distinct preferences for gene-dense regions, where active transcription occurs (Schroder et al., 2002). However, transcription activity is not the sole determinant of the HIV-1 integration site landscape (Lucic et al., 2019). For instance, the most favored region of HIV-1 integration is an intergenic locus, and despite the lower probability of integration, HIV-1 proviruses are observed in non-genic regions (Einkauf et al., 2022; Schroder et al., 2002). These HIV-1 integration into the 'gene desert' regions of the host genome

challenges the traditional understanding of HIV-1 integration site selection mechanism, and are rather more problematic as they are selected to preserve in the host genome during prolonged antiretroviral therapies (Einkauf et al., 2022) (Figure 2). The mechanism by which HIV-1 integrates into gene-silent regions of the genome is not fully understood. This indicates the possibility of there being an undiscovered mechanism or determinant that composes the correct genomic environment for HIV-1 integration.

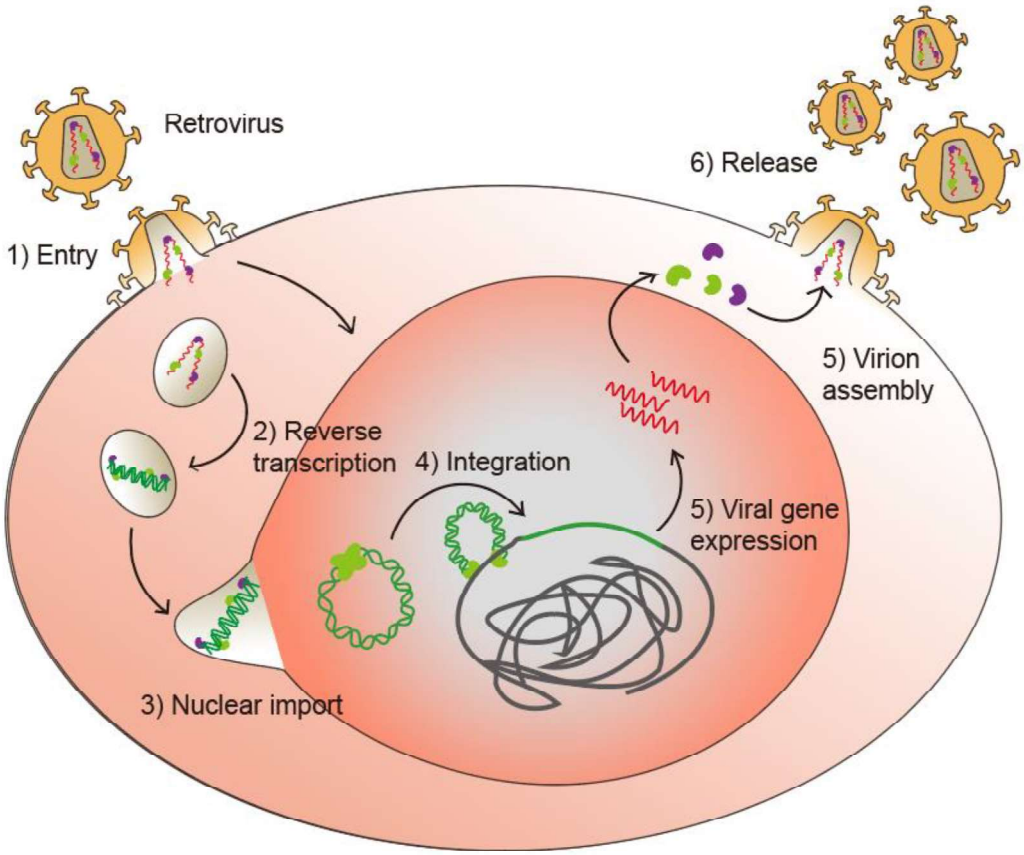
Retrovirus
6) Release
1) Entry
5) Virion
assembly
2) Reverse
transcription
4) Integration
5) Viral gene
expression
3) Nuclear import

Figure 1. The retrovirus life cycle.

Retrovirus virion binds to the host cellular receptor and enter by endocytosis. The internal core is released into host cellular cytoplasm and the two RNA viral genome undergo reverse transcription. The single copy of cDNA viral genome is then transported to the nucleus and the cDNA is integrated into the host cellular chromosome. The integrated genome (provirus) undergoes transcription producing viral RNA genome copies. Viral genes expressed from the provirus are also translated into structural and enzymatic proteins. These viral materials are then assembled into virions that are released from the infected host cell.

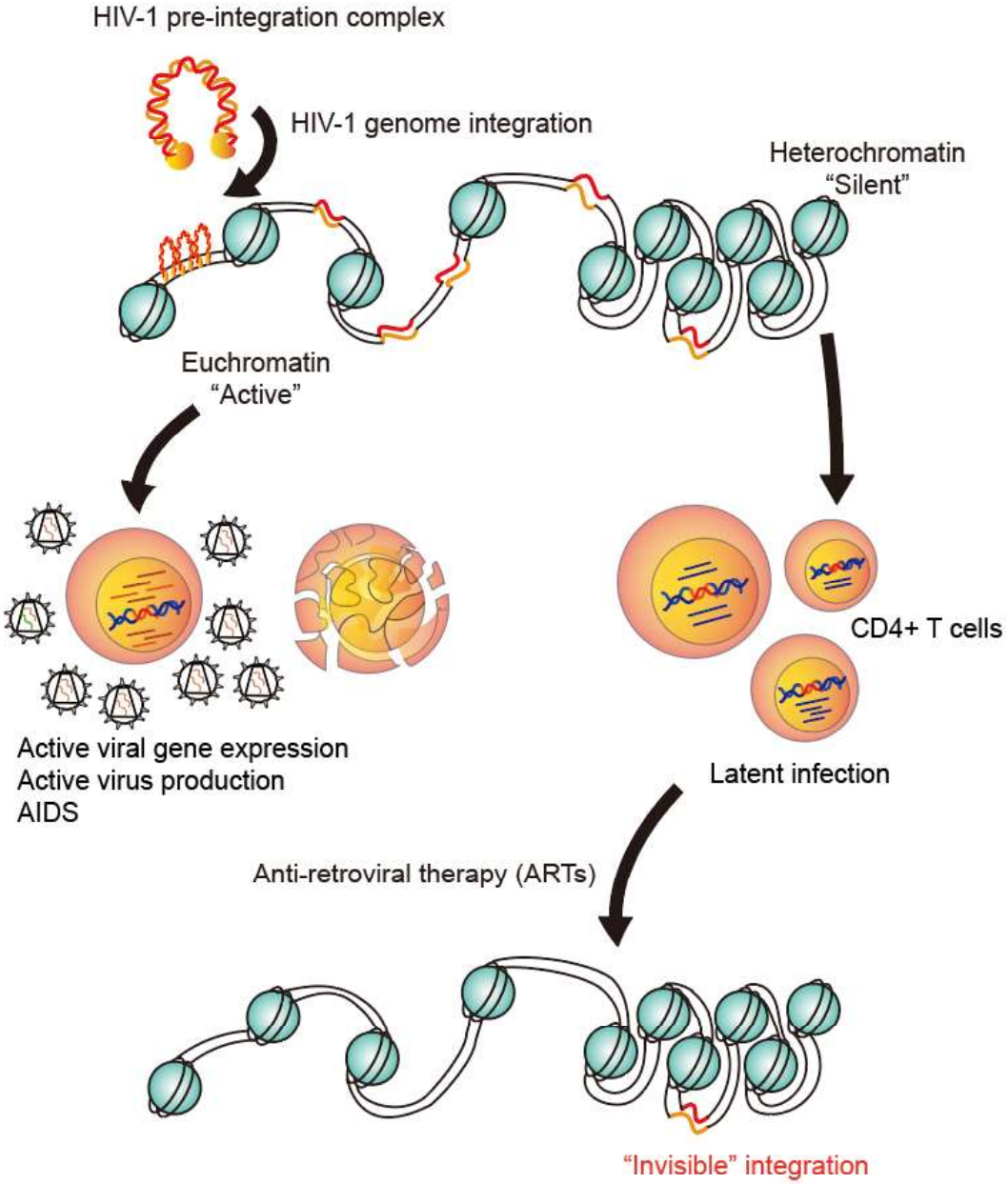
HIV-1 pre-integration complex
HIV-1 genome integration
Heterochromatin
"Silent"
Euchromatin
"Active"
CD4+ T cells
Active viral gene expression
Active virus production
AIDS
Latent infection
Anti-retroviral therapy (ARTs)
"Invisible" integration

Figure 2. Chromosomal landscape of HIV-1 integration and persistence of integrated proviruses.

HIV-1 viral genome integration into the host genomic region of active gene transcription result in active viral gene expression and outbreaks AIDS. HIV-1 also integrates into heterochromatin regions of the host genome where are the gene transcription is limited. The HIV-1 integration into non-genic or silenced genomic regions establishes latent infection and persist disparate of prolonged anti-retroviral therapy.

1.2. R-loop in cellular genome

An R-loop is a three-stranded nucleic acid structure that comprises a DNA-RNA hybrid and displaced strand of DNA, and has long been considered a transcription byproduct (Niehrs and Luke, 2020; Petermann et al., 2022). R-loops are nucleic acid structures that are enriched in actively transcribed genes as they occur naturally during transcription (Hamperl et al., 2017; Niehrs and Luke, 2020), but R-loop formation is not limited to gene body regions and is widespread in the genome (Niehrs and Luke, 2020). As a result of *in trans* R-loop formation, R-loops are also abundant in non-genic regions, such as intergenic regions, repetitive sequences, including transposable elements, centromeres, or telomeres (Arora et al., 2014; Ginno et al., 2012; Lim et al., 2015; Niehrs and Luke, 2020), independently of transcription of the genes harboring the R-loops (Figure 3). Although R-loops are identified as critical intermediates and regulators in various biological processes (Garcia-Muse and Aguilera, 2019; Niehrs and Luke, 2020; Petermann et al., 2022), molecular mechanisms and the role played by cellular R-loops in various pathological contexts remain unrevealed.

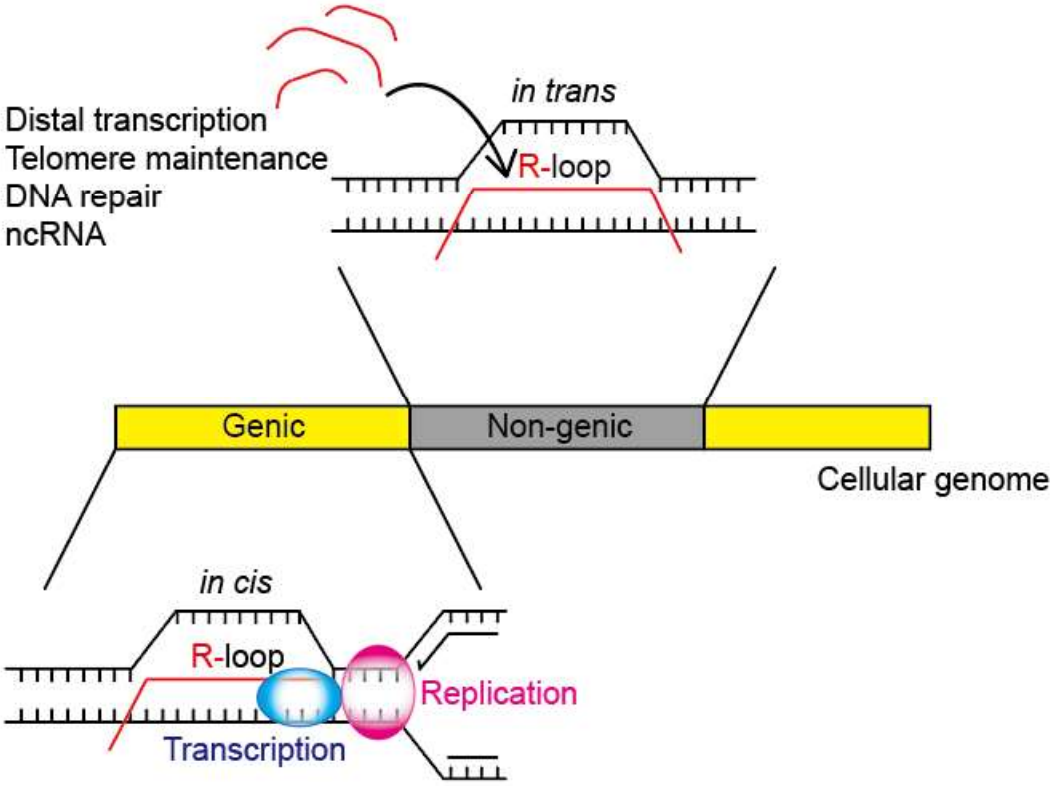

Figure 3. Genome-wide R-loop formation

R-loops are formed genome-wide both in *cis* and in *trans* manner. Transcription activity is the most prevalent predisposing factor of R-loop formation. R-loops are found at promoter regions, transcription termination regions, GC rich regions and at regions of transcription-replication collision. R-loops also form at non-genic regions where distal transcribed RNAs base pair with the complementary DNA strand, telomerase maintenance occur and at the sites of DNA damage. ncRNA, non-coding RNA.

1.3. Host factor and host genomic R-loops in HIV-1 integration site determination

Various host proteins, such as a transcription activator or a protein interacting with active epigenetic markers (Achuthan et al., 2018; Sowd et al., 2016), and genomic elements, such as super enhancers (Lucic et al., 2019), were identified as important host contributors to HIV-1 integration site selection. In fact, these host factors also play roles in R-loop biology. LEDGF/p75 (Cherepanov et al., 2003; Schrijvers et al., 2012; Sowd et al., 2016) and CPSF6 (Achuthan et al., 2018; Sowd et al., 2016) are two decisive host factors that direct HIV-1 integration into gene-dense regions and have recently been identified as potential R-loop binding proteins in DNA-RNA interactome analysis (Cristini et al., 2018) and R-loop proximity proteomics (Mosler et al., 2021), respectively. The Fanconi anemia pathway (Garcia-Rubio et al., 2015; Giannini et al., 2020), a well-known R-loop regulatory pathway, has been proposed as an HIV-1 integration regulatory factor exploited by HIV-1 (Fu et al., 2022). Additionally, R-loop rich regions are frequently associated to open chromatin marks and active enhancers (Chedin, 2016; Sanz et al., 2016), which are also distributed over HIV-1 integration sites (Schroder et al., 2002). In a recent study, non-canonical B-form DNA motifs have been revealed as important factors in HIV-1 integration and provirus reactivation (Ajoge et al., 2022; McAllister et al., 2014). R-loops are prevalent non-canonical B-form DNA structures (Chedin and Benham, 2020). Together, the accumulated

evidence suggests that R-loops potentially play notable roles in HIV-1 integration site selection.

Here, I discovered a novel role of R-loops in the interaction between HIV-1 and its host, specifically in HIV-1 integration. HIV-1-infection induced host genomic R-loops are favored by HIV-1 integration. These results suggest that R-loops are an important composer of host genomic environment for HIV-1 integration site determination and may be a potential target for therapeutic intervention in HIV-1 elimination strategies.

2. MATERIAL AND METHODS

2.1. Cell culture

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

2.2. Virus production and infection

VSV-G-pseudotyped HIV-1 virus stocks were prepared by performing standard polyethylenimine-mediated transfection of HEK293T monolayers with pNL4-3 ⊿Env EGFP (NIH AIDS Reagent Program 11100) or pNL4-3. Luc.R-E (NIH AIDS Reagent Program, 3418) along with pVSV-G at a ratio of 5:1. HIV-IN-EGFP virions were produced by performing polyethylenimine-mediated transfection of HEK293T cells with 6 μg of pVpr-IN-EGFP, 6 μg of HIV-1 NL4-3 non-infectious molecular clone (pD64E; NIH AIDS Reagent Program 10180), and 1 μg of pVSV-G. The cells were incubated for 4 h before the medium was replaced with fresh complete medium. Virion-containing supernatants were collected after 48 h, filtered through a 0.45-μm syringe filter, and pelleted using the Lenti-X Concentrator (631232; Clontech) according to the

manufacturer’s instructions. The multiplicity of infection (MOI) of virus stocks was determined by transducing a known number of HeLa cells with a known amount of virus particles and then counting GFP-positive cells using flow cytometry. For luciferase reporter HIV-1 virus, reverse transcriptase mutant virus (D185A/D186A), or integrase mutant virus (D116N), the HIV-1 p24 antigen content in viral stock were quantified using the HIV1 p24 ELISA kit (Abcam, ab218268), according to the manufacturer’s instruction. For virus infection, HeLa cells were seeded at a density of 0.5-4 × 10^5 cells/mL on the day before infection. The culture medium was replaced with fresh complete culture medium 2 hpi. The infected cells were washed twice with PBS and harvested at the indicated time points. Jurkat cells were seeded at a density of 1× 10^6 cells/mL and inoculated with 300ng/p24 capsid antigen. The plates were centrifuged at 1000 *g* at 30° C for 1 h. The medium was replaced with fresh RPMI 2 h after infection.

2.3. Primary cell isolation, culture, T cell activation, and infection

For $CD4^+$ T cells isolation, human PBMC (ST70025, STEMCELL Technologies) was mixed and incubated with MACS CD4 MicroBeads (130-045-101, Miltenyi Biotec) and FITC-conjugated mouse anti-CD4 (561005, BD Bioscience) according to the manufacturer’s instructions. Then the $CD4^+$ T cells were enriched by using LS Columns (130-042-401, Miltenyi Biotec) and MidiMACS Separator

(130-042-302, Miltenyi Biotec). The efficiency of magnetic separation was analyzed by using Flow-Activated Cell Sorter Canto II (BD Bioscience) and Flowjo software (Flowjo).

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

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

2.4. DRIPc-seq library construction

DRIP followed by library preparation, next-generation sequencing, and peak calling were performed as described earlier(Sanz and Chedin, 2019). Briefly, the corresponding cells were harvested and their gDNA was extracted. The extracted nucleic acids were

fragmented using a restriction enzyme cocktail with BsrB I (NEB, R0102S), HindIII (NEB, R0136L), Xba I (NEB, R0145L), and EcoRI (NEB, R3101L) overnight at 37° C. Half of the fragmented nucleic acids were digested with RNase H (New England Biolabs) overnight at 37° C to serve as a negative control. The digested nucleic acids were cleaned using standard phenol-chloroform extraction and resuspended in DNase/RNase-free water. DNA-RNA hybrids were immunoprecipitated from total nucleic acids using mouse anti-DNA-RNA hybrid S9.6 (Kerafast, ENH001) DRIP binding buffer and incubated overnight at 4° C. Dynabeads Protein A (Invitrogen, 10001D) was used to pull down the DNA-antibody complexes by incubation for 4 h at 4° C. The isolated complexes were washed twice with DRIP binding buffer before elution. For elution, the isolated complexes were incubated in an elution buffer containing proteinase K for 45 min at 55 ° C. Subsequently, DNA was purified using the standard phenol-chloroform extract method and subjected to DNase I (Takara, 2270 B) treatment and reverse transcription for DRIPc-seq library construction. DRIPc-seq was performed in biological replicates. Table 1 shows details of the oligonucleotides used for DRIPc-seq library construction. DRIPc-seq libraries were analyzed using 150 bp paired-end sequencing on a HiSeqX Illumina instrument.

2.5. Immunofluorescence microscopy

For immunofluorescence assays of S9.6 nuclear signals, when indicated, the cells were pre-extracted with cold 0.5% NP-40 for 3 min on ice. Cells were fixed with 100% ice-cold methanol for 10 min on ice and then incubated with 100% ice-cold acetone for 1 min. The slides were washed with 1× PBS and incubated with or without 60 U/mL RNase H (M0297S, NEB) at 37° C for 36 h or left untreated. The slides were subsequently briefly rinsed thrice with 2% BSA/0.05% Tween (in PBS) and incubated with mouse anti-DNA-RNA hybrid S9.6 (Kerafast, ENH001; 1:100) and rabbit anti-nucleolin (Abcam, ab22758; 1:300) in 2% BSA/0.05% Tween (in PBS) for 4 h at 4° C. The slides were then washed with 2% BSA/0.05% Tween (in PBS) and incubated with goat anti-rabbit AlexaFluor-488-conjugated (Invitrogen, A-11008) and goat anti-mouse AlexaFluor-568-conjugated (Molecular Probes, A11004) secondary antibodies (1:200) for 2 h at room temperature. The slides were then washed with 2% BSA/0.05% Tween (in PBS) and mounted using the ProLong Gold AntiFade reagent (Invitrogen). Images were obtained using an inverted microscope Nikon Eclipse Ti2, equipped with a 1.45 numerical aperture, plan apochromat lambda 100× oil objective, and an scientific complementary metal-oxide-semiconductor camera (Photometrics prime 95 B 25 mm). For each field of view, images were obtained with DAPI395, GFP488, and Alexa594 channels using the NIS-Elements software. For quantification analysis, binary masks of nuclei and nucleoli were generated using the ROI manager and auto local thresholding using the ImageJ software. The intensity

of nuclear signals for DNA-RNA hybrids and nucleolin was then quantified. The final DNA-RNA hybrid signals in the nucleus were calculated by subtracting the nucleolin signals from the DNA-RNA hybrid signals.

2.6. HIV-1 integration site sequencing library construction

HIV-1 integration site sequencing library construction was performed as described earlier (Achuthan et al., 2018; Sowd et al., 2016). Summarily, HeLa cells were infected with VSV-G-pseudotyped HIV-1 NL4-3 ⊿Env EGFP virus at a MOI of 0.6 and harvested 5 days post infection. gDNA was isolated using a DNA purification kit (Qiagen, 51106), according to the manufacturer's instructions. gDNA (10 μg) was digested overnight at 37° C with 100 U each of the restriction endonucleases MseI (NEB, R0525L) and BglII (NEB, R0144L). Linker oligonucleotides, which were compatible for ligation with the MseI-generated DNA ends, were ligated with gDNA overnight at 12° C in reactions containing 1.5 μM ligated linker, 1 μg fragmented DNA, and 800 U T4 DNA ligase (NEB, M0202S). Viral LTR-host DNA junctions were amplified using semi-nested PCR with a unique linker-specific primer and LTR primers. The second round of PCR was carried out with primers binding to the LTR and the linkers for next-generation sequencing. Two PCRs were performed in parallel for the first round of PCR and five PCRs were performed in parallel for the second round of PCR to

enhance library diversity. Table 2 presents details of the oligonucleotides used for HIV-1 integration site sequencing library construction. HIV-1 integration site sequencing was performed in biological replicates. Integration site libraries were analyzed using 150 bp paired-end sequencing on a HiSeqX Illumina instrument.

2.7. pgR-rich and -poor cell line generation with piggyBac transposition

I adapted and modified an elegantly designed episomal system that induces defined R-loops with controlled transcription levels(Hamperl et al., 2017) for R-loop-forming or non-R-loop-forming sequence subcloning into the piggyBac transposon vector. HeLa cells were seeded at a density of 5×10^4 cells/ml in a 6-well plate. The next day, cells were transfected with 0.2 μg of Super PiggyBac Transposase Expression Vector (System Biosciences, PB210PA-1) and 0.2, 1, or 2 μg of transposon vectors with appropriate "cargo" sub cloned using Lipofectamine 3000 (Invitrogen) according to the manufacturer's instructions. After 3 days, the cells were treated with 10 μg/ml blasticidin S (Gibco, A1113903) for selection. Cells with positive integrants for more than 7 days were validated using immunoblotting or RT-qPCR following treatment with DOX. Jurkat cells were seeded at a density of 8×10^5 cells/ml in a 6-well plate and transfected with 0.2 μg of transposase and 1 μg of corresponding transposon vectors with Lipofectamine 3000, like HeLa cells. After 3 days, the cells were

treated with 10 μg/ml blasticidin S (Gibco, A1113903) for selection. For each passage, cells were cushioned onto Ficoll-Pacque (Cytiva, 17144002) to separate live cells from dead cell debris. The cells over the cushion were washed with PBS and incubated in cell culture medium with 10 μg/ml of blasticidin for further selection for at least 14 days. Cells with positive integrants were validated by immunoblotting after treatment with DOX. Quantification of successfully integrated piggyBac transposons was performed using a piggyBac qPCR copy number kit (System Biosciences, PBC100A-1) according to the manufacturer's instructions.

2.8. Co-immunoprecipitation of DNA-RNA hybrid

DNA-RNA hybrid immunoprecipitation was performed as described earlier (Cristini et al., 2018). Summarily, non-crosslinked HeLa cells transfected with the pFlag-IN codon-optimized plasmid were lysed in 85 mM KCl, 5 mM PIPES (pH 8.0), and 0.5% NP-40 for 10 min on ice, and then, the lysates were centrifuged at 750 *g* for 5 min to pellet the nuclei. The pelleted nuclei were resuspended in sodium deoxycholate, SDS, and sodium lauroyl sarcosinate in RSB buffer and were sonicated for 10 min (Diagenode Bioruptor). Extracts were then diluted (1:4 in RSB + T buffer) and subjected to immunoprecipitation with the S9.6 antibody overnight at 4° C. Antibody-bound complexes were incubated with Protein A Dynabeads (Invitrogen) for 4 h at 4° C for immunoprecipitation. Normal mouse IgG antibodies (Santa Cruz, sc-2025) were used as negative controls. RNase A

(Thermo Scientific, EN0531) was added during immunoprecipitation at 0.1 ng RNase A per μg gDNA. Beads were washed four times with RSB + T; twice with RSB, and eluted either in 2× LDS (Novex, NP0007), 100 mM DTT for 10 min at 70° C (for western blot), or 1% SDS and 0.1 M $NaHCO_3$ for 30 min at room temperature (for DNA–RNA hybrid dot blot).

For co-immunoprecipitation of DNA-RNA hybrids with RNase H treatment (Cristini et al., 2018), gDNA containing RNA-DNA hybrids was isolated from HeLa cells transfected with a pFlag-IN codon-optimized plasmid using a QIAmp DNA Mini Kit (Qiagen, 51304). gDNA was sonicated for 10 min (Diagenode Bioruptor) and then treated with 5.5 U RNase H (NEB, M0297) per μg of DNA overnight at 37 ° C. A fraction of gDNA was stored as "nucleic acid input" for dot blot analysis. gDNA was cleaned using standard phenol-chloroform extraction, resuspended in DNase/RNase-free water, enriched for DNA-RNA hybrids using immunoprecipitation with the S9.6 antibody (overnight at 4° C), isolated with Protein A Dynabeads (Invitrogen; 4 h at 4° C), washed with RSB+T. The immunoprecipitated complexes were incubated with nuclear extracts of HeLa cells transfected with the pFlag-IN codon-optimized plasmid for 2 h at 4° C with diluted HeLa nuclear extracts. The cell lysate containing proteins were pre-treated with 0.1 mg/ml RNase A (Thermo Scientific, EN0531) for 1 h at 37° C to degrade all RNA-DNA hybrids, and the excess of RNase A was blocked by adding 200 U of SUPERase in RNase inhibitor (Invitrogen, AM2694) for

immunoprecipitation. In addition, 100 μL fraction of diluted and RNase A pre-treated extracts prior to immunoprecipitation was stored as "protein input" for western blotting. Beads were washed four times with RSB + T; twice with RSB, and eluted either in 2× LDS (Novex, NP0007), 100 mM DTT for 10 min at 70° C (for western blot), or 1% SDS, and 0.1 M $NaHCO_3$ for 30 min at room temperature (for DNA-RNA hybrid dot blot).

2.9. Recombinant Sso7d-IN protein purification

Sso7d-integrase active site mutant E152Q was expressed in Escherichia coli BL21-AI and purified as previously described (Passos et al., 2017). Briefly, Sso7d-IN (E152Q) expressed BL21-AI cells were lysis in lysis buffer (20 mM HEPES pH 7.5, 2 mM 2-mercaptoethanol, 1 M NaCl, 10% (w/v) glycerol, 20 mM imidazole, 1 mg RNase A, and 1000 U DNase I) and purified by nickel affinity chromatography (Qiagen, 30210). Protein were first loaded on HeparinHP column (GE Healthcare) equilibrated with equilibrated with 20 mM Tris, pH 8.0, 0.5 mM TCEP, 200 mM NaCl, 10% glycerol prior to the size exclusion chromatography. Proteins were eluted with a linear gradient of NaCl from 200 mM to 1 M. Eluted fractions were pooled and then separated on a Superdex-200 PC 10/300 GL column (GE Healthcare) equilibrated with 20 mM Tris pH 8.0, 0.5 mM TCEP, 500 mM NaCl and 6% (w/v) glycerol. The purified protein was flash-frozen in liquid nitrogen and stored at -80° C.

2.10. Electrophoretic mobility shift assay for R-loop

binding of Sso7d-IN

To test the binding affinity of Sso7d-tagged HIV-1 IN to different types of nucleic acid substrates, I prepared R-loop, dsDNA, RNA-DNA hybrid with exposed ssDNA (R:D+ssDNA), RNA-DNA hybrid (Hybrid), ssDNA, and ssRNA by annealing different combinations of Cy3, Cy5 or non-labeled oligonucleotides following the previous protocol (Nguyen et al., 2017). 10 nM of DNA substrate was incubated with Sso7d-IN at different concentrations in assembly buffer (20 mM HEPES pH 7.5, 5 mM $CaCl_2$, 8 mM 2-mercaptoethanol, 4 uM $ZnCl_2$, 100 mM NaCl, 25% (w/v) glycerol and 50 mM 3-(Benzyldimethylammonio) propanesulfonate (NDSB-256)), for 1 h at 30° C then incubated for 15 min on ice. All the reactants were run on 4.5% non-denaturing PAGE in 1× TBE and then Cy3 or Cy5 fluorescence signal was imaged by ChemiDoc MP imaging system (Bio-Rad). Table 3 presents details of the oligonucleotide sequence used for EMSA.

2.11. Proximity Ligation Assay (PLA)

For PLA, HeLa cells were grown on coverslips and infected with HIV-IN-EGFP virions. At 6 hpi, cells were pre-extracted with cold 0.5% NP-40 for 3 min on ice. The cells were fixed with 4% paraformaldehyde in PBS for 15 min at 4 ° C. The cells were then blocked with 1× blocking solution (Merck, DUO92102) for 1 h at 37° C in a humidity chamber. After blocking, cells were incubated with the following primary antibodies overnight at 4° C for S9.6-

HIV-1-IN_PLA: mouse anti-DNA-RNA hybrid S9.6 (1:250; Kerafast, ENH001) and rabbit anti-GFP (1:500; Abcam, ab6556). The following day, after washing with once with buffer A twice (Merck, DUO92102), cells were incubated with pre-mixed Duolink PLA plus (anti-mouse) and PLA minus probes (anti-rabbit) antibodies for 1 h at 37° C. The subsequent steps in the proximal ligation assay were performed using the Duolink PLA Fluorescence kit (Sigma) according to the manufacturer's instructions. To obtain images, the mounted specimens were visually scanned and representative images were acquired using a Zeiss LSM 710 laser scanning confocal microscope (Carl Zeiss). The number of intranuclear PLA puncta was quantified using the ImageJ software. For each biological replicate and experiment, a PLA with a single antibody was performed as a negative control under the same conditions.

2.12. DRIPc-Seq data processing and peak calling

DRIPc-seq reads were quality-controlled using FastQC v0.11.9(Andrews, 2010), and sequencing adapters were trimmed using Trim Galore! v0.6.6 (Felix Krueger, 2021) based on Cutadapt v2.8 (Martin, 2011). Trimmed reads were aligned to the hg38 reference genome using bwa v0.7.17-r1188(Li and Durbin, 2009). Read deduplication and peak calling were performed using MACS v2.2.7.1(Zhang et al., 2008). Because R-loops appear as both narrow and broad peaks in DRIPc-seq read alignment owing to its variable

length, two independent "MACS2 callpeak" runs were performed for narrow and broad peak calling. The narrow and broad peaks were merged using Bedtools v2.26.0(Quinlan and Hall, 2010). To increase the sensitivity of DRIPc-seq peak identification, peaks were called after pooling the two biological replicates of the DRIPc-seq sequencing data for each condition.

2.13. Consensus R-loop peak calling

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

2.14. HIV-1 integration site sequencing data processing

Quality control of HIV-1 integration site-sequencing reads was performed using FastQC v0.11.9. To discard primers and linkers specific for integration site-sequencing from reads, Cutadapt v2.8 with the following option: "-u 49 -U 38 --minimum-length 36 -

-pair-filter any --action trim -q0,0 -a linker -A TGCTAGAGATTTTCCACACTGACTGGGTCTGAGGG -A GGGTCTGAGGG --no-indels --overlap 12" were used. This allowed the first position of the read alignment to directly represent the genomic position of HIV-1 integration. Processed reads were aligned to the hg38 reference genome using bwa v0.7.17-r1188, and integration sites were identified using an in-house Python script. Genomic positions supported by more than five read alignments were regarded as HIV-1 integration sites. For Jurkat cells, integration site sequencing data of HIV-1 infected wild type Jurkat cells from SRR12322252 (Li et al., 2020b) were adopted.

2.15. Co-localization analysis of R-loops and integration sites

Enrichment of integration sites near the R-loop peaks was tested using a randomized permutation test. Randomized R-loop peaks were generated using "bedtools shuffle" command, thus preserving the number and the length distribution of the R-loop peaks during the randomization process. Similarly, integration sites were randomized using the "bedtools shuffle" command. Randomization was performed 100 times. ENCODE blacklist regions (Amemiya et al., 2019) were excluded while shuffling the R-loops and integration sites to exclude inaccessible genomic regions from the analysis. For each of the observed (or randomized) integration sites, the closest observed (or randomized) R-loop peak and the corresponding

genomic distance were identified using the "bedtools closest" command. The distribution of the genomic distances was displayed to show the local enrichment of integration sites in terms of the increased proportion of integration sites within the 30-kb window centered on R-loops compared to their randomized counterparts.

2.16. DNA plasmid construction and transfection

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

2.17. DNA-RNA hybrid dot blotting

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

2.18. DRIP-qPCR

DRIP was performed as described for the construction of the DRIPc-seq library. After the elution of isolated complexes, nucleic acids were purified using the standard phenol-chloroform extract method and used for qPCR. Table 4 presents details of the primer sequences used for DRIP-qPCR analysis.

2.19. RNA-seq library construction

For RNA-seq, HeLa cells were infected with VSV-G-pseudotyped HIV-1 NL4-3 ⊿Env EGFP virus at a MOI of 0.6 and harvested at 0, 3, 6, and 12 hpi. Sequencing was performed with biological replicates. Total RNA was extracted using TRIzol reagent (Invitrogen), according to the manufacturer's instructions. An mRNA sequencing library was constructed using Illumina adaptors harboring p5 and p7 sequences and Rd1 SP and Rd2 SP sequences. Sequencing was performed using the HiSeq2500 system (Illumina).

2.20. Luciferase assay

HeLa cells infected with VSV-G-pseudotyped pNL4-3.Luc.R-E HIV-1 viruses were harvested at 48 hpi, and luminescence was measured using the Dual-Luciferase Reporter Assay System (Promega) according to the manufacturer's instructions. Briefly, 250 μl of passive lysis buffer was used to lyse cells for each sample,

20 μl of the lysate was mixed with 100 μl of the Luciferase Assay Reagent II, and the luminescence of firefly luciferase was measured using a microplate luminometer (Berthold). The luminescence signal were normalized with total protein content, measured by BCA assay.

2.21. Quantitative real-time PCR (qPCR)

For RT (reverse transcription)-qPCR, 1 μg of RNA was reverse-transcribed using the ReverTra Ace qPCR RT Kit (TOYOBO) following the manufacturer's instructions. For qPCR, DNA extracts were prepared using a DNA purification kit (Qiagen, 51106) according to the manufacturer's instructions. Equivalent amounts of purified gDNA from each sample were analyzed using qPCR. qPCR was performed using TOPreal qPCR PreMIX (Enzynomics, RT500M). The reactions were performed in duplicate or triplicate for technical replicates. PCR was performed using the iCycler iQ real-time PCR detection system (Bio-Rad). All the primers used for qPCR are listed in Table 4.

2.22. Chromatin immunoprecipitation (ChIP) of Flag-tagged codon-optimized integrase

For chromatin immunoprecipitation of Flag-tagged E152A mutant codon-optimized integrase proteins followed by RT-qPCR analysis, HeLa cells transfected with piggyBac transposon vector were induced by 1 ug/ml DOX treatment for Flag-tagged E152A mutant

codon-optimized integrase expression and infected with pNL4-3 ⊿Env EGFP virus at MOI of 0.6, and harvested at 6 hpi. For chromatin immunoprecipitation of Flag-tagged codon-optimized integrase proteins, cells were prepared by transfecting HeLa cells with pFlag-IN codon-optimized plasmid. Nuclear fraction was isolated from each sample by cell lysis with 85 mM KCl, 5 mM PIPES (pH 8.0), and 0.5% NP-40 for 10 min on ice, and then, the lysates were centrifuged at 750 *g* for 5 min to pellet the nuclei. The pelleted nuclei were resuspended in sodium deoxycholate, SDS, and sodium lauroyl sarcosinate in RSB buffer and were sonicated for 10 min (Diagenode Bioruptor). Extracts were then diluted (1:4 in RSB + T buffer) and subjected to immunoprecipitation with the FLAG M2 (Sigma, F3165) antibody overnight at 4° C. Antibody-bound complexes were incubated with Protein G Dynabeads (Invitrogen) for 4 h at 4° C for immunoprecipitation. Normal mouse IgG antibodies (Santa Cruz, sc-2025) were used as negative controls. Beads were washed four times with RSB + T; twice with RSB, and eluted either in 2× LDS (Novex, NP0007), 100 mM DTT for 10 min at 70° C (for western blot), or 1% SDS and 0.1 M $NaHCO_3$ for 30 min at room temperature (for DNA-RNA hybrid dot blot). If indicated, nucleic acid elutes were treated with 10U of DNase I (Thermo Fisher Scientific, EN0525) at 37° C for 30 min. The reaction was inactivated by adding final concentration of 5mM EDTA and incubating at 75° C for 15 min. RNA fraction was precipitated by ethanol precipitation and reverse transcribed with iScript™ cDNA Synthesis Kit (Bio-Rad, 1708890) according to the

manufacturer’s instructions.

2.23. Immunoblotting

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

2.24. RNA-seq data processing

RNA-seq reads were quality-controlled and adapter-trimmed as in

DRIPc-seq processing. To quantify the expression levels of protein-coding genes, processed reads were aligned to the hg38 reference genome with GENCODE v37 gene annotation (Frankish et al., 2021) using STAR v2.7.3a (Dobin et al., 2013). Gene expression quantification was performed using RSEM v1.3.1. To quantify the expression levels of transposable elements (TEs), TEtranscripts v2.2.1 (Jin et al., 2015) was used. Processed reads were first aligned to the hg38 reference genome using GENCODE v37 and RepeatMasker TE annotation using STAR v2.7.3a. In this case, STAR options were modified as follows to utilize multimapping reads in downstream analyses: "--outFilterMultimapNmax 100 --winAnchorMultimapNmax 100 --outMultimapperOrder random --runRNGseed 77 --outSAMmultNmax 1 --outFilterType BySJout --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000". Expression levels of TEs were quantified as read counts with the "TEcount" command.

2.25. Genome annotations

All bioinformatic analyses were performed using the hg38 reference genome and GENCODE v37 gene annotation. Promoters were defined as a 2-kb region centered at the transcription start sites of the APPRIS principal isoform of protein-coding genes. TTS regions were defined as the 2-kb region centered at the 3′ terminals of protein-coding transcripts. CpG island annotations were downloaded

from the UCSC table browser. CpG shores were defined as 2-kb regions flanking CpG islands, excluding the regions overlapping with CpG islands. Similarly, CpG shelves were defined as 2-kb regions flanking the stretch of CpG islands and shores while excluding the regions overlapping with CpG islands and shores. Annotations for LINE, SINE, and LTR were extracted from the RepeatMasker track in the UCSC table browser.

2.26. Identification of viral sequencing reads in DRIPc-seq

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

2.27. Code availability

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

Table 1. 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, Jurkat 0hpi Input replicate 1
Index Adapter 2	GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCCGTCTTCTGCTTG	HeLa, Jurkat 3hpi Input replicate 1
Index Adapter 3	GATCGGAAGAGCACACGTCTGAACTCCAGTCACTTAGGCATCTCGTATGCCGTCTTCTGCTTG	HeLa, Jurkat 6hpi Input replicate 1
Index Adapter 4	GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATGCCGTCTTCTGCTTG	HeLa, Jurkat 12hpi Input replicate 1
Index Adapter 5	GATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGCCGTCTTCTGCTTG	HeLa, Jurkat 0hpi RNH-IP replicate 1
Index Adapter 6	GATCGGAAGAGCACACGTCTGAACTCCAGTCACGCCAATATCTCGTATGCCGTCTTCTGCTTG	HeLa, Jurkat 3hpi RNH-IP replicate 1
Index Adapter 7	GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGATCATCTCGTATGCCGTCTTCTGCTTG	HeLa, Jurkat 6hpi RNH-IP replicate 1
Index Adapter 8	GATCGGAAGAGCACACGTCTGAACTCCAGTCACACTTGAATCTCGTATGCCGTCTTCTGCTTG	HeLa, Jurkat 12hpi RNH-IP replicate 1
Index Adapter 9	GATCGGAAGAGCACACGTCTGAACTCCAGTCACGATCAGATCTCGTATGCCGTCTTCTGCTTG	HeLa, Jurkat 0hpi RNH+IP replicate 1
Index Adapter 10	GATCGGAAGAGCACACGTCTGAACTCCAGTCACTAGCTTATCTCGTATGCCGTCTTCTGCTTG	HeLa, Jurkat 3hpi RNH+IP replicate 1
Index Adapter 11	GATCGGAAGAGCACACGTCTGAACTCCAGTCACGGCTACATCTCGTATGCCGTCTTCTGCTTG	HeLa, Jurkat 6hpi RNH+IP replicate 1
Index Adapter 12	GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTTGTAATCTCGTATGCCGTCTTCTGCTTG	HeLa, Jurkat 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
Index Adapter 13	GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGTCAAATCTCGTATGCCGTCTTCTGCTTG	Jurkat 0hpi Input replicate 2
Index Adapter 14	GATCGGAAGAGCACACGTCTGAACTCCAGTCACAGTTCCATCTCGTATGCCGTCTTCTGCTTG	Jurkat 3hpi Input replicate 2
Index Adapter 15	GATCGGAAGAGCACACGTCTGAACTCCAGTCACATGTCAATCTCGTATGCCGTCTTCTGCTTG	Jurkat 6hpi Input replicate 2
Index Adapter 16	GATCGGAAGAGCACACGTCTGAACTCCAGTCACCCGTCCATCTCGTATGCCGTCTTCTGCTTG	Jurkat 12hpi Input replicate 2
Index Adapter 17	GATCGGAAGAGCACACGTCTGAACTCCAGTCACGTAGAGATCTCGTATGCCGTCTTCTGCTTG	Jurkat 0hpi RNH-IP replicate 2

Index Adapter 18	GATCGGAAGAGCACACGTCTGAACTCCAGTCACG TCCGCATCTCGTATGCCGTCTTCTGCTTG	Jurkat 3hpi RNH-IP replicate 2
Index Adapter 19	GATCGGAAGAGCACACGTCTGAACTCCAGTCACG TGAAAATCTCGTATGCCGTCTTCTGCTTG	Jurkat 6hpi RNH-IP replicate 2
Index Adapter 20	GATCGGAAGAGCACACGTCTGAACTCCAGTCACG TGGCCATCTCGTATGCCGTCTTCTGCTTG	Jurkat 12hpi RNH-IP replicate 2
Index Adapter 21	GATCGGAAGAGCACACGTCTGAACTCCAGTCACG TTTCGATCTCGTATGCCGTCTTCTGCTTG	Jurkat 0hpi RNH+IP replicate 2
Index Adapter 22	GATCGGAAGAGCACACGTCTGAACTCCAGTCACC GTACGATCTCGTATGCCGTCTTCTGCTTG	Jurkat 3hpi RNH+IP replicate 2
Index Adapter 23	GATCGGAAGAGCACACGTCTGAACTCCAGTCACG AGTGGATCTCGTATGCCGTCTTCTGCTTG	Jurkat 6hpi RNH+IP replicate 2
Index Adapter 24	GATCGGAAGAGCACACGTCTGAACTCCAGTCACG GTAGCATCTCGTATGCCGTCTTCTGCTTG	Jurkat 12hpi RNH+IP replicate 2
Index Adapter 25	GATCGGAAGAGCACACGTCTGAACTCCAGTCACA CTGATATCTCGTATGCCGTCTTCTGCTTG	CD4 0hpi Input donor 1
Index Adapter 26	GATCGGAAGAGCACACGTCTGAACTCCAGTCACA TGAGCATCTCGTATGCCGTCTTCTGCTTG	CD4 3hpi Input donor 1
Index Adapter 27	GATCGGAAGAGCACACGTCTGAACTCCAGTCACA TTCCTATCTCGTATGCCGTCTTCTGCTTG	CD4 6hpi Input donor 1
Index Adapter 28	GATCGGAAGAGCACACGTCTGAACTCCAGTCACC AAAAGATCTCGTATGCCGTCTTCTGCTTG	CD4 12hpi Input donor 1
Index Adapter 29	GATCGGAAGAGCACACGTCTGAACTCCAGTCACC AACTAATCTCGTATGCCGTCTTCTGCTTG	CD4 0hpi RNH-IP donor 1
Index Adapter 30	GATCGGAAGAGCACACGTCTGAACTCCAGTCACC ACCGGATCTCGTATGCCGTCTTCTGCTTG	CD4 3hpi RNH-IP donor 1
Index Adapter 31	GATCGGAAGAGCACACGTCTGAACTCCAGTCACC ACGATATCTCGTATGCCGTCTTCTGCTTG	CD4 6hpi RNH-IP donor 1
Index Adapter 32	GATCGGAAGAGCACACGTCTGAACTCCAGTCACC ACTCAATCTCGTATGCCGTCTTCTGCTTG	CD4 12hpi RNH-IP donor 1
Index Adapter 33	GATCGGAAGAGCACACGTCTGAACTCCAGTCACC AGGCGATCTCGTATGCCGTCTTCTGCTTG	CD4 0hpi RNH+IP donor 1
Index Adapter 34	GATCGGAAGAGCACACGTCTGAACTCCAGTCACC ATGGCATCTCGTATGCCGTCTTCTGCTTG	CD4 3hpi RNH+IP donor 1
Index Adapter 35	GATCGGAAGAGCACACGTCTGAACTCCAGTCACC ATTTTATCTCGTATGCCGTCTTCTGCTTG	CD4 6hpi RNH+IP donor 1
Index Adapter 36	GATCGGAAGAGCACACGTCTGAACTCCAGTCACC CAACAATCTCGTATGCCGTCTTCTGCTTG	CD4 12hpi RNH+IP donor 1
Index Adapter 37	GATCGGAAGAGCACACGTCTGAACTCCAGTCACC GGAATATCTCGTATGCCGTCTTCTGCTTG	CD4 0hpi Input donor 2
Index Adapter 38	GATCGGAAGAGCACACGTCTGAACTCCAGTCACC TAGCTATCTCGTATGCCGTCTTCTGCTTG	CD4 3hpi Input donor 2
Index Adapter 39	GATCGGAAGAGCACACGTCTGAACTCCAGTCACC TATACATCTCGTATGCCGTCTTCTGCTTG	CD4 6hpi Input donor 2
Index Adapter 40	GATCGGAAGAGCACACGTCTGAACTCCAGTCACC TCAGAATCTCGTATGCCGTCTTCTGCTTG	CD4 12hpi Input donor 2
Index Adapter 41	GATCGGAAGAGCACACGTCTGAACTCCAGTCACG ACGACATCTCGTATGCCGTCTTCTGCTTG	CD4 0hpi RNH-IP donor 2
Index Adapter 42	GATCGGAAGAGCACACGTCTGAACTCCAGTCACT AATCGATCTCGTATGCCGTCTTCTGCTTG	CD4 3hpi RNH-IP donor 2
Index Adapter 43	GATCGGAAGAGCACACGTCTGAACTCCAGTCACT ACAGCATCTCGTATGCCGTCTTCTGCTTG	CD4 6hpi RNH-IP donor 2
Index Adapter 45	GATCGGAAGAGCACACGTCTGAACTCCAGTCACT CATTCATCTCGTATGCCGTCTTCTGCTTG	CD4 12hpi RNH-IP donor 2
Index Adapter 46	GATCGGAAGAGCACACGTCTGAACTCCAGTCACT CCCGAATCTCGTATGCCGTCTTCTGCTTG	CD4 0hpi RNH+IP donor 2
Index Adapter 47	GATCGGAAGAGCACACGTCTGAACTCCAGTCACT CGAAGATCTCGTATGCCGTCTTCTGCTTG	CD4 3hpi RNH+IP donor 2
Index Adapter 48	GATCGGAAGAGCACACGTCTGAACTCCAGTCACT CGGCAATCTCGTATGCCGTCTTCTGCTTG	CD4 6hpi RNH+IP donor 2
Index Adapter 49	GATCGGAAGAGCACACGTCTGAACTCCAGTCACA ACAACATCTCGTATGCCGTCTTCTGCTTG	CD4 12hpi RNH+IP donor 2

Table 2. Oligonucleotides used for HIV-1 integration site sequencing library construct.

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	GTAATACGACTCACTATAGGGCCTCCG CTTAAGGGAC	replicate 1 5dpi Linker long / replicate 1 pgR-poor DOX- Linker long / CD4+ donor 1 Linker long
AE6382	CAAGCAGAAGACGGCATACGAGATCGG TCTCGGCATTCCTGCTGAACCGCTCTT CCGATCTGTAATACGACTCACTATAGG GC	replicate 1 5dpi Linker primer / replicate 1 pgR-poor DOX- Linker primer / CD4+ donor 1 Linker primer
AE6404	AATGATACGGCGACCACCGAGATCTAC ACTCTTTCCCTACACGACGCTCTTCCG ATCTCGATGTGAGATCCCTCAGACCCT TTTAGTCAG	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	GTAATACGACTCACTATAGGGCCTCCG CTTAAGGGAC	replicate 2 5dpi Linker long / replicate 2 pgR-poor DOX+ Linker long
AE6382	CAAGCAGAAGACGGCATACGAGATCGG TCTCGGCATTCCTGCTGAACCGCTCTT CCGATCTGTAATACGACTCACTATAGG GC	replicate 2 5dpi Linker primer / replicate 2 pgR-poor DOX+ Linker prime
AE6404-1	AATGATACGGCGACCACCGAGATCTAC ACTCTTTCCCTACACGACGCTCTTCCG ATCTTTAGGCGAGATCCCTCAGACCCT TTTAGTCAG	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	GAGAATCCATGAGTATGCTCACGCGTC ACCGTCATAG	replicate 1 pgR-rich DOX- Linker long / CD4+ donor 2 Linker long
AE6388	CAAGCAGAAGACGGCATACGAGATCGG TCTCGGCATTCCTGCTGAACCGCTCTT CCGATCTGAGAATCCATGAGTATGCTC AC	replicate 1 pgR-rich DOX- Linker primer / CD4+ donor 2 Linker primer
AE6406	AATGATACGGCGACCACCGAGATCTAC ACTCTTTCCCTACACGACGCTCTTCCG ATCTACAGTGGAGATCCCTCAGACCCT TTTAGTCAG	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	GACGTACATACTGATCGCATAGCAGAC TGCGTCAGTC	replicate 1 pgR-poor DOX+ Linker long
AE6458	CAAGCAGAAGACGGCATACGAGATCGG TCTCGGCATTCCTGCTGAACCGCTCTT CCGATCTGACGTACATACTGATCGCAT AG	replicate 1 pgR-poor DOX+ Linker primer
AE6405	AATGATACGGCGACCACCGAGATCTAC ACTCTTTCCCTACACGACGCTCTTCCG ATCTTGACCAGAGATCCCTCAGACCCT TTTAGTCAG	replicate 1 pgR-poor DOX+ Second round LTR primer
AE6386	TACTATGACGGTGACGC-NH2	replicate 2 pgR-rich DOX+ Linker short
AE6387	GAGAATCCATGAGTATGCTCACGCGTC ACCGTCATAG	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

Table 3. Oligonucleotides used for electrophoretic mobility shift assay substrate preparation.

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
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	R-loop, R:D+ssDNA
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 G	Hybrid
Hybrid RNA	5' -[Cy5]-ACC CGG AGC UUG GAC AGC CAA	Hybrid
oligo 5	5′ GCAGTAGCATGACGCTGCTGAATTCTACCAC GCTATGCT CTCGTCTAGGTTCACTCCGT CCCTGCGATTCATACCTGTCGTGCCAGCTGC	R:D+ssDNA

Table 4. Primers used for qPCR.

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 (mAIRN Fwd)	CGAGAGAGGCTAAGGGTGAA
SH21 (ECFP/mAIRN Rev)	ACATGGTCCTGCTGGAGTTC
C1 Fwd	TCCTCTGTCCTCTTCCCAGTT
C1 Rev	GAGAGGCTTTGATGGCGATAC
C2 Fwd	GCAGAGAGTCCTCACTCCAAG
C2 Rev	TGCAATTGTCAGGGTCCACT
C3 Fwd	ACACAAGTTTTGCCTATGTGGC
C3 Rev	TGTTTTGGTCCTTGGTGAATGT
C4 Fwd	TGCTCAGGTGTCAATCCTCG
C4 Rev	CCTTTCCCTGCTGACCAACT
C5 Fwd	TCTTGGAAAGTTCTGGACGCT
C5 Rev	AGGTCTGGGGCCTGAACA
I1 Fwd	CTATGGTGGTGGTGGGAACTAT
I1 Rev	CTGTAGCCCTTTTCTGCTGTTT
I2 Fwd	AAATAAATTGCCTCCCTTCCTC
I2 Rev	CCTTTTATGTCCTTCCCAGTCA
I3 Fwd	TGCTTTCTTTCTTCACCTCCTC
I3 Rev	CGGAAGTTTGTAATGGAAAAGG
I4 Fwd	GATGAGTTGGGGGAGTTAATGA
I4 Rev	AAGAGAGAGCGAAAGAGCAAGA
I5 Fwd	AGGGGAGATTTTACCTCTCCTG
I5 Rev	TCAGTTGGGGTATGTGAGTGAG
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

3. RESULT

3.1. DRIPc-seq analysis of host genomic R-loops dynamics upon HIV-1 infection

To investigate the relationship between HIV-1 infection and host cellular R-loops, I first analyzed R-loop dynamics in different types of cells infected with HIV-1 at early post-infection time points using DNA-RNA immunoprecipitation followed by cDNA conversion coupled to high-throughput sequencing (DRIPc-seq) using a DNA-RNA hybrid-specific binding antibody, anti-S9.6 (Sanz and Chedin, 2019). This recently invented high-throughput R-loop mapping method includes DNase I digestion of S9.6-immunoprecipitated material, which not only removes the nonspecific DNA and DNA portion of the R-loops but only recovers the RNA moiety. Therefore, DRIPc-seq is strand-specific and has a higher resolution than the conventional DRIP-seq (Sanz and Chedin, 2019). HeLa cells, primary $CD4^+$ T cells isolated from two individual donors and $CD4^+/CD8^-$ T cell lymphoma Jurkat cell line were infected with VSV-G-pseudotyped HIV-1-EGFP and harvested at 0, 3, 6, and 12 h post infection (hpi) for DRIPc-seq library construction (Figure 4 and Figure 5A-C). Our DRIPc-seq analysis yielded loci specific R-loop signals at the referenced R-loop-positive loci (RPL13A and CALM3) and an R-loop-negative locus (SNRPN) (Sanz and Chedin, 2019) that were both strand-specific and highly sensitive to pre-immunoprecipitation in vitro RNase H treatment (Table 5-7).

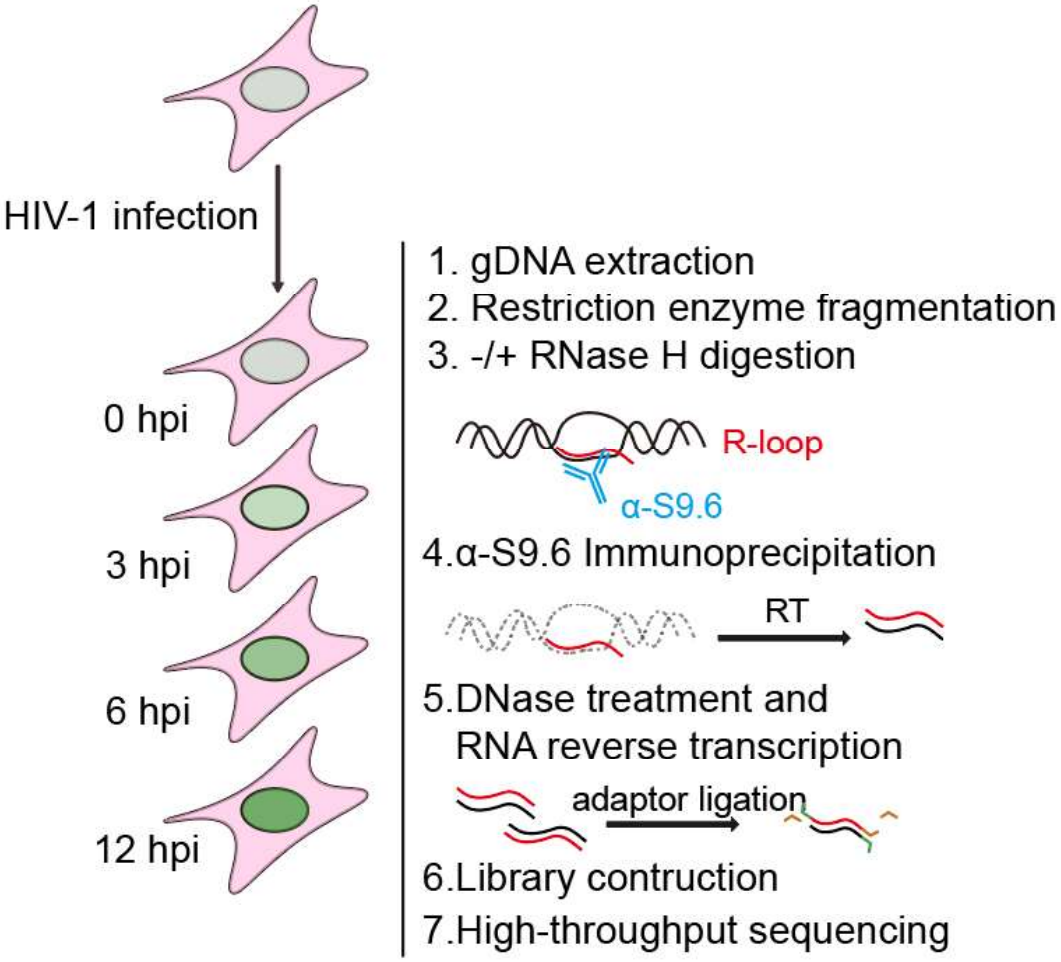
HIV-1 infection
0 hpi
3 hpi
6 hpi
12 hpi
1. gDNA extraction
2. Restriction enzyme fragmentation
3. -/+ RNase H digestion
R-loop
α-S9.6
4.α-S9.6 Immunoprecipitation
RT
5.DNase treatment and
RNA reverse transcription
adaptor ligation
6.Library contruction
7.High-throughput sequencing

Figure 4. Summary of experimental design for DRIPc-seq in HeLa cells, primary $CD4^+$ T cells and Jurkat cells infected with HIV-1.

For DRIPc-seq library construction and analysis of HIV-1 infected cells, cells were infected with VSV-G-pseudotyped HIV-1-EGFP and harvested at 0, 3, 6, and 12 h post infection (hpi). Genomic DNA (gDNA) were extracted and fragmented by restriction enzymes. For DNA-RNA hybrid negative control, RNase H enzymes were treated to the half of fragmented gDNA pre-immunoprecipitates then subjected to immunoprecipitation by S9.6 antibodies. Immunoprecipitated materials were treated with DNase I enzyme and reverse transcribed to only recover the RNA moiety of R-loop.

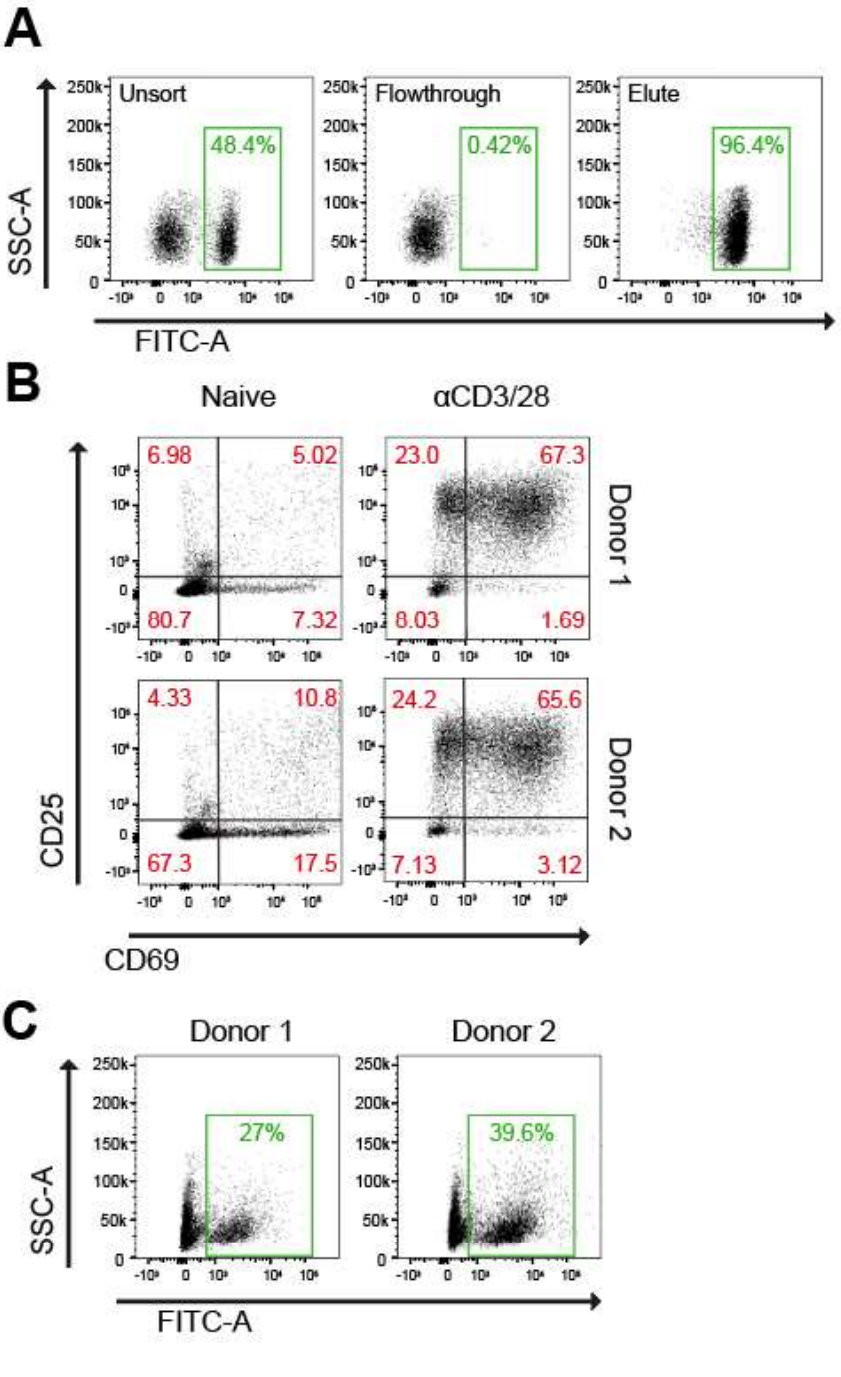
A
Unsort
48.4%
Flowthrough
0.42%
Elute
96.4%
SSC-A
FITC-A
B
Naive
αCD3/28
6.98
5.02
80.7
7.32
23.0
67.3
8.03
1.69
4.33
10.8
67.3
17.5
24.2
65.6
7.13
3.12
Donor 1
Donor 2
CD25
CD69
C
Donor 1
Donor 2
27%
39.6%
SSC-A
FITC-A

Figure 5. Primary $CD4^+$ T cell isolation and HIV-1 infection.

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

Table 5. Chromosomal position and DRIPc-seq signal for R-loop-positive and -negative reference regions in HeLa cells.

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
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
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

Table 6. Chromosomal position and DRIPc-seq signal for R-loop-positive and -negative reference regions in primary $CD4^+$ T 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
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
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

Table 7. Chromosomal position and DRIPc-seq signal for R-loop-positive and -negative reference regions in primary Jurkat 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
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
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

Notably, the number of DRIPc-seq peaks mapped to the human reference genome increased remarkably during early post infection of HIV-1, at 3 hpi and 6 hpi for HeLa cells (Figure 6A). Most of the peaks mapped in cells harvested at 0 hpi were commonly found in all other samples, but a significant numbers of unique peaks were observed after infection (Fig. 6B). Importantly, nearly 100% of DRIPc-seq reads were aligned to the host cellular genome, but not on that of HIV-1, which forms DNA-RNA hybrid during its viral life cycle (Figure 6C).

$CD4^+$ T cells are the physiological targets of HIV-1 infection. In T cells, the number of DRIPc-seq peaks mapped to the human reference genome increased significantly at 6 and 12 hpi (Figure 7A and 7B). Most of the peaks mapped in cells harvested at 0 and 3 hpi were commonly found in all other samples, but significant numbers of unique peaks were observed at 6 and 12 hpi (Figure 7C and 7D). Importantly, most of DRIPc-seq reads were aligned to the host cellular genome at all post infection time points (Figure 7E and 7F). Together, I suggest that host genomic R-loops accumulate considerably in multiple types of cells upon HIV-1 infection, particularly during the early post-infection time points.

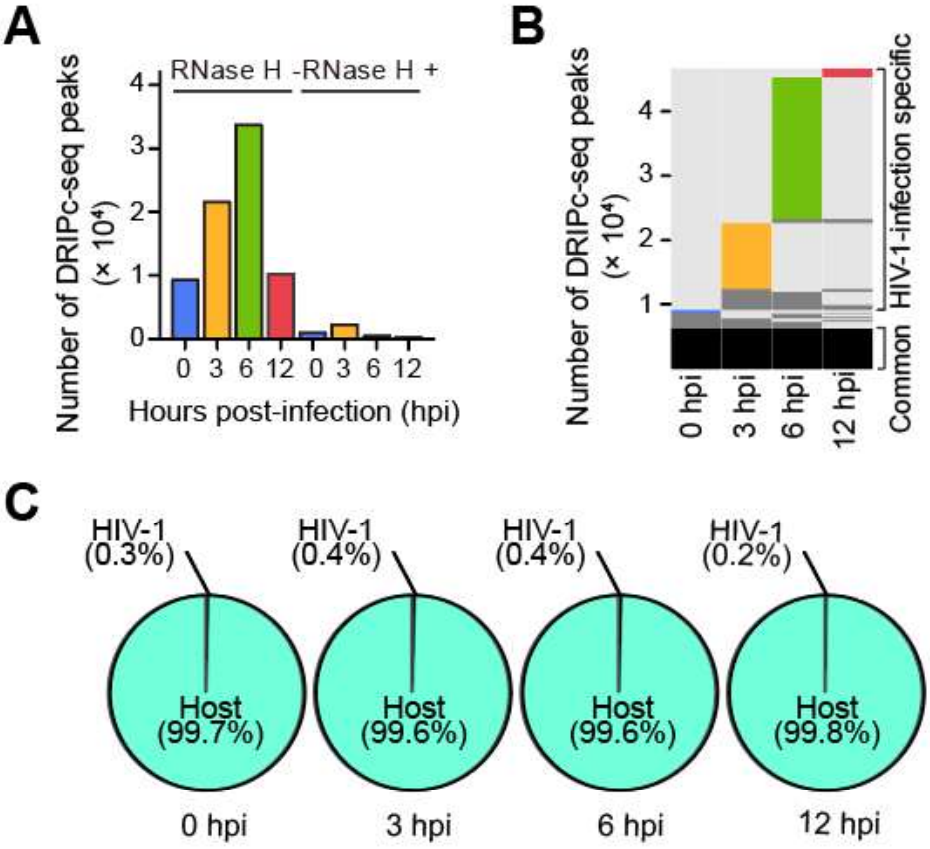
A
RNase H -
RNase H +
Number of DRIPc-seq peaks (× 10⁴)
0 3 6 12 0 3 6 12
Hours post-infection (hpi)
B
Number of DRIPc-seq peaks (× 10⁴)
HIV-1-infection specific
Common
0 hpi
3 hpi
6 hpi
12 hpi
C
HIV-1 (0.3%)
HIV-1 (0.4%)
HIV-1 (0.4%)
HIV-1 (0.2%)
Host (99.7%)
Host (99.6%)
Host (99.6%)
Host (99.8%)
0 hpi
3 hpi
6 hpi
12 hpi

Figure 6. DRIPc-seq analysis in HIV-1-infected HeLa cells at early post infection.

(A) Bar graph indicating DRIPc-seq peak counts for HIV-1-infected HeLa cells with MOI of 0.6 harvested at the indicated hours post infection (hpi). Pre-immunoprecipitated samples were untreated (-) or treated (+) with RNase H, as indicated. Each bar corresponds to pooled datasets from two biologically independent experiments. (B) All genomic loci overlapping a DRIPc-seq peak from HIV-1 infected HeLa cells in at least one sample are stacked vertically; the position of each peak in a stack is constant horizontally across samples. Each hpi occupies a vertical bar, as indicated. Each bar corresponds to pooled datasets from two biologically independent experiments. Common peaks for all samples are represented in black, and in dark gray for those common for at least two samples. The lack of a DRIP signal over a given peak in any sample is shown in light gray. The sample-unique peaks are colored blue, yellow, green, and red at 0, 3, 6, and 12 hpi, respectively. (C) Pie graphs indicating the percentage of DRIPc-seq reads aligned to host cellular genome (aquamarine) or to HIV-1 viral genome (gray), out of the total consensus DRIPc-seq peaks from HIV-infected HeLa cells.

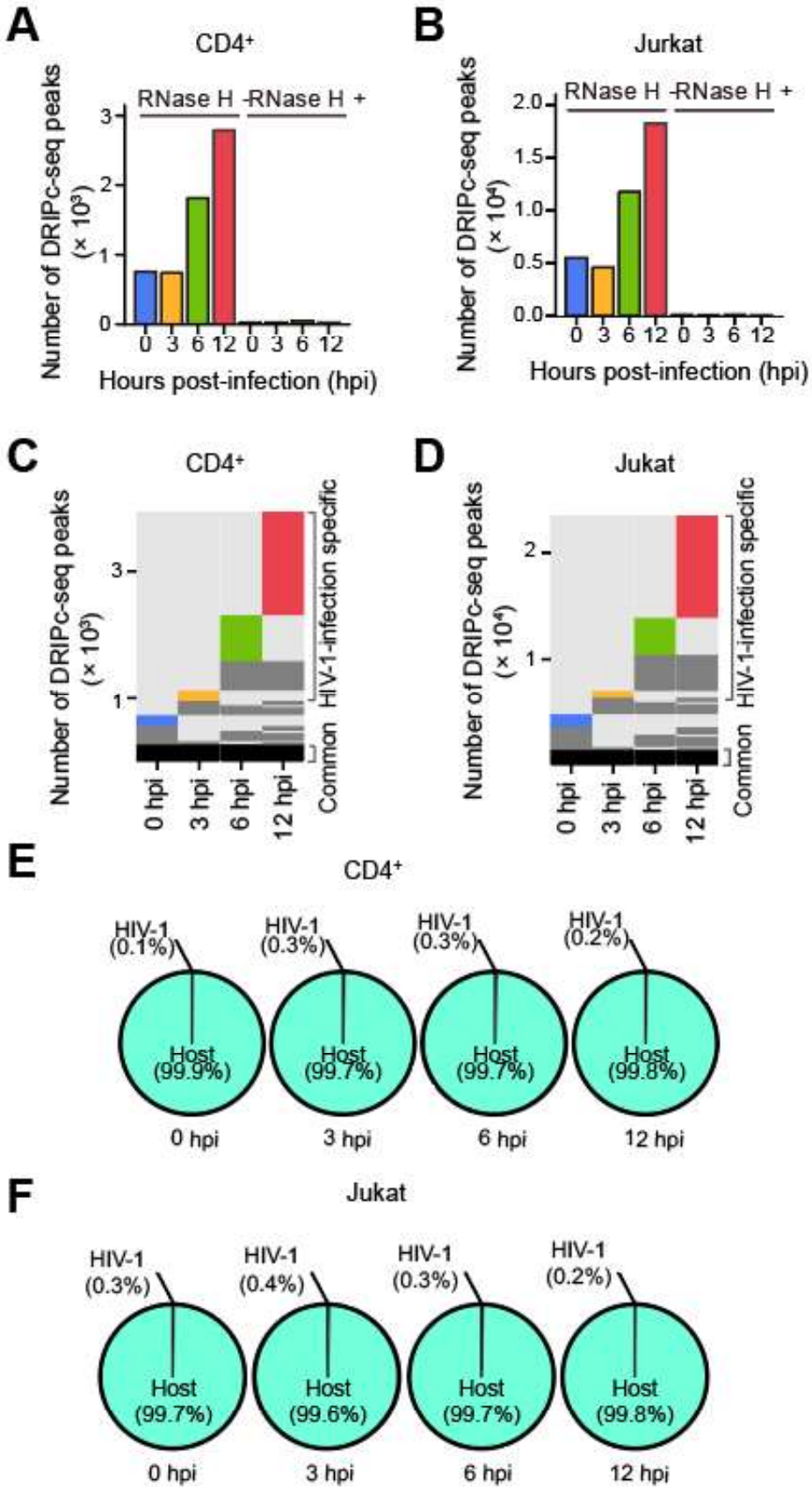
A
CD4+
RNase H -
RNase H +
Number of DRIPc-seq peaks (× 10³)
0 3 6 12 0 3 6 12
Hours post-infection (hpi)
B
Jurkat
RNase H -
RNase H +
Number of DRIPc-seq peaks (× 10⁴)
0 3 6 12 0 3 6 12
Hours post-infection (hpi)
C
CD4+
Number of DRIPc-seq peaks (× 10³)
HIV-1-infection specific
Common
0 hpi
3 hpi
6 hpi
12 hpi
D
Jukat
Number of DRIPc-seq peaks (× 10⁴)
HIV-1-infection specific
Common
0 hpi
3 hpi
6 hpi
12 hpi
E
CD4+
HIV-1 (0.1%)
HIV-1 (0.3%)
HIV-1 (0.3%)
HIV-1 (0.2%)
Host (99.9%)
Host (99.7%)
Host (99.7%)
Host (99.8%)
0 hpi
3 hpi
6 hpi
12 hpi
F
Jukat
HIV-1 (0.3%)
HIV-1 (0.4%)
HIV-1 (0.3%)
HIV-1 (0.2%)
Host (99.7%)
Host (99.6%)
Host (99.7%)
Host (99.8%)
0 hpi
3 hpi
6 hpi
12 hpi

Figure 7. DRIPc-seq analysis in HIV-1-infected primary $CD4^+$ T cells and Jurkat cells at early post infection.

(A and B) Bar graph indicating DRIPc-seq peak counts for primary $CD4^+$ T cells infected with 600ng/p24 of HIV-1 (A) and Jurkat T cells infected with 300ng/p24 of HIV-1 (B) per 1× 106 cells/mL, harvested at the indicated hours post infection (hpi). Pre-immunoprecipitated samples were untreated (-) or treated (+) with RNase H, as indicated. Each bar corresponds to pooled datasets from two biologically independent experiments. (C and D) All genomic loci overlapping a DRIPc-seq peak from HIV-1 infected primary $CD4^+$ T cells (C) and Jurkat T cells (D) in at least one sample are stacked vertically; the position of each peak in a stack is constant horizontally across samples. Each hpi occupies a vertical bar, as indicated. Each bar corresponds to pooled datasets from two biologically independent experiments. Common peaks for all samples are represented in black, and in dark gray for those common for at least two samples. The lack of a DRIP signal over a given peak in any sample is shown in light gray. The sample-unique peaks are colored blue, yellow, green, and red at 0, 3, 6, and 12 hpi, respectively. (E and F) Pie graphs indicating the percentage of DRIPc-seq reads aligned to host cellular genome (aquamarine) or to HIV-1 viral genome (gray), out of the total consensus DRIPc-seq peaks from HIV-infected $CD4^+$ (E) and Jukat (F) T cells.

3.2. Host cellular R-loops accumulate after HIV-1 infection in HeLa cells

To strengthen my DRIPc-seq data analysis, I used a number of different biochemical approaches to examine R-loops in HeLa cells. First, R-loop accumulation in HIV-1-infected cells was observed using DNA-RNA hybrid dot blots with the anti-S9.6 antibodies (Figure 8A). The dot intensity increased significantly upon HIV-1 infection at 6 hpi, and the enhanced R-loop signals on dot blots of HIV-1-infected cells were highly sensitive to in vitro treatment with RNase H. This result was highly consistent with my DRIPc-seq data analysis results. Subsequently, I observed HIV-1-induced R-loops using an immunofluorescence assay by probing HIV-1-infected or non-infected control cells with S9.6 antibody at 6 hpi (Figure 8B, left). I quantified R-loop accumulation in HIV-1-infected cells by subtracting the nucleolar signal (green) from the S9.6 signal (red) intensity per nucleus. The nuclear fluorescence signal associated with the R-loops was significantly enhanced in cells infected with HIV-1 (Figure 8B, right). Importantly, R-loop signal was enriched even in cells when the reverse transcription or integration of HIV-1 is blocked by enzyme inhibitors like Nevirapine (NVP) or Raltegravir (RAL), respectively (Figure 9A and 9B.). This result emphasizes that the enrichment of R-loop signals in cells are originated from the host genome but not by DNA-RNA hybrid formation during the viral life cycle or transcriptional burst from infected HIV-1 proviruses.

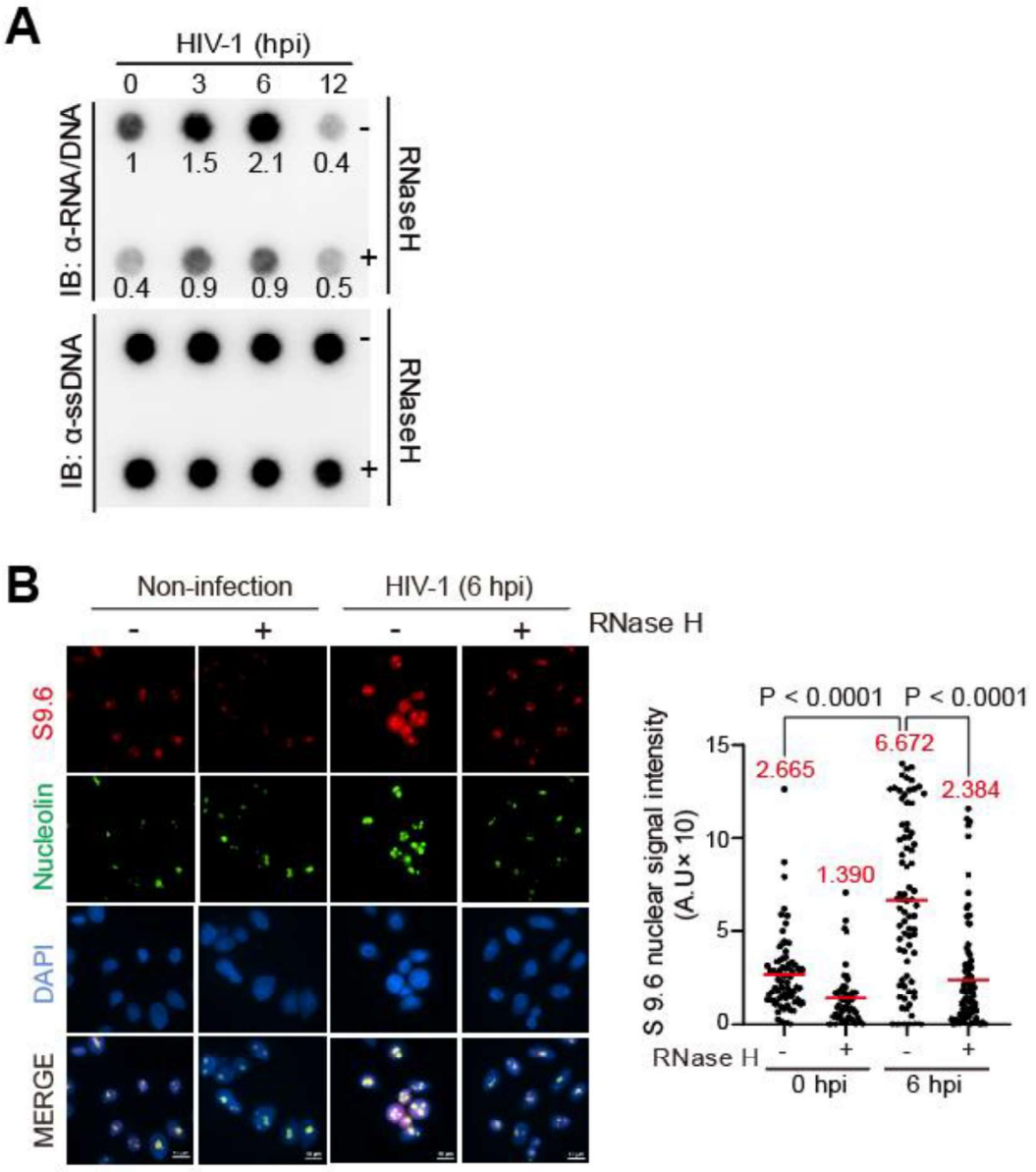
A
HIV-1 (hpi)
0
3
6
12
IB: α-RNA/DNA
1
1.5
2.1
0.4
0.4
0.9
0.9
0.5
RNaseH
IB: α-ssDNA
RNaseH
B
Non-infection
HIV-1 (6 hpi)
RNase H
S9.6
Nucleolin
DAPI
MERGE
P < 0.0001
P < 0.0001
S 9.6 nuclear signal intensity (A.U× 10)
2.665
1.390
6.672
2.384
RNase H
0 hpi
6 hpi

Figure 8. HIV-1 infection induces cellular R-loop accumulation in cells at early post-infection.

(A) Dot blot analysis of the R-loop in gDNA extracts from HIV-1 infected HeLa cells with MOI of 0.6 harvested at the indicated hpi. gDNAs were probed with anti-S9.6. gDNA extracts were incubated with or without RNase H in vitro before membrane loading (anti-RNA/DNA signal). Fold-induction was normalized to the value of harvested cells at 0 hpi by quantifying the dot intensity of the blots and calculating the ratios of the S9.6 signal to the total amount of gDNA (anti-ssDNA signal). (B) Representative images of the immunofluorescence assay of S9.6 nuclear signals in HIV-1 infected HeLa cells with MOI of 0.6 harvested at 6 hpi. The cells were pre-extracted of cytoplasm and co-stained with anti-S9.6 (red), anti-nucleolin antibodies (green), and DAPI (blue). The cells were incubated with or without RNase H in vitro before staining with anti-S9.6 antibodies, as indicated. Quantification of S9.6 signal intensity per nucleus after nucleolar signal subtraction for the immunofluorescence assay. The mean value for each data point is indicated by the red line. Statistical significance was assessed using one-way ANOVA ($n > 53$).

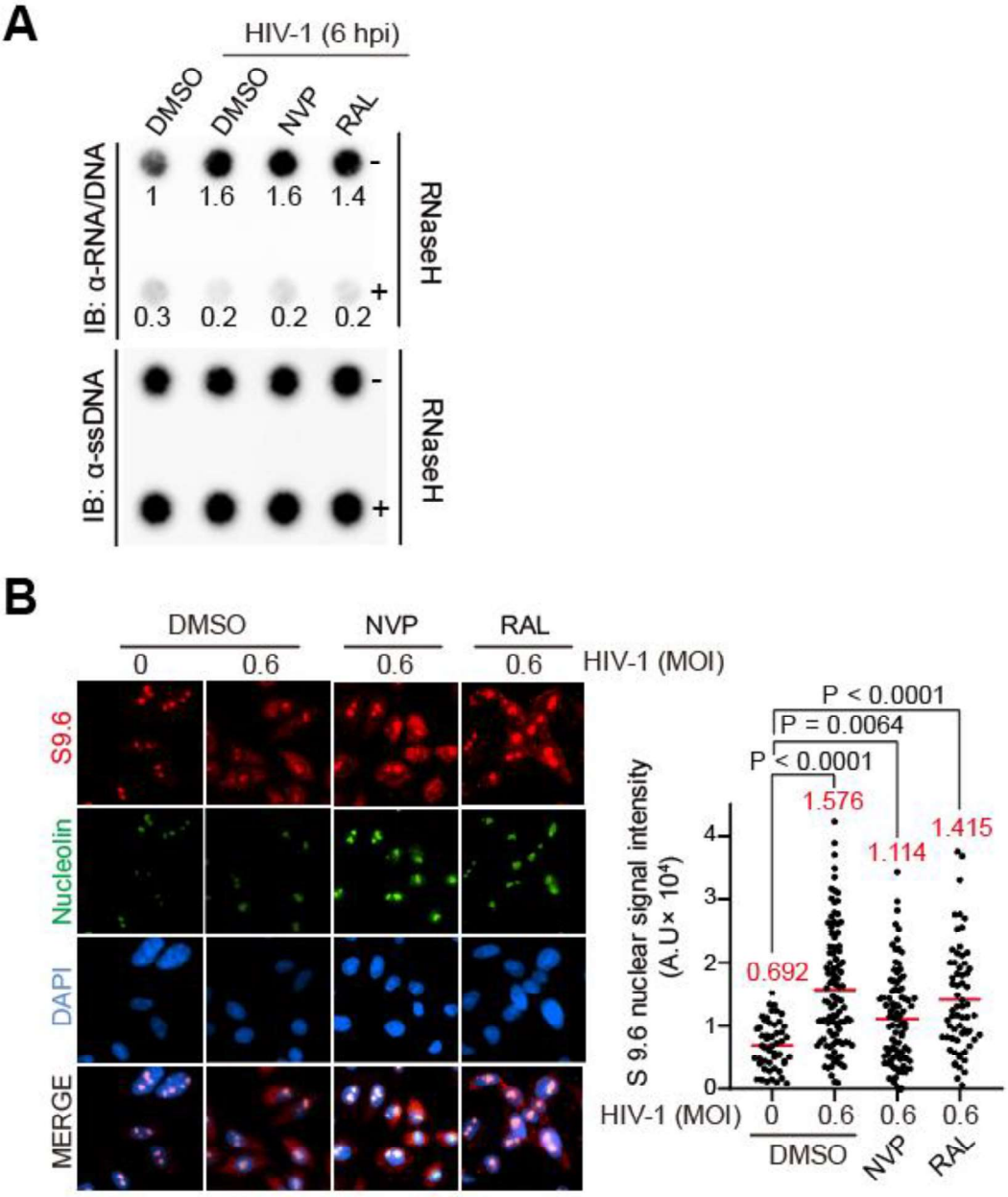
A
HIV-1 (6 hpi)
DMSO
DMSO
NVP
RAL
IB: α-RNA/DNA
RNaseH
1
1.6
1.6
1.4
0.3
0.2
0.2
0.2
IB: α-ssDNA
RNaseH
B
DMSO
NVP
RAL
0
0.6
0.6
0.6
HIV-1 (MOI)
S9.6
Nucleolin
DAPI
MERGE
P < 0.0001
P = 0.0064
P < 0.0001
1.576
1.415
1.114
0.692
S 9.6 nuclear signal intensity (A.U× 10⁴)
HIV-1 (MOI)
0
0.6
0.6
0.6
DMSO
NVP
RAL

Figure 9. HIV-1 infection still induces cellular R-loop accumulation when its reverse transcription or integration was inhibited.

(A) Dot blot analysis of the R-loop in gDNA extracts from HIV-1 infected HeLa cells with MOI of 0.6 harvested at 6hpi. The cells were treated with DMSO, 10uM of Nevirapine (NVP), or 10uM of Raltegravir (RAL) for 24 h before infection, as indicated. gDNAs were probed with anti-S9.6. gDNA extracts were incubated with or without RNase H in vitro before membrane loading (anti-RNA/DNA signal). Fold-induction was normalized to the value of harvested cells at 0 hpi by quantifying the dot intensity of the blots and calculating the ratios of the S9.6 signal to the total amount of gDNA (anti-ssDNA signal). (B) Representative images of the immunofluorescence assay of S9.6 nuclear signals in HIV-1 infected HeLa cells with MOI of 0.6 at 6 hpi. The cells were pre-extracted of cytoplasm and co-stained with anti-S9.6 (red), anti-nucleolin antibodies (green), and DAPI (blue). The cells were treated with DMSO, 10uM of Nevirapine (NVP), or 10uM of Raltegravir (RAL) for 24 h before infection, as indicated. Quantification of S9.6 signal intensity per nucleus after nucleolar signal subtraction for the immunofluorescence assay. The mean value for each data point is indicated by the red line. Statistical significance was assessed using one-way ANOVA ($n > 51$).

3.3. R-loops induced by HIV-1 are widely distributed in both genic and non-genic regions regardless of the expression

To investigate the distribution of cellular genomic R-loops during HIV-1 infection, I conducted a genome-wide analysis, in HeLa cells. I observed a significant accumulation of R-loops in diverse genomic compartments at 3 and 6 hpi, while the R-loops from the 0 and 12 hpi samples did not exhibit any distinct pattern of induction in the indicated genomic compartments (Figure 10A). The presence of R-loops is often correlated with high transcriptional activity. Consistent with this observation, I found that the gene body regions had the highest numbers of DRIPc-seq peaks, and their enrichment was evident upon HIV-1 infection at both 3 and 6 hpi (Figure 10B). However, I also observed a significant number of DRIPc-seq peaks mapped to intergenic or repeat regions at both 3 and 6 hpi, including short interspersed nuclear elements (SINEs), long interspersed nuclear elements (LINEs), and long terminal repeat (LTR) retrotransposons, where transcription is typically repressed (Figure 10B).

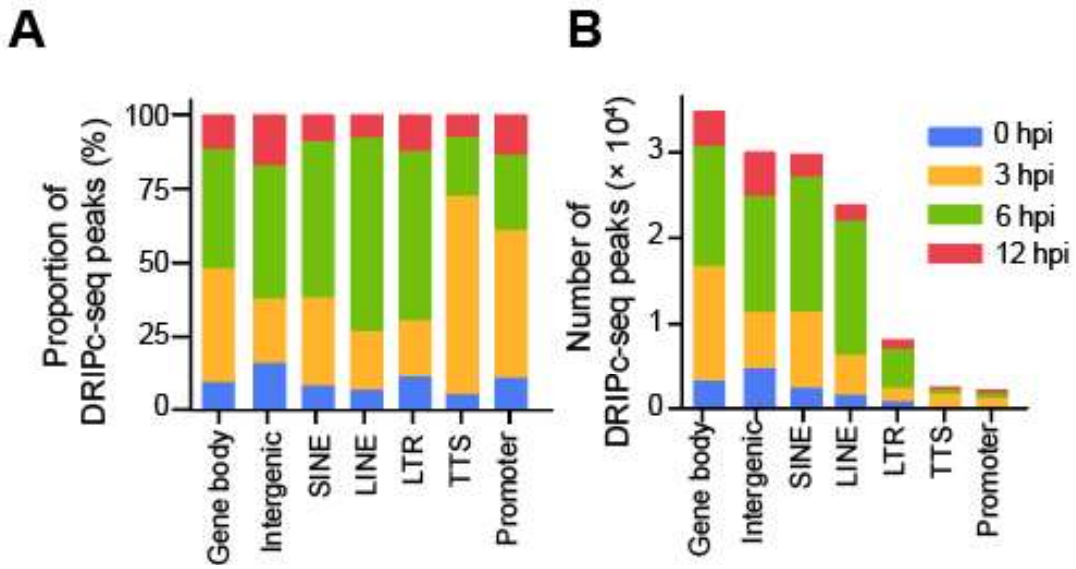
A
B
Proportion of DRIPc-seq peaks (%)
100
75
50
25
0
Gene body
Intergenic
SINE
LINE
LTR
TTS
Promoter
Number of DRIPc-seq peaks (× 10⁴)
3
2
1
0
0 hpi
3 hpi
6 hpi
12 hpi

Figure 10. HIV-1-induced R-loops are enriched at both transcriptionally active and silent regions.

(A) Stacked bar graphs indicating the proportion of DRIPc-seq peaks mapped for HIV-1-infected HeLa cells harvested at the indicated hpi over different genomic features. (B) Stacked bar graphs indicating the number of DRIPc-seq peak counts for HIV-1-infected HeLa cells harvested at the indicated hpi over different genomic features

Although the expression of repetitive elements, including SINE, LINE, and LTR, is mostly repressed during normal cellular activities, HIV-1 infection could activate endogenous retroviral promoters (Jones et al., 2013; Srinivasachar Badarinarayan et al., 2020). To investigate the possibility that R-loop induction in gene-silent regions is associated with transcriptome changes during HIV-1 infection, I performed RNA sequencing (RNA-seq) for HIV-1-infected HeLa cells at 0, 3, 6, and 12 hpi, similar to the DRIPc-seq analysis. Consistent with previous reports, I observed an increase in the expression levels of repetitive elements at later time points post-infection (Figure 10; 12 hpi). In contrast, I 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 (Figure 11).

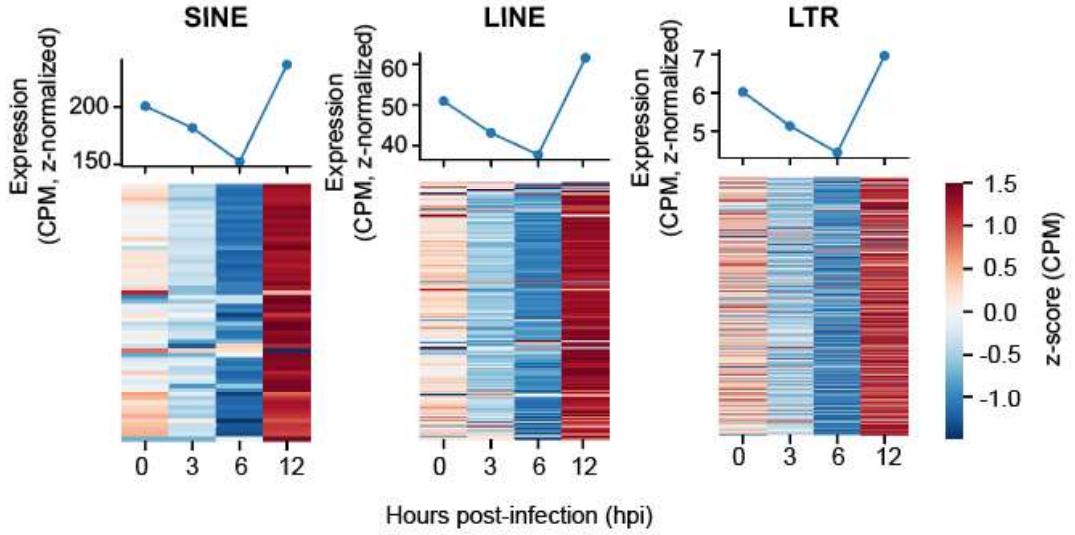
SINE
LINE
LTR
Expression (CPM, z-normalized)
200
150
60
50
40
7
6
5
0 3 6 12
Hours post-infection (hpi)
1.5
1.0
0.5
0.0
-0.5
-1.0
z-score (CPM)

Figure 11. R-loop induction by HIV-1 infection in repetitive elements does not follow transcriptome changes.

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.

I further examined the expression profile of genes containing HIV-1-induced R-loops at 3 and 6 hpi because they yielded the highest number of HIV-1-induced R-loops. 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 = 1.08×10^{-84}; Figure 12A) and at 6 hpi samples (Pearson's r = -0.34, P = 2.40×10^{-228}; Figure 12A). Because unique DRIPc-seq peaks at 3 and 6 hpi represent a large proportion of the total consensus DRIPc-seq peaks induced by HIV-1 infection and the respective samples consistently display distinct R-loop enriched features compared to the 0 and 12 hpi samples, I defined unique DRIPc-seq peaks at 3 and 6 hpi as"HIV-1-induced R-loops" and all other consensus DRIPc-seq peaks found at all hpi as "constitutive R-loops".

I compared GC skew values of the constitutive and HIV-1-induced R-loops. A High GC skew value is a well-established predisposing factor for R-loop formation upon transcription (Ginno et al., 2013; Lim et al., 2015). HIV-1-induced R-loops showed significantly lower absolute GC skew values (Figure 12B). Moreover, I compared nucleotide features of the constitutive and HIV-1-induced R-loops. I observed abrupt flips in the polarity of AT and GC skew shifting from low to high skew absolute values at the centers of the R-loop peaks, only for constitutive R-loops (Figure 12C and 12D). This implies that the constitutive R-loops are DNA-RNA hybrids of two

strands in head-on (convergent) orientation (Crossley et al., 2023), but HIV-1-induec R-loops are not.

Furthermore, I observed R-loop enrichment in diverse genomic compartments including gene body, intergenic and repeat regions, at 6 and 12 hpi, in primary $CD4^+$ T cells and Jurkat cells (Figure 13A and 13B). These findings demonstrate that R-loop accumulation occurs throughout the genome, including both genic and non-genic regions, during HIV-1 infection and in different cell types including T cells. This accumulation is not limited to regions where transcriptomic changes are induced by HIV-1 infection and implies a more complex interplay between viral infection and R-loop formation. Together, these result suggest that HIV-1-induced R-loops are non-canonical, and formed in diverse genomic regions but independently of the transcription activation, perhaps *in-trans* manner.

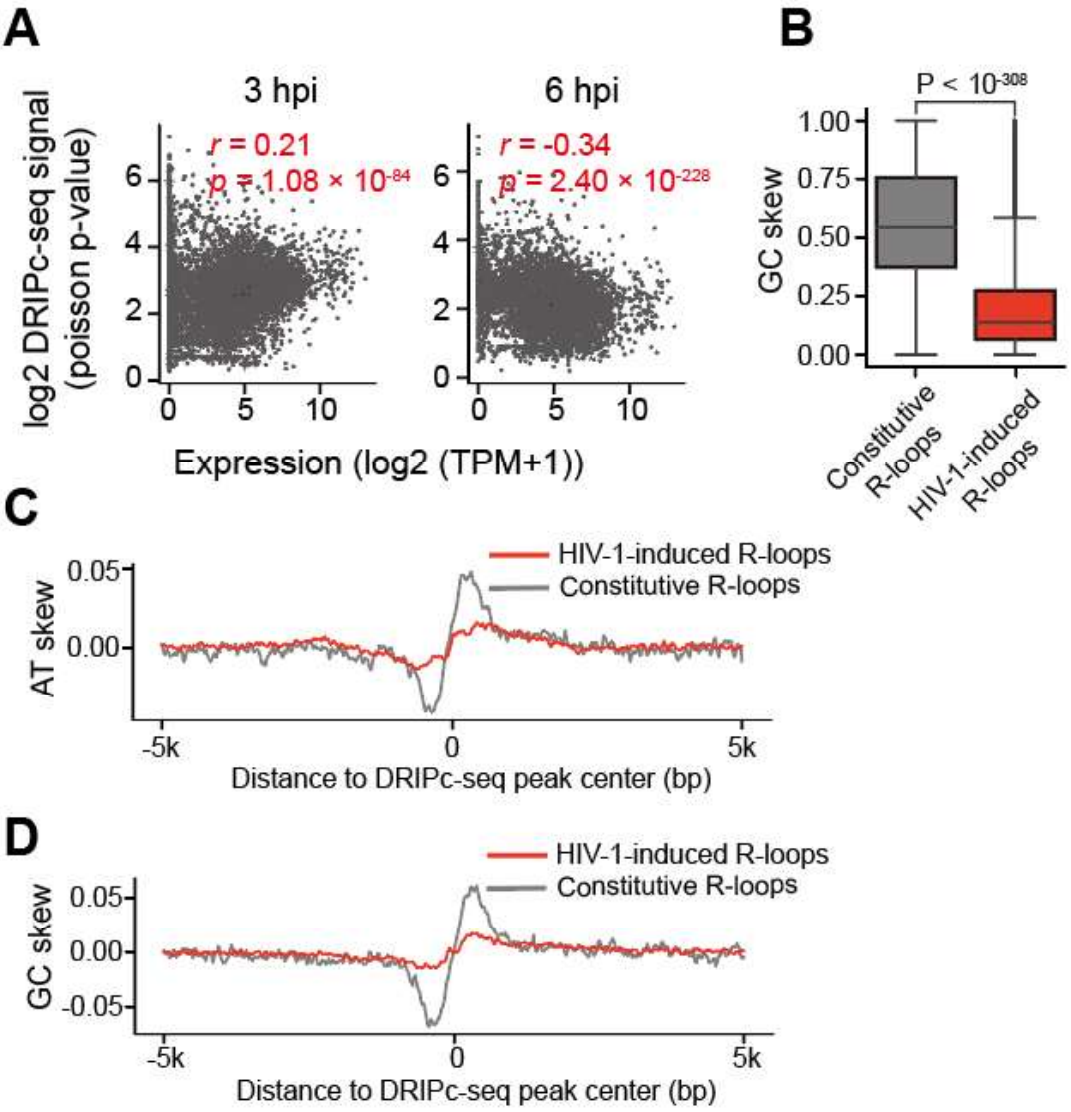
A
3 hpi
6 hpi
r = 0.21
p = 1.08 × 10-84
r = -0.34
p = 2.40 × 10-228
log2 DRIPc-seq signal (poisson p-value)
Expression (log2 (TPM+1))
B
P < 10-308
GC skew
Constitutive R-loops
HIV-1-induced R-loops
C
HIV-1-induced R-loops
Constitutive R-loops
AT skew
Distance to DRIPc-seq peak center (bp)
D
HIV-1-induced R-loops
Constitutive R-loops
GC skew
Distance to DRIPc-seq peak center (bp)

Figure 12. Host genomic R-loop accumulation is not limited to regions where transcriptomic changes are induced by HIV-1 infection.

(A) Correlation between gene expression and DRIPc-seq signals of HIV-1-infected HeLa cells with MOI of 0.6 harvested at the indicated hpi. Statistical significance was assessed using Pearson's r and p-values. (B) Box plot indicating the GC skewed absolute values of the mapped DRIPc-seq peaks for HIV-1-induced or constitutive R-loops. Statistical significance was assessed using two-sided independent t-tests. (C and D) AT skew (C) and GC skew (D) of HIV-1-induced (red solid lines) or constitutive R-loops (gray solid lines) in 10-kb windows.

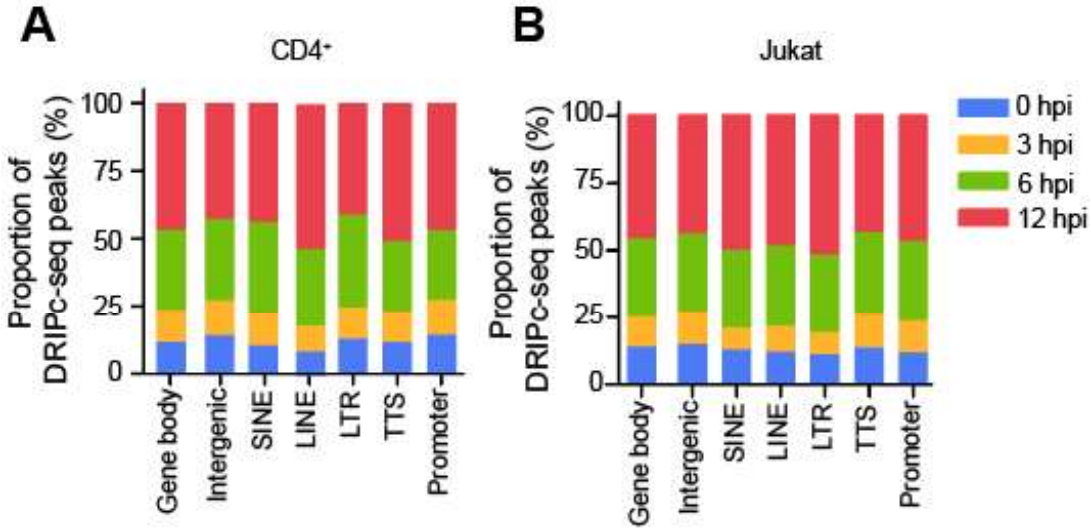
A
CD4+
Proportion of DRIPc-seq peaks (%)
100
75
50
25
0
Gene body
Intergenic
SINE
LINE
LTR
TTS
Promoter
B
Jukat
0 hpi
3 hpi
6 hpi
12 hpi

Figure 13. Genome-wide R-loop induction by HIV-1 infection in T cells.

(A and B) Stacked bar graphs indicating the proportion of DRIPc-seq peaks mapped for HIV-1-infected primary $CD4^+$ (A) and Jurkat (B) T cells harvested at the indicated hpi over different genomic features.

3.4. Host genomic R-loops regulate HIV-1 integration

To investigate the role of R-loops in HIV-1 life cycle, I examined HIV-1 infectivity in HeLa and Jurkat cells ectopically expressing Flag-tagged RNase H1 enzyme (RNH1), which specifically degraded the RNAs of DNA-RNA hybrids (Figure 14A and 14B). When the cells were infected with VSV-G-pseudotyped HIV-1-luciferase viruses, cells expressing wild-type RNH1 showed significantly lower luciferase activity than that of enzymatic inactive mutant RNH1 ($RNH1^{D10R/E48R}$) expressing cells (Figure 15).

R-loops are the important modulators and composers of the cellular genomic dynamics. HIV-1 completes its infection by integrating its viral genome into the host' s and closely interact with the host genome particularly during integration. Besides, as HIV-1 infection induced R-loop accumulation at early post infection hours when HIV-1 integration may initiate(Albanese et al., 2008; Brussel and Sonigo, 2003), I hypothesized that host genomic R-loops play a role in HIV-1 integration, and possibly in integration site selection. I carried DRIPc-sequencing and HIV-1 integration site sequencing in HeLa cells ectopically expressing wild-type and mutant RNH1. When cellular R-loops were removed by wild-type RNH1 expression (Figure 16A), the HIV-1 integration events at R-loop regions were decreased by approximately two-folds (Figure 16B). Interestingly, the integrated HIV-1 viral genomes were found farther away from

the HIV-1-induced R-loops in cells expressing wild-type RNH1 ectopically (mean distance = 4.6 $\log_{10}$ bp; Figure 16C) than in cells expressing mutant RNH1 (median distance = 4.8 $\log_{10}$ bp; Figure 16C, HIV-1-induced). Notably, the distance between HIV-1 integration sites and constitutive R-loops was comparable, regardless of R-loop resolution (Figure 16C, Constitutive). This finding substantiates the preferential integration of HIV-1 into HIV-1-induced R-loops within the host genome.

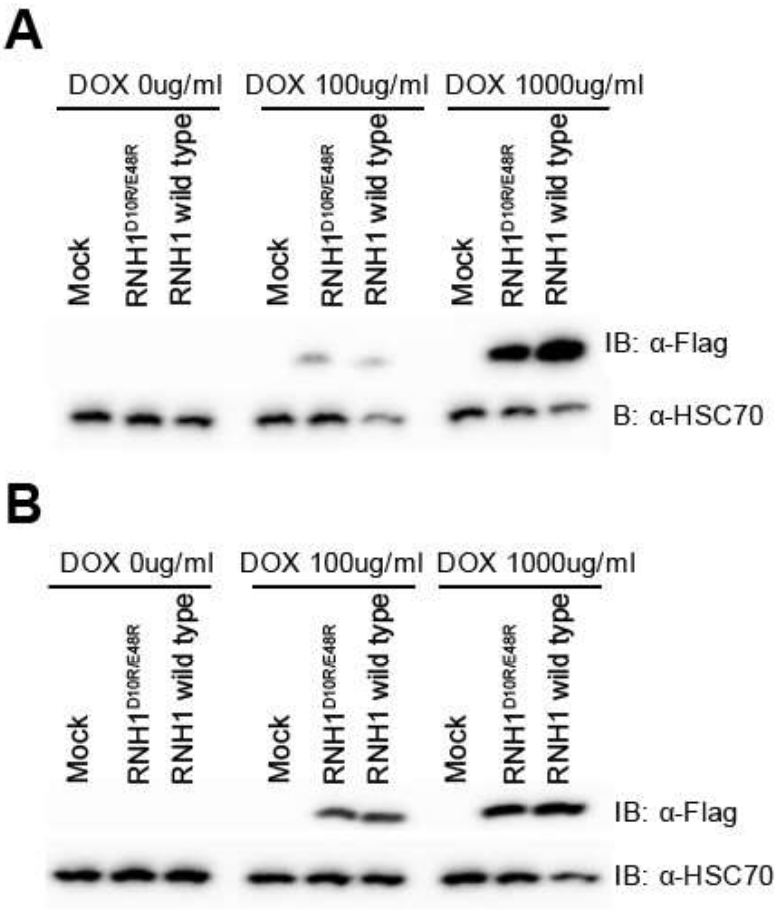
A
DOX 0ug/ml
DOX 100ug/ml
DOX 1000ug/ml
Mock
RNH1D10R/E48R
RNH1 wild type
IB: α-Flag
B: α-HSC70
B
DOX 0ug/ml
DOX 100ug/ml
DOX 1000ug/ml
Mock
RNH1D10R/E48R
RNH1 wild type
IB: α-Flag
IB: α-HSC70

Figure 14. Regulation of cellular R-loops by RNase H1 expression. (A and B) Immunoblots of FLAG-tag (α-FLAG) and HSC70 (α-HSC70) in DOX-inducible mock or indicated RNH1-expressing HeLa (A) or Jurkat (B) cells after incubation with or without DOX for 24 h.

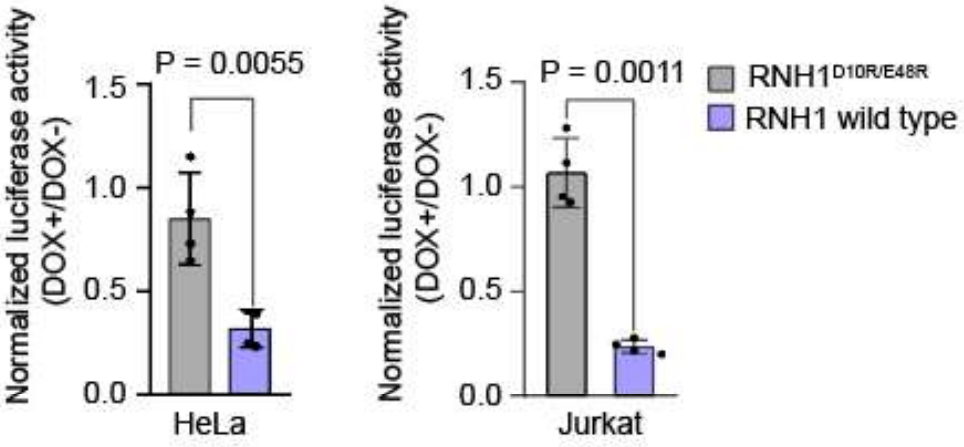
Normalized luciferase activity
(DOX+/DOX-)
P = 0.0055
1.5
1.0
0.5
0.0
HeLa
Normalized luciferase activity
(DOX+/DOX-)
P = 0.0011
1.5
1.0
0.5
0.0
Jurkat
RNH1D10R/E48R
RNH1 wild type

Figure 15. HIV-1 infectivity in cells ectopically expressing RNase H1.

Bar graphs indicating luciferase activity at 48 hpi in DOX-inducible $RNH1^{D10R/E18R}$ or RNH1 wild type-expressing HeLa and Jurkat T cells infected with100ng/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 = 4).

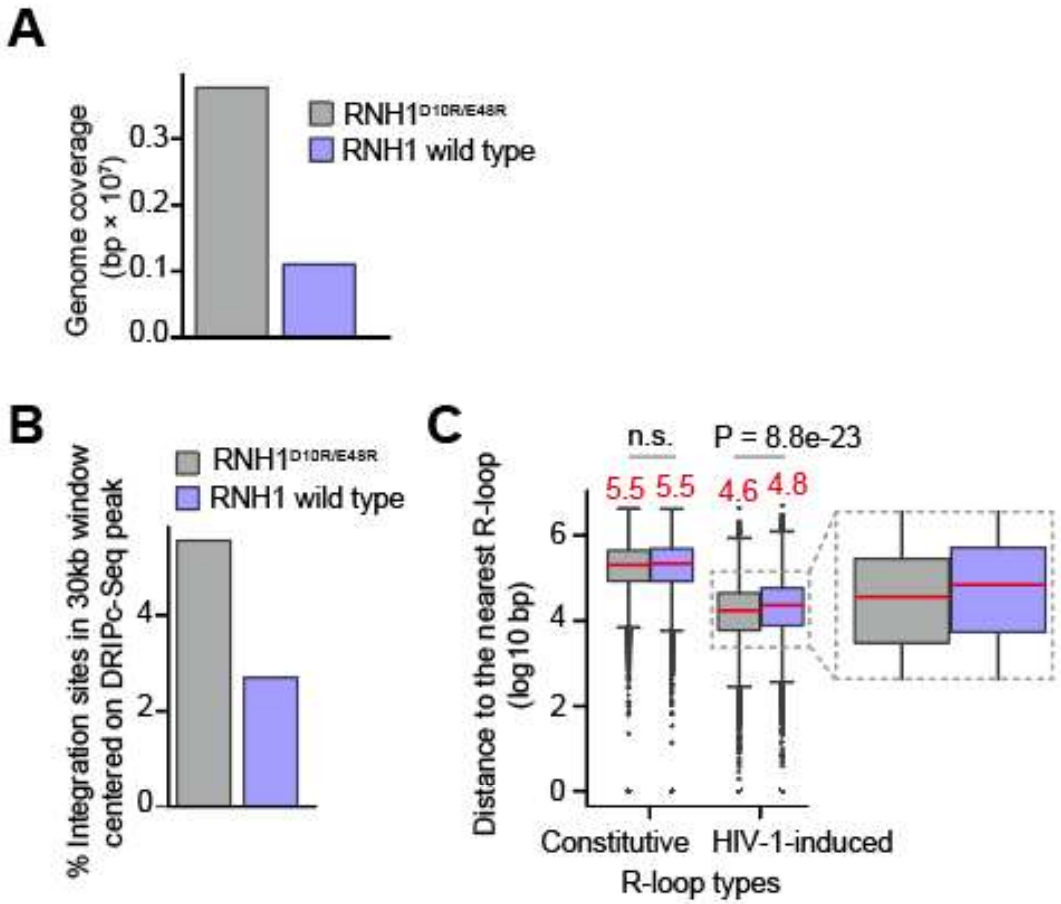
A
Genome coverage (bp × 10⁷)
0.3
0.2
0.1
0.0
RNH1D10R/E48R
RNH1 wild type
B
% Integration sites in 30kb window centered on DRIPc-Seq peak
4
2
0
RNH1D10R/E48R
RNH1 wild type
C
n.s.
P = 8.8e-23
5.5 5.5
4.6 4.8
Distance to the nearest R-loop (log10 bp)
6
4
2
0
Constitutive
HIV-1-induced
R-loop types

Figure 16. Host genomic R-loops and HIV-1 integration sites in cells ectopically expressing RNase H1.

(A) Bar graph indicating the genome coverage of DRIPc-seq peaks of uninfected HeLa cells ectopically expressing RNH1 wild type (purple), or RNH1$^{D10R/E18R}$ control (light gray). (B) Bar graph showing quantified proportion of HIV-1 integration within the 30-kb windows centered on DRIPc-seq peaks, in the host cell genome of HeLa cells ectopically expressing RNH1 wild type (purple), or RNH1$^{D10R/E18R}$ control (light gray). (C) Box plot indicating the distance from HIV-1 integration sites to each group of R-loops (constitutive or HIV-1-induced R-loops) in HeLa cells ectopically expressing RNH1 wild type (purple), or RNH1$^{D10R/E18R}$ control (light gray). The log10 mean distance from the individual HIV-1 integration sites to the HIV-1-induced R-loop region is indicated in red. Statistical significance was assessed using a two-tailed independent t-test.

3.5. HIV-1 integration sites are enriched at systemically induced sequence-specific R-loop regions in cell model

To more directly assess the relationship between host genomic R-loops and HIV-1 integration, I adapted and modified an elegantly designed episomal system that induces sequence specific R-loops through DOX-inducible promoters (Hamperl et al., 2017). Rather than simply adapting the episomal R-loop forming vector system, I integrated the R-loop forming and non-R-loop forming control sequences into the cellular genome by piggyBac transposon-transposase system to most closely mimic the host genomic R-loop induction during HIV-1 infection. I subcloned the R-loop-forming portion of the mouse gene encoding AIRN (mAIRN) (Ginno et al., 2012) or non-R-loop-forming ECFP sequence with a DOX-inducible promoter into the piggyBac transposon vector. I expressed the piggyBac transposase in HeLa cells transfected with piggyBac transposon vector with R-loop forming or non-R-loop forming sequence, which are non-human sequences that can be distinguished from uncontrollable cellular R-loops sequences. I designated the pool of cells with the R-loop forming sequence (mAIRN) inserted into its genome as "pgR-rich (piggyBac R-loop rich)" cell line and the pool of cells with the non-R-loop forming sequence (ECFP) inserted into its genome as "pgR-poor (piggyBac R-loop poor)" cell line (Figure 17).

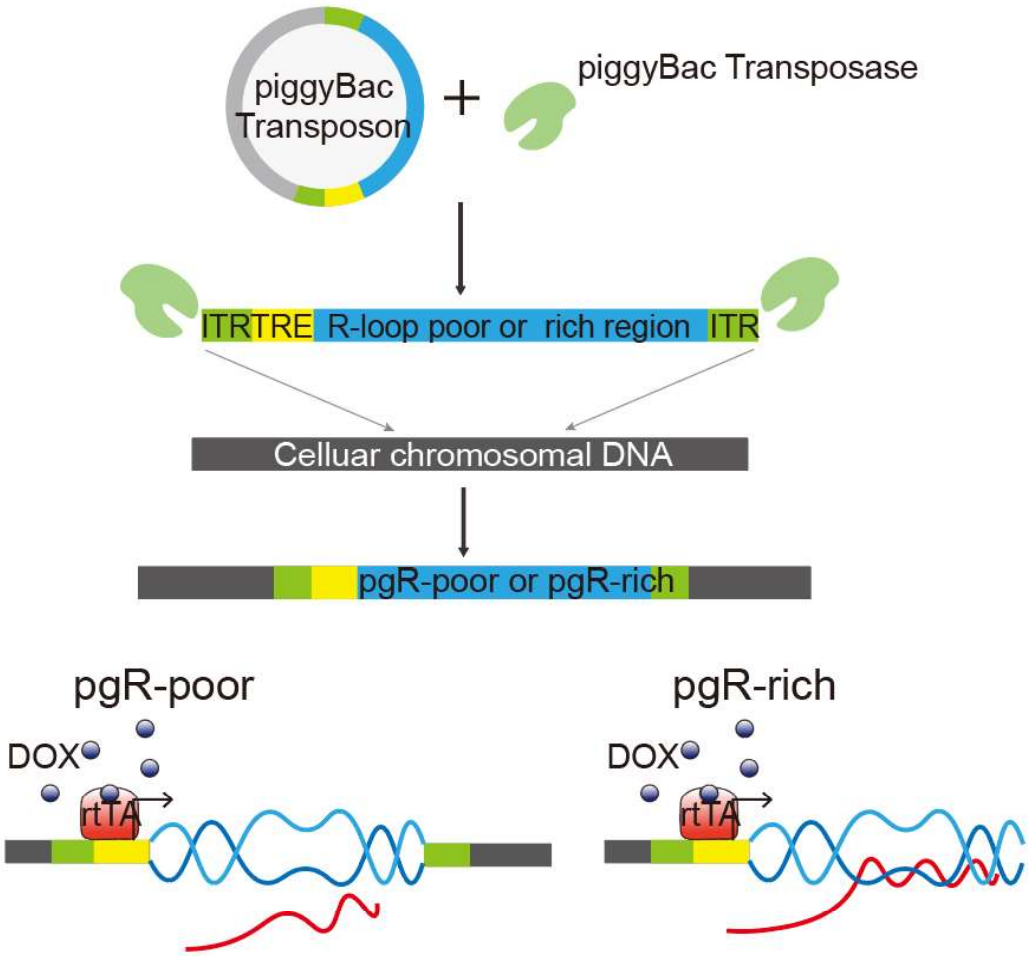
piggyBac
Transposon
piggyBac Transposase
ITRTRE R-loop poor or rich region ITR
Celluar chromosomal DNA
pgR-poor or pgR-rich
pgR-poor
DOX
rtTA
pgR-rich
DOX
rtTA

Figure 17. Summary of the experimental design for R-loop inducible cell lines, pgR-poor and pgR-rich.

R-loop-forming portion of the mouse gene encoding AIRN (mAIRN) or non-R-loop-forming ECFP sequence (blue) with a DOX-inducible promoter (yellow) were subcloned into the piggyBac transposon vector (light gray). Co-expression of the piggyBac Transposon vectors and piggyBac Transposases integrate the transposons into cellular chromosomal DNA (dark gray). DOX treatment activates transcription and systemically induced genomic R-loops only in pgR-rich cell line with R-loop-forming sequence (mAIRN) form R-loops but not in pgR-poor cell lines.

A similar number of the copies of piggyBac transposon was successfully delivered to the genome of each cell line (Table 8), and DOX treatment strongly induced the transcriptional activity of mAIRN or ECFP without affecting the transcription of endogenous loci in both cell lines (Figure 18A and 18B). Although the transcription of mAIRN or ECFP was strongly induced upon DOX treatment, the activity did not exceed that of endogenous loci in both cell lines (Figure 19A and 19B).

Table 8. Copy number of piggyBac transposon inserts in each cell line constructed by transfecting the transposon vector and transposase-expressing vector.

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

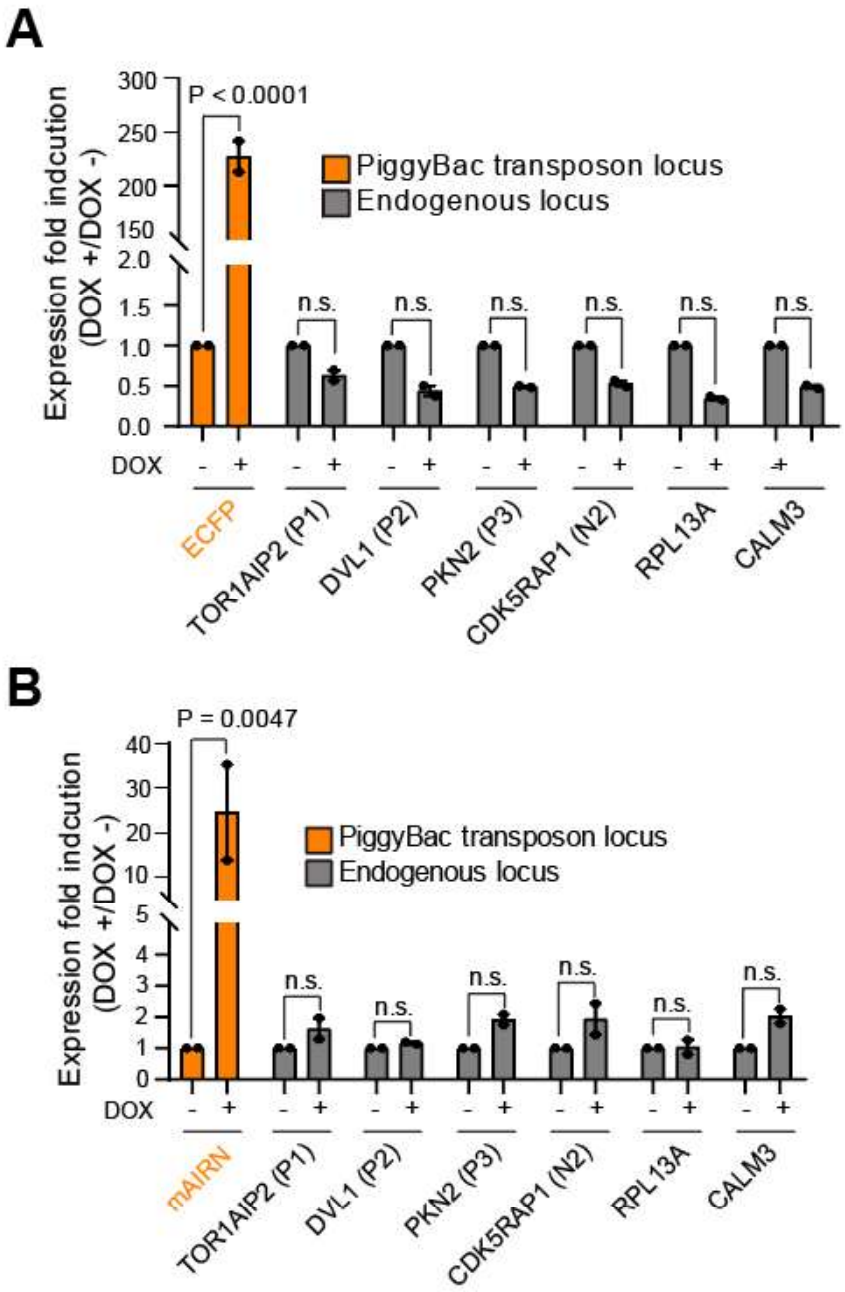
A
P < 0.0001
PiggyBac transposon locus
Endogenous locus
Expression fold indcution
(DOX +/DOX -)
n.s.
DOX
ECFP
TOR1AIP2 (P1)
DVL1 (P2)
PKN2 (P3)
CDK5RAP1 (N2)
RPL13A
CALM3
B
P = 0.0047
PiggyBac transposon locus
Endogenous locus
Expression fold indcution
(DOX +/DOX -)
n.s.
DOX
mAIRN
TOR1AIP2 (P1)
DVL1 (P2)
PKN2 (P3)
CDK5RAP1 (N2)
RPL13A
CALM3

Figure 18. Relative gene expression of piggyBac transposon and endogenous loci upon DOX treatment in pgR−poor and pgR−rich HeLa cells.

(A and B) Relative gene expression of the indicated genes as measured by RT−qPCR in DOX−treated (+) or DOX−untreated (−) pgR−poor cells (A) or pgR−rich cells (B). Data represent mean ± SEM, n = 2, P values were calculated according to two−way ANOVA. P > 0.05; n.s, not significant.

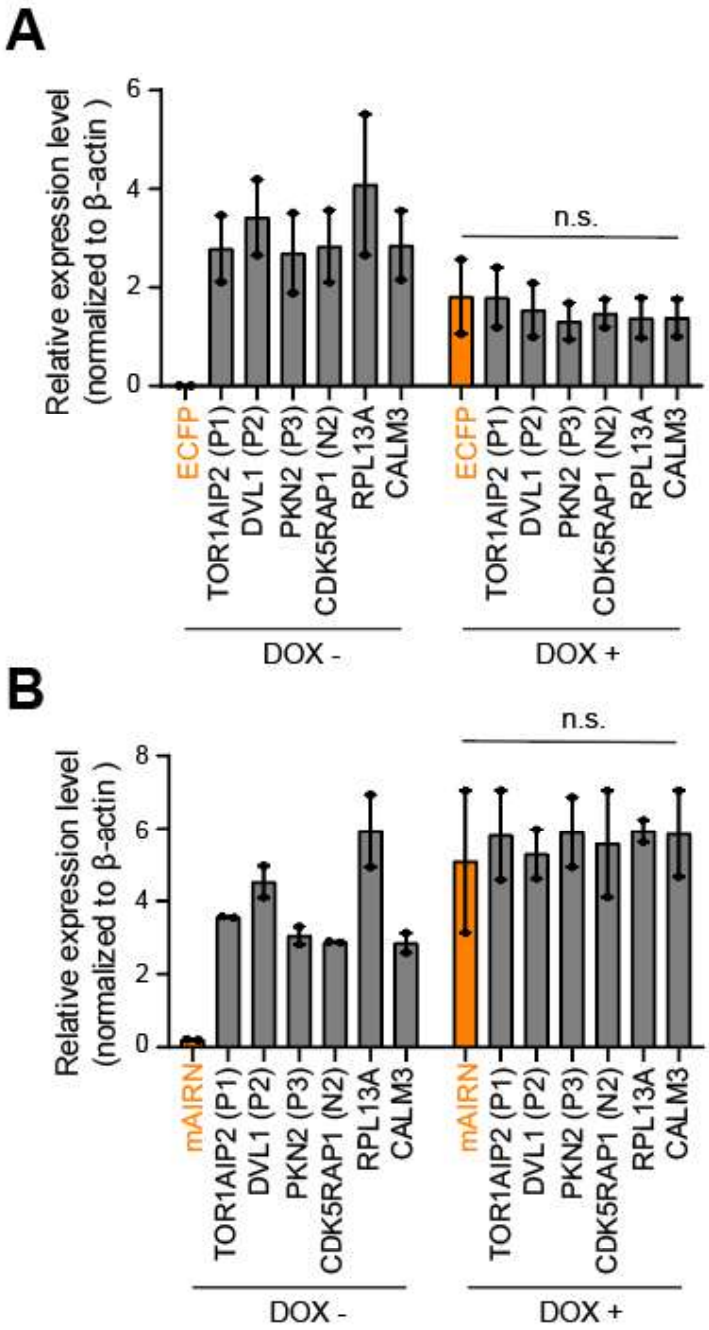
A
Relative expression level
(normalized to β-actin)
6
4
2
0
n.s.
ECFP
TOR1AIP2 (P1)
DVL1 (P2)
PKN2 (P3)
CDK5RAP1 (N2)
RPL13A
CALM3
ECFP
TOR1AIP2 (P1)
DVL1 (P2)
PKN2 (P3)
CDK5RAP1 (N2)
RPL13A
CALM3
DOX -
DOX +
B
Relative expression level
(normalized to β-actin)
8
6
4
2
0
n.s.
mAIRN
TOR1AIP2 (P1)
DVL1 (P2)
PKN2 (P3)
CDK5RAP1 (N2)
RPL13A
CALM3
mAIRN
TOR1AIP2 (P1)
DVL1 (P2)
PKN2 (P3)
CDK5RAP1 (N2)
RPL13A
CALM3
DOX -
DOX +

Figure 19. Fold induction of gene expression of piggyBac transposon and endogenous loci upon DOX treatment in pgR−poor and pgR−rich HeLa cells.

(A and B) 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 (A) or pgR−rich cells (B). Data represent mean ± SEM, n = 2, P values were calculated according to two−way ANOVA. $P > 0.05$; n.s, not significant.

To determine whether DOX-dependent transcription induces the sequence specific R-loop formation only in pgR-rich cells, I performed RT-qPCR and DNA-RNA immunoprecipitation followed by real-time polymerase chain reaction (qPCR; DRIP-qPCR) using primers that were specific to mAIRN or ECFP sequences. While two cell lines showed comparable level of DOX-inducible transcription activity (Figure 20A), only pgR-rich cells exhibited robust RNase H-sensitive stable R-loop formation upon DOX treatment (Figure 19B, mAIRN). By contrast, R-loops were weakly formed in the pgR-poor cells (Figure 20B, ECFP).

To examine whether the formation of 'extra' R-loops in the host genome influence HIV-1-infection to the host cells, I infected both cell lines with VSV-G-pseudotyped HIV-1-luciferase viruses and harvested at 48 hpi for HIV-1 luciferase activity examination. Interestingly, I found that pgR-rich cells showed significantly high luciferase activity only when R-loops were induced by DOX treatment, whereas pgR-poor cells showed comparable luciferase activity regardless of transcription activation by DOX treatment (Figure 21A). This data indicates that R-loop formation in the host genome positively affect HIV-1 infectivity or viral gene expression. I conducted HIV-1 integration site sequencing in HIV-1-infected pgR-poor and pgR-rich cells to directly quantify site-specific integration events at sequence-specific R-loop regions. I aligned the integration site sequencing reads to the human reference genome as

well as the piggyBac transposon cargo sequences. Remarkably, integration events were significantly higher in pgR-rich cells only when R-loops were induced by DOX treatment (Figure 21B). However, HIV-1 integration frequency within non-R-loop forming sequence in pgR-poor cells remained very low, even with transcription activation by DOX treatment (Figure 21B). HIV-1 integration frequency was much higher at the vicinity of R-loop forming regions only in pgR-rich cell line upon DOX treatment (Figure 21C and 21D). This cell-based R-loop inducing system with independent control over transcription and R-loop formation enabled the direct measurement of HIV-1 integration events at the defined R-loop regions, and the results indicate that host genomic R-loops are targeted by HIV-1 integration. Moreover, these data suggest that transcription itself is not sufficient for HIV-1 integration site determination, but the presence of R-loops accounts for HIV-1 integration site selection.

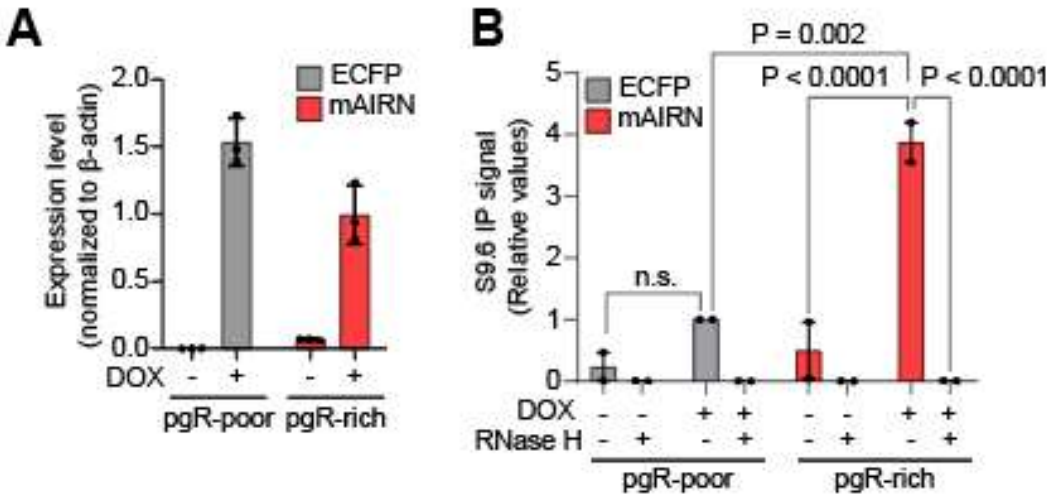
A
Expression level (normalized to β-actin)
2.0
1.5
1.0
0.5
0.0
ECFP
mAIRN
DOX - + - +
pgR-poor pgR-rich
B
S9.6 IP signal (Relative values)
5
4
3
2
1
0
ECFP
mAIRN
P = 0.002
P < 0.0001
P < 0.0001
n.s.
DOX - - + + - - + +
RNase H - + - + - + - +
pgR-poor
pgR-rich

Figure 20. R-loop inducible cell lines induce loci specific R-loop formation independently of gene expression level.

(A) 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. (B)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).

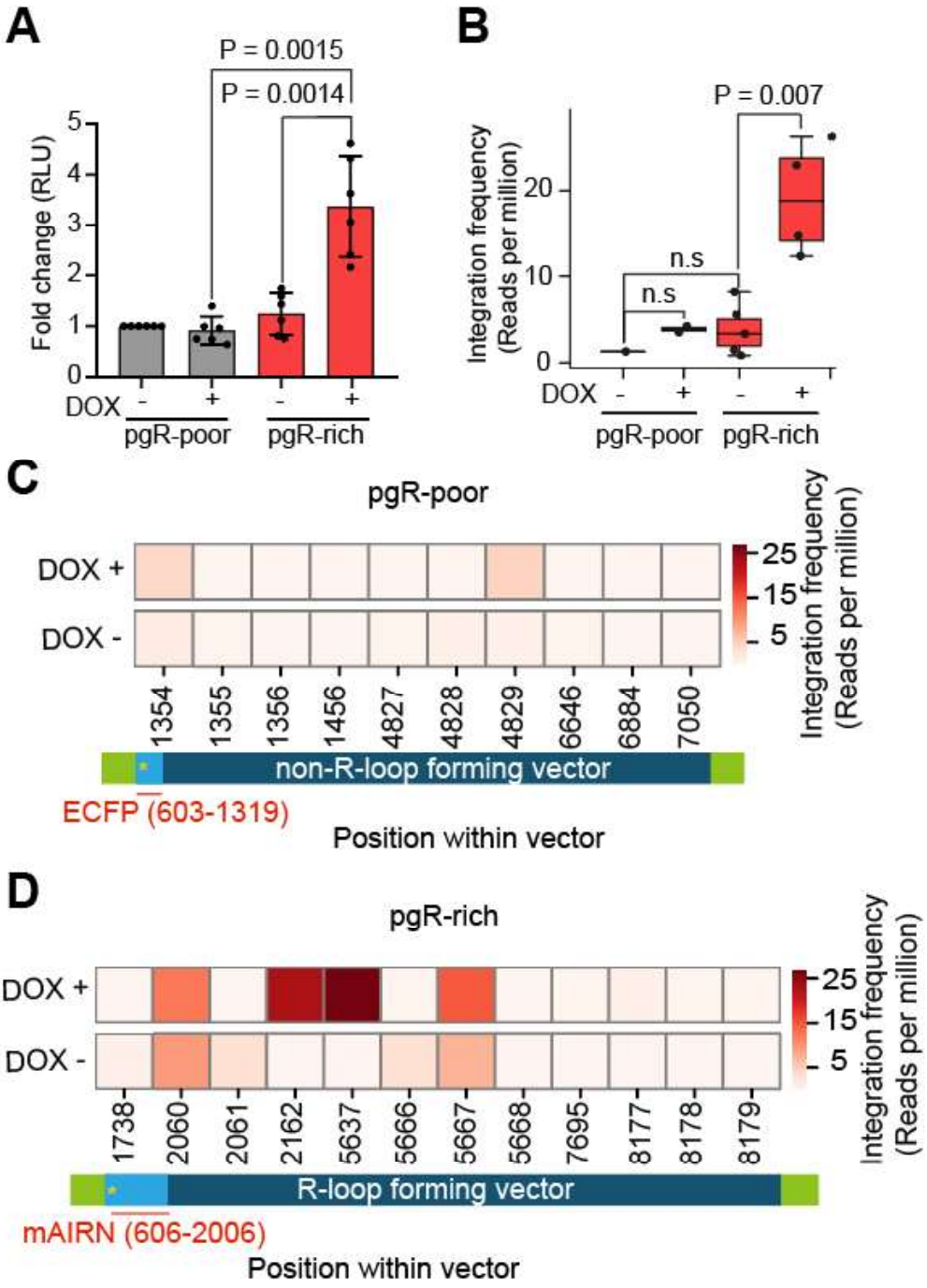
A
P = 0.0015
P = 0.0014
Fold change (RLU)
DOX
pgR-poor
pgR-rich
B
P = 0.007
n.s
n.s
Integration frequency (Reads per million)
C
pgR-poor
DOX +
DOX -
1354
1355
1356
1456
4827
4828
4829
6646
6884
7050
non-R-loop forming vector
ECFP (603-1319)
Position within vector
D
pgR-rich
1738
2060
2061
2162
5637
5666
5667
5668
7695
8177
8178
8179
R-loop forming vector
mAIRN (606-2006)

Figure 21. R-loop inducible cell line model directly addresses R-loop-mediated HIV-1 integration site selection.

(A) 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× 105 cells/mL. Data are presented as the mean ± SEM; P values were calculated using one-way ANOVA (n = 6). (B) 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 capsid antigen of luciferase reporter HIV-1 virus per 1× 105 cells/mL. Each bar corresponds to pooled datasets from three biologically independent experiments (n =3). In each boxplot, the centerline denotes the median, the upper and lower box limits denote the upper and lower quartiles, and the whiskers denote the 1.5 × interquartile range. Statistical significance was assessed using a two-sided Mann–Whitney U test. (C and D) Heat maps representing HIV-1 integration frequency across pgR-poor (C) or pgR-rich (D) transposon sequence in untreated (-) or DOX-treated (+) pgR-poor (C) or pgR-rich (D) cell line. Each rectangular box corresponds to the pooled integration frequency from three biologically independent experiments (n =3) at the indicated position within pgR-poor (C) or pgR-rich (D) transposon vector. Each light blue box represents actual position of R-loop forming or non-R-loop forming sequence (ECFP or mAIRN) and the yellows stars indicate TRE promoter position within vector.

3.6. HIV-1 exploits the HIV-1-induced host genomic R-loops for viral DNA integration

I further validate the global relationship between R-loops and the HIV-1 integration site selection. I performed HIV-1 integration site sequencing on naive HeLa cells, primary $CD4^+$ T cells and Jurkat cells infected with VSV-G-pseudotyped HIV-1-EGFP and analyzed the sequencing data combined with the DRIPc-seq data. I mapped the HIV-1 integration site sequencing on the 30-kb windows centered on DRIPc-seq peaks at 0, 3, 6, and 12 hpi. Interestingly, a significantly higher proportion of HIV-1 integration occurred within the R-loop windows at 3 and 6 hpi (33% and 35%, respectively) than at 0 and 12 hpi (8% and 10%, respectively; Figure 22A) for HeLa cells, Higher proportions of HIV-1 integration occurred within the R-loop windows at 6 and 12 hpi than at 0 and 3 hpi, in primary $CD4^+$ T cells and Jurkat cells (Figure 22B and 22C). To investigate the extent to which R-loops influence HIV-1 integration site selection, I counted and compared the number of successfully integrated proviruses in the R-loop regions (the combined genomic regions within 30-kb windows centered on DRIPc-seq peaks from 0, 3, 6, and 12 hpi) to those in non-R-loop forming regions (the total genomic regions outside of the 30-kb windows centered on DRIPc-seq peaks). Notably, I found that approximately three times more integration sites were detected in the R-loop regions as in other genomic regions without R-loops (Figure 22D), in HeLa cells. R-loop regions were preferred by HIV-1 integration more than three-

folds in both primary $CD4^+$ T cells and Jurkat cells (Figure 22E and 22F). Overall, these results from bioinformatics analysis using naïve host cells infected with HIV-1 are consistent with the idea that the virus has a preference for targeting R-loops for integration (Figure 21).

I then compared the proportion of R-loop-dependent HIV-1 integration events around constitutive and HIV-1-induced R-loops, in HeLa cells. Interestingly, I found that a significant proportion of R-loop-dependent HIV-1 integration sites were within the 30-kb windows centered on HIV-1-induced R-loops (85.5%) rather than within the 30-kb windows centered on constitutive R-loops (3.6%; Figure 23A). The chromosomal locations of constitutive R-loops were not particularly related to HIV-1 integration sites (Figure 23B, left), but HIV-1 integration sites tended to be located in the center and nearby areas of the R-loops induced by HIV-1 infection (Figure 23B, right).

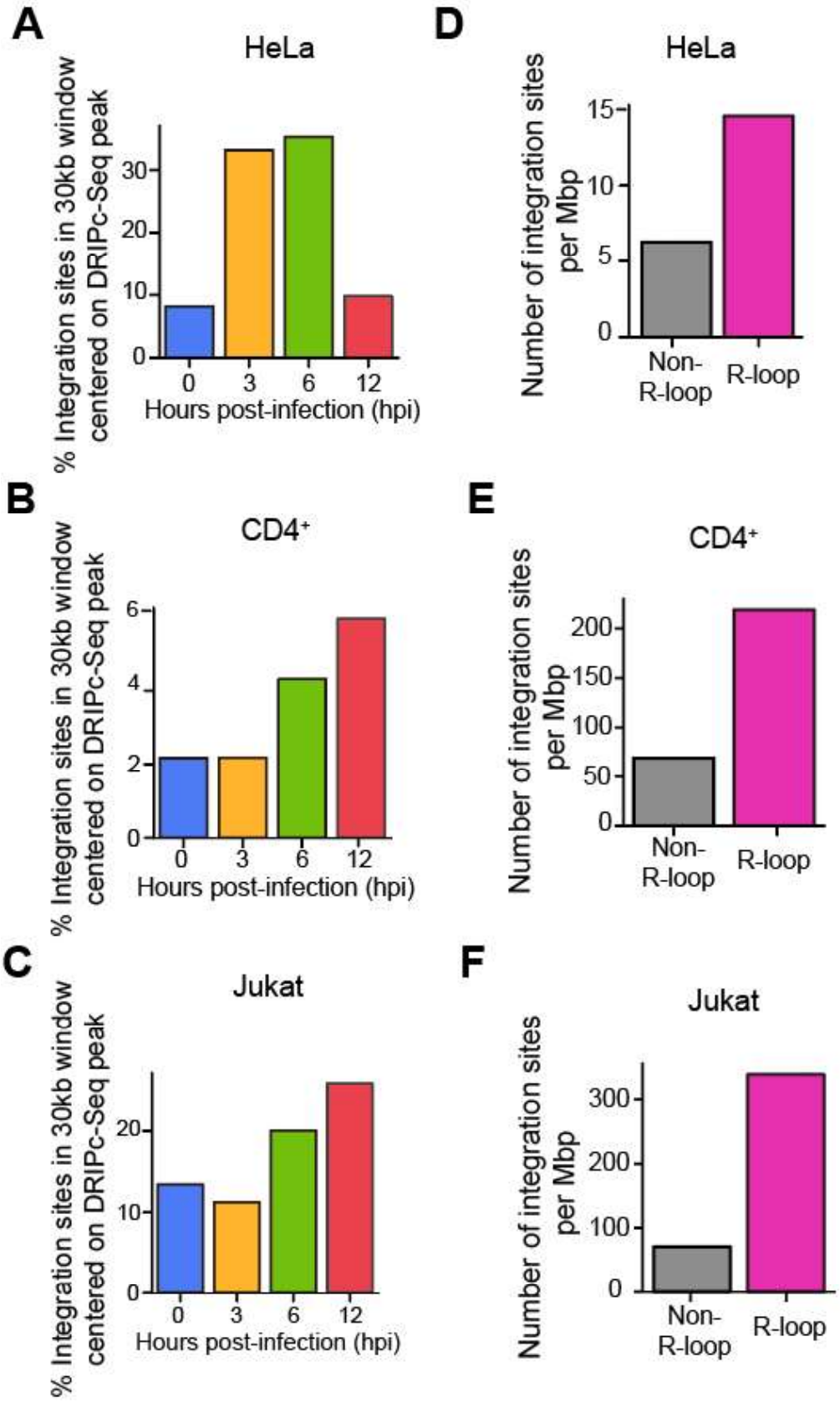
A
HeLa
% Integration sites in 30kb window centered on DRIPc-Seq peak
0
10
20
30
0
3
6
12
Hours post-infection (hpi)
B
CD4+
% Integration sites in 30kb window centered on DRIPc-Seq peak
0
2
4
6
0
3
6
12
Hours post-infection (hpi)
C
Jukat
% Integration sites in 30kb window centered on DRIPc-Seq peak
0
10
20
0
3
6
12
Hours post-infection (hpi)
D
HeLa
Number of integration sites per Mbp
0
5
10
15
Non-R-loop
R-loop
E
CD4+
Number of integration sites per Mbp
0
50
100
150
200
Non-R-loop
R-loop
F
Jukat
Number of integration sites per Mbp
0
100
200
300
Non-R-loop
R-loop

Figure 22. HIV-1 targets host genomic R-loop for its viral cDNA integration.

(A-C) Bar graph showing the quantified proportion of HIV-1 integration within the 30-kb windows centered on R-loops from each indicated hpi (blue, 0 hpi; yellow, 3 hpi; green, 6 hpi; red, 12 hpi), in HeLa cells (A), primary $CD4^+$ T cells (B) and Jurkat cells (C). (D-E) Bar graph 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 (D), primary $CD4^+$ T cells (E) and Jurkat cells (F) (magenta) or non-R-loop region in the cellular genome (gray).

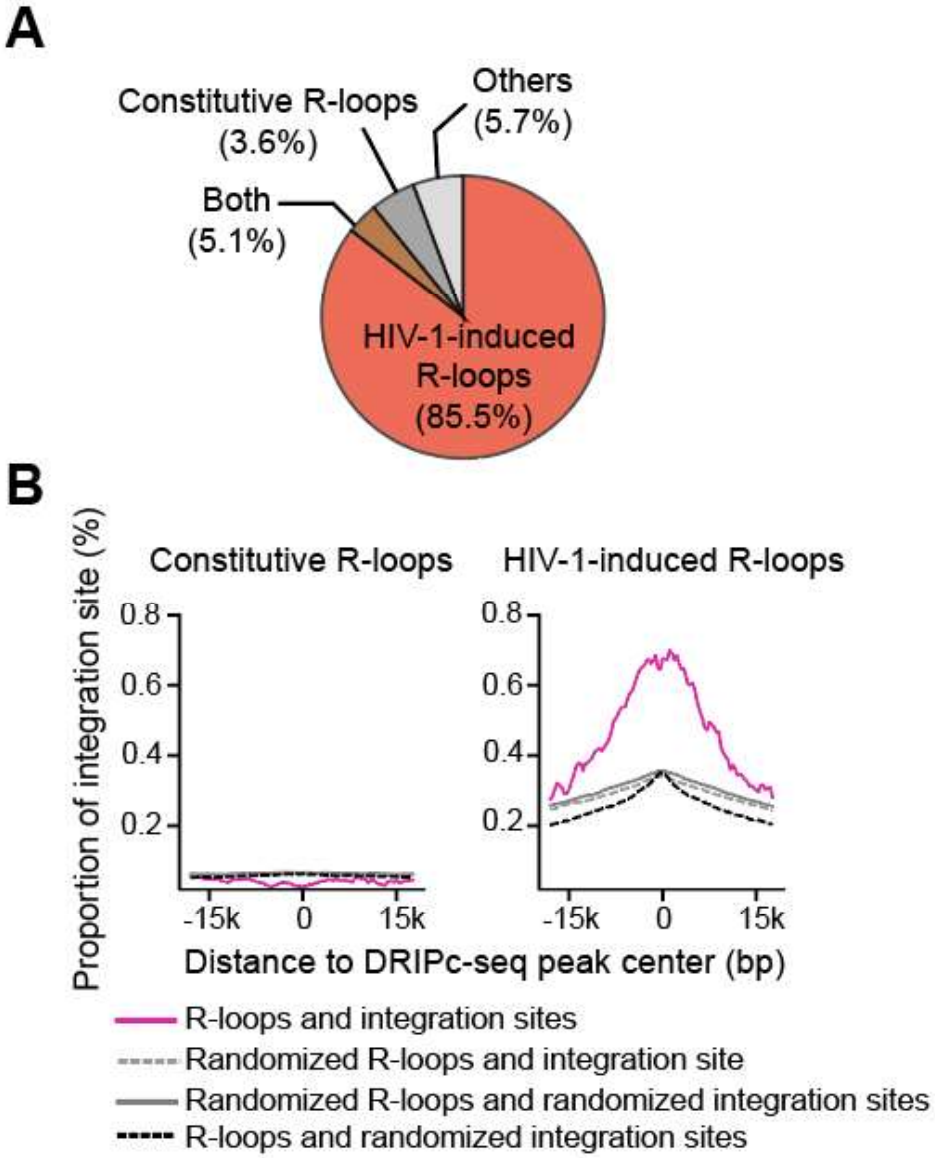
A
Constitutive R-loops
(3.6%)
Others
(5.7%)
Both
(5.1%)
HIV-1-induced
R-loops
(85.5%)
B
Constitutive R-loops
HIV-1-induced R-loops
Proportion of integration site (%)
0.8
0.6
0.4
0.2
-15k
0
15k
Distance to DRIPc-seq peak center (bp)
R-loops and integration sites
Randomized R-loops and integration site
Randomized R-loops and randomized integration sites
R-loops and randomized integration sites

Figure 23. HIV-1 preferentially integrate it viral genome into HIV-1-induced-R-loops in HeLa cells.

(A) A pie graph indicating the percentage of HIV-1 integration site within the 30-kb windows centered on HIV-1-induced R-loops (red), constitutive R-loops (light gray), both types of R-loops (brown), and unannotated other R-loops (light gray), out of the 30-kb windows centered on the total consensus DRIPc-seq peaks from HIV-infected HeLa cells. (B) Proportion of integration sites within the 30-kb windows centered on constitutive or HIV-1-induced 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.

I validated the global analysis of the relationship between host genomic R-loops and HIV-1 integration in a genome-site specific manner, in HeLa cells. First, I verified R-loop induction in HIV-1-infected cells using DRIP-qPCR. In this experiment, the S9.6 signal was determined for three and two HIV-1-induced-R-loop-positive (P1, P2, and P3) and -negative regions (N1 and N2), respectively, which were defined by DRIPc-seq data analysis. I detected significantly increased R-loop signals in the P1, P2, and P3 regions in HIV-1-infected cells at 6 hpi compared to those in the control (cells harvested at 0 hpi) (Figure 24A). However, the HIV-1-induced R-loop-negative regions, N1 and N2, did not show significant R-loop accumulations (Figure 24A). I conducted RT-qPCR and analyzed RNA-seq data for genes harboring the R-loop regions. The transcription activity of the genes harboring HIV-1-infection induced R-loops were not significantly altered (Figure 24B). I observed biases for HIV-1 integration in HIV-1-induced R-loop-positive regions where showed sufficient R-loop inductions in DRIP-qPCR, P2 and P3 (Figure 25A and 25B). By contrast, HIV-1 integration sites were not detected in R-loop-negative regions, N1 and N2 (Figure 25C and 25D).

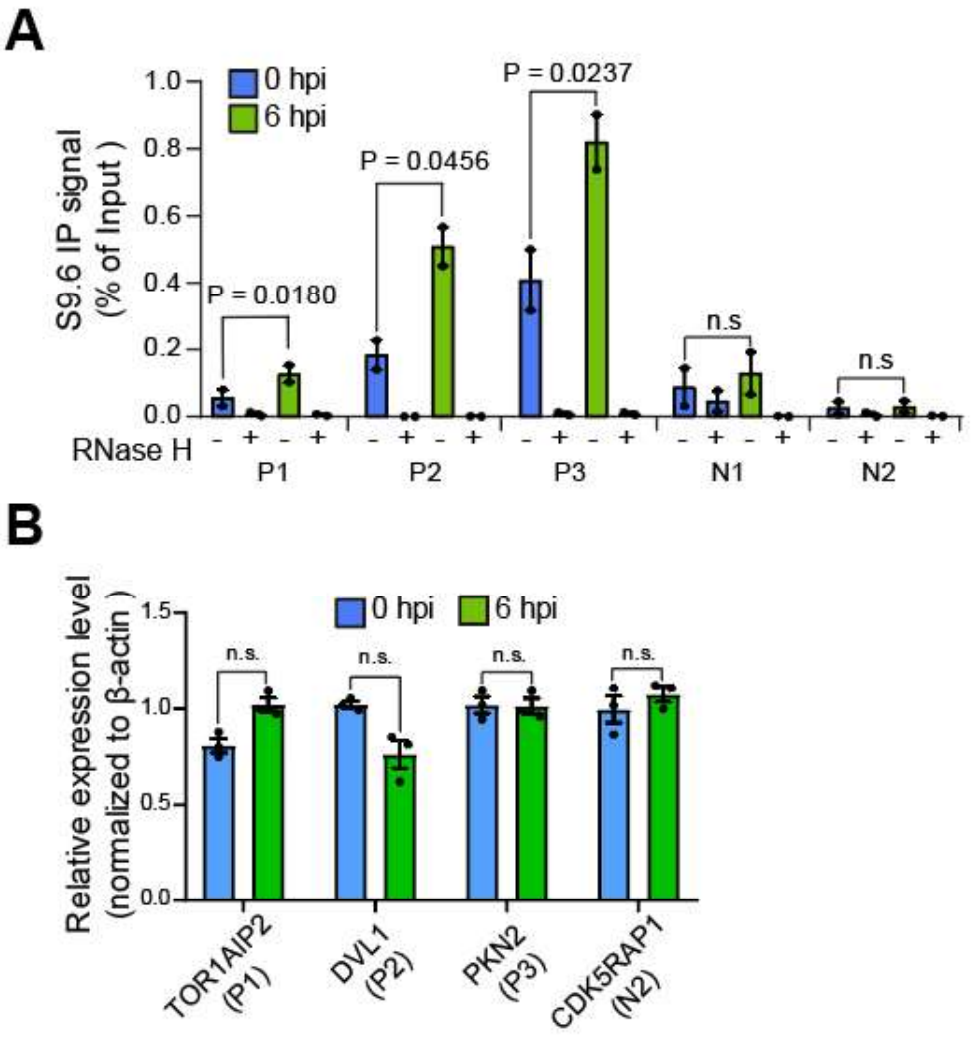
A
S9.6 IP signal
(% of Input)
0 hpi
6 hpi
P = 0.0180
P = 0.0456
P = 0.0237
n.s
n.s
RNase H
P1
P2
P3
N1
N2
B
Relative expression level
(normalized to β-actin)
0 hpi
6 hpi
n.s.
TOR1AIP2
(P1)
DVL1
(P2)
PKN2
(P3)
CDK5RAP1
(N2)

Figure 24. Endogenous loci specific HIV-1-induced-R-loops formation in HeLa cells.

(A) DRIP-qPCR using 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 indicated hpi (blue, 0 hpi; green, 6 hpi). Pre-immunoprecipitated materials were untreated (-) or treated (+) with RNase H, as indicated. Data are presented as the mean ± SEM; P-values were calculated using one-way ANOVA (n = 2). (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.

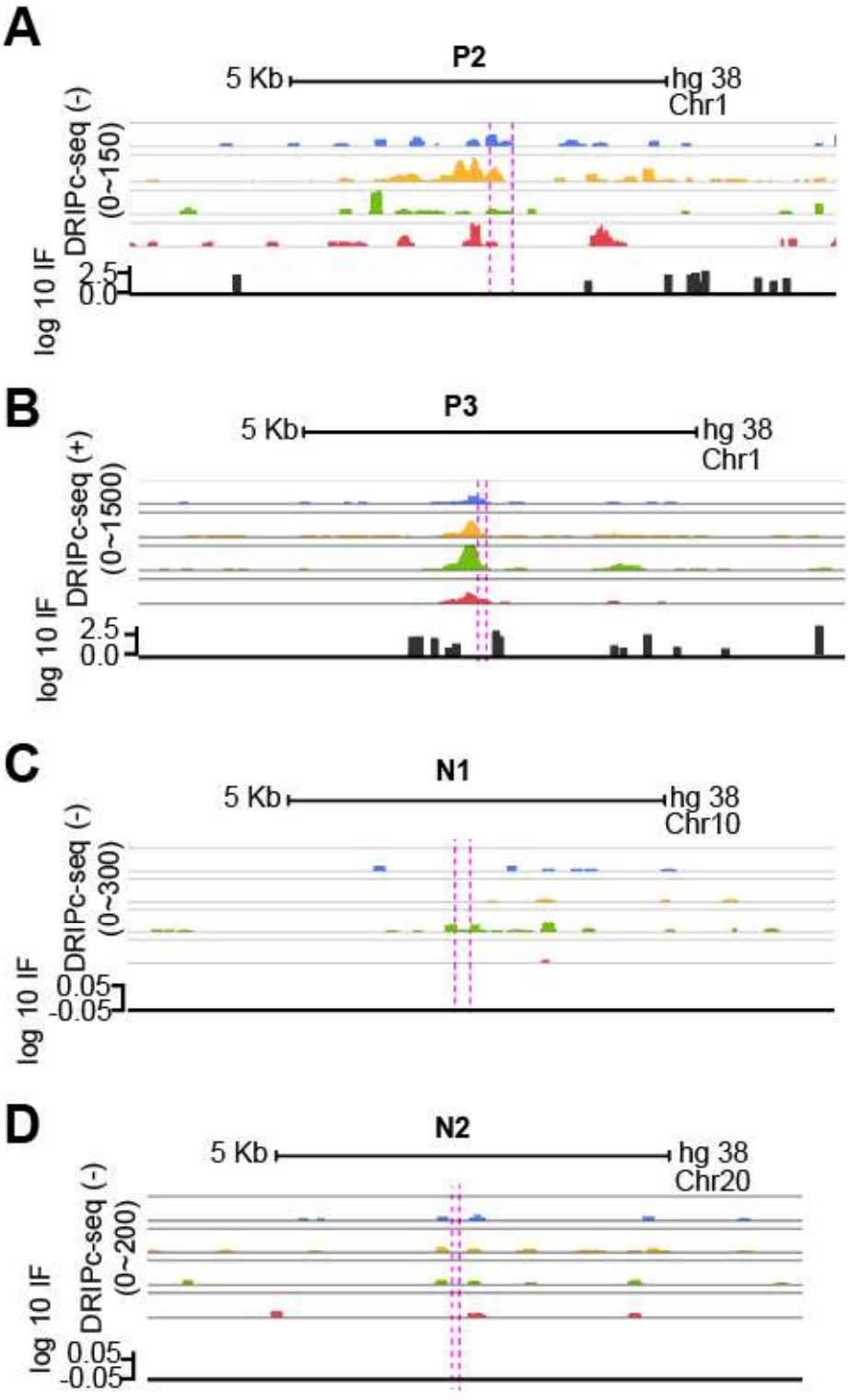
A
P2
5 Kb
hg 38
Chr1
DRIPc-seq (-)
(0~150)
log 10 IF
2.5
0.0
B
P3
5 Kb
hg 38
Chr1
DRIPc-seq (+)
(0~1500)
log 10 IF
2.5
0.0
C
N1
5 Kb
hg 38
Chr10
DRIPc-seq (-)
(0~300)
log 10 IF
0.05
-0.05
D
N2
5 Kb
hg 38
Chr20
DRIPc-seq (-)
(0~200)
log 10 IF
0.05
-0.05

Figure 25. Endogenous loci specific HIV-1-induced-R-loops formation and HIV-1 viral genome integration in HeLa cells.

(A-D) Superimpositions of HIV-1-induced R-loop positive chromatin regions, P2 and P3 (A and B), 773 and HIV-1-induced R-loop negative chromatin regions, N1 and N2 (C and D), on DRIPc-seq (blue, 0 hpi; yellow, 3 hpi; 774 green, 6 hpi; red, 12 hpi) and HIV-1 integration frequency (IF, black).

3.7. HIV-1 integrase physically interacts with R-loops on the host genome

HIV-1 pre-integration complexes (PICs) tether to the host genome for its viral cDNA integration. PICs consist of HIV-1 viral cDNA and HIV-1 coding protein, integrases. HIV-1 preferentially integrated into R-loops on the host genome, thus I hypothesized that the HIV-1 integrase protein could directly bind to the cellular R-loops. To test this hypothesis, I performed DNA-RNA hybrid immunoprecipitation using S9.6 antibodies in FLAG-tagged HIV-1 integrase-expressing HeLa cells (Figure 26). Under these experimental conditions, R-loops were reproducibly immunoprecipitated (Figure 27A) and HIV-1 integrase proteins co-immunoprecipitated with R-loops (Figure 27B). DNA-RNA hybrids also co-immunoprecipitated with the positive control H3 but not with the negative control LaminA/C and Actin (Figure 27B). To verify the specificity of these co-immunoprecipitation results for R-loops and HIV-1 integrases, I performed DNA-RNA hybrid immunoprecipitation with RNase H treatment (Figure 28). The S9.6 signal of immunoprecipitated nucleic acids was highly sensitive to RNase H treatment of pre-immunoprecipitates (Figure 29A). Accordingly, the blotting signal of the co-immunoprecipitated HIV-1 integrase and H3 proteins was significantly reduced when the pre-immunoprecipitates were subjected to RNase H treatment (Figure 29B).

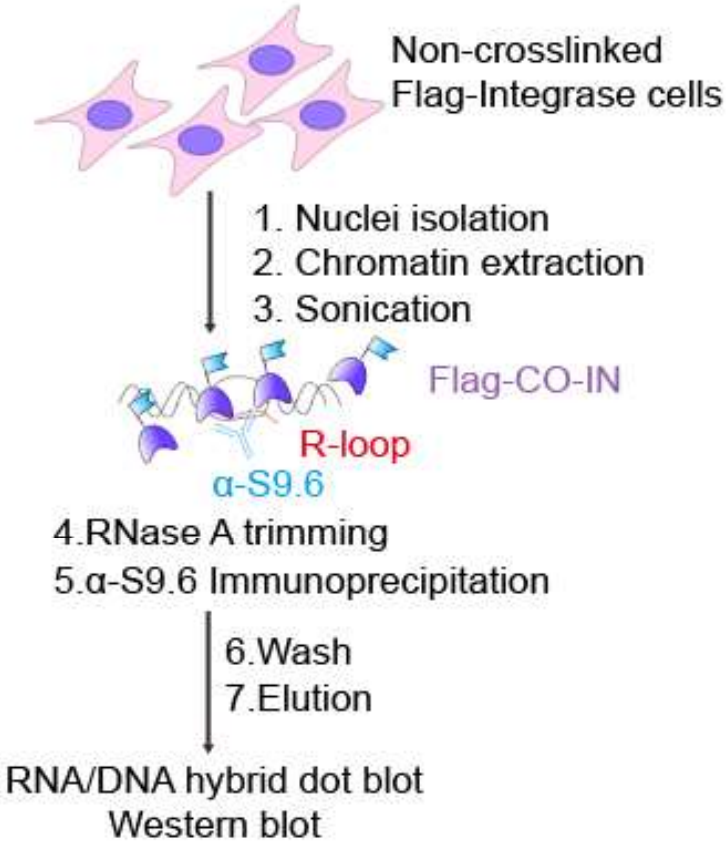
Non-crosslinked
Flag-Integrase cells
1. Nuclei isolation
2. Chromatin extraction
3. Sonication
Flag-CO-IN
R-loop
α-S9.6
4.RNase A trimming
5.α-S9.6 Immunoprecipitation
6.Wash
7.Elution
RNA/DNA hybrid dot blot
Western blot

Figure 26. Summary of the experimental design for R-loop immunoprecipitation using S9.6 antibody in FLAG-tagged HIV-1 integrase protein-expressing HeLa cells.

Nuclei of non-crosslinked HeLa cells ectopically expressing FLAG-tagged HIV-1 integrase proteins were isolated and the chromatin extracts were sonicated. RNase A were treated to the pre-immunoprecipitate materials to remove ssRNA, which non-specifically bind to S9.6 antibodies. After immunoprecipitation by using S9.6 antibodies, the nucleic acid extracts were assessed by dot blotting and the protein extracts were examined by western blotting.

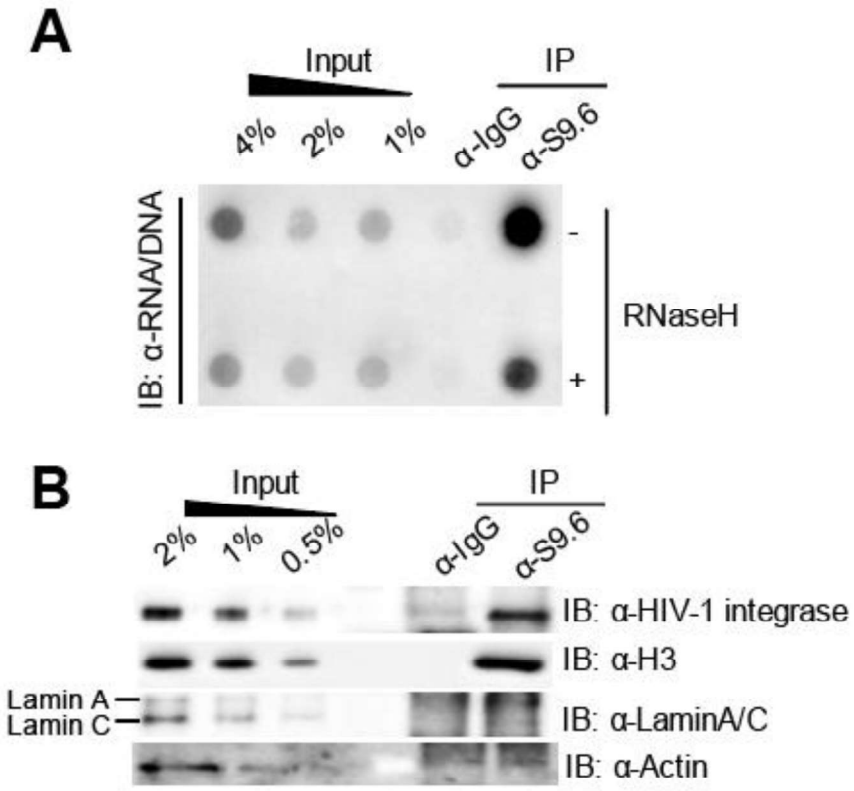
A
Input
IP
4%
2%
1%
α-IgG
α-S9.6
IB: α-RNA/DNA
-
+
RNaseH
B
Input
IP
2%
1%
0.5%
α-IgG
α-S9.6
IB: α-HIV-1 integrase
IB: α-H3
Lamin A
Lamin C
IB: α-LaminA/C
IB: α-Actin

Figure 27. FLAG-tagged HIV-1 integrases are immunoprecipitated by R-loop immunoprecipitation using S9.6 antibody.

(A) gDNA–RNA hybrid extracts from FLAG-HIV-1-integrase-expressing cells were immunoprecipitated using S9.6 antibody. gDNA was precipitated from the elutes of immunoprecipitation and subjected to DNA–RNA hybrid dot blotting. Where indicated, the gDNA extracts were either untreated (–) or treated (+) with RNase H after elution of immunoprecipitated materials. (B) Western blotting for HIV-1 integrase protein, H3, and LaminA/C of DNA–RNA hybrid immunoprecipitation using the S9.6 antibody.

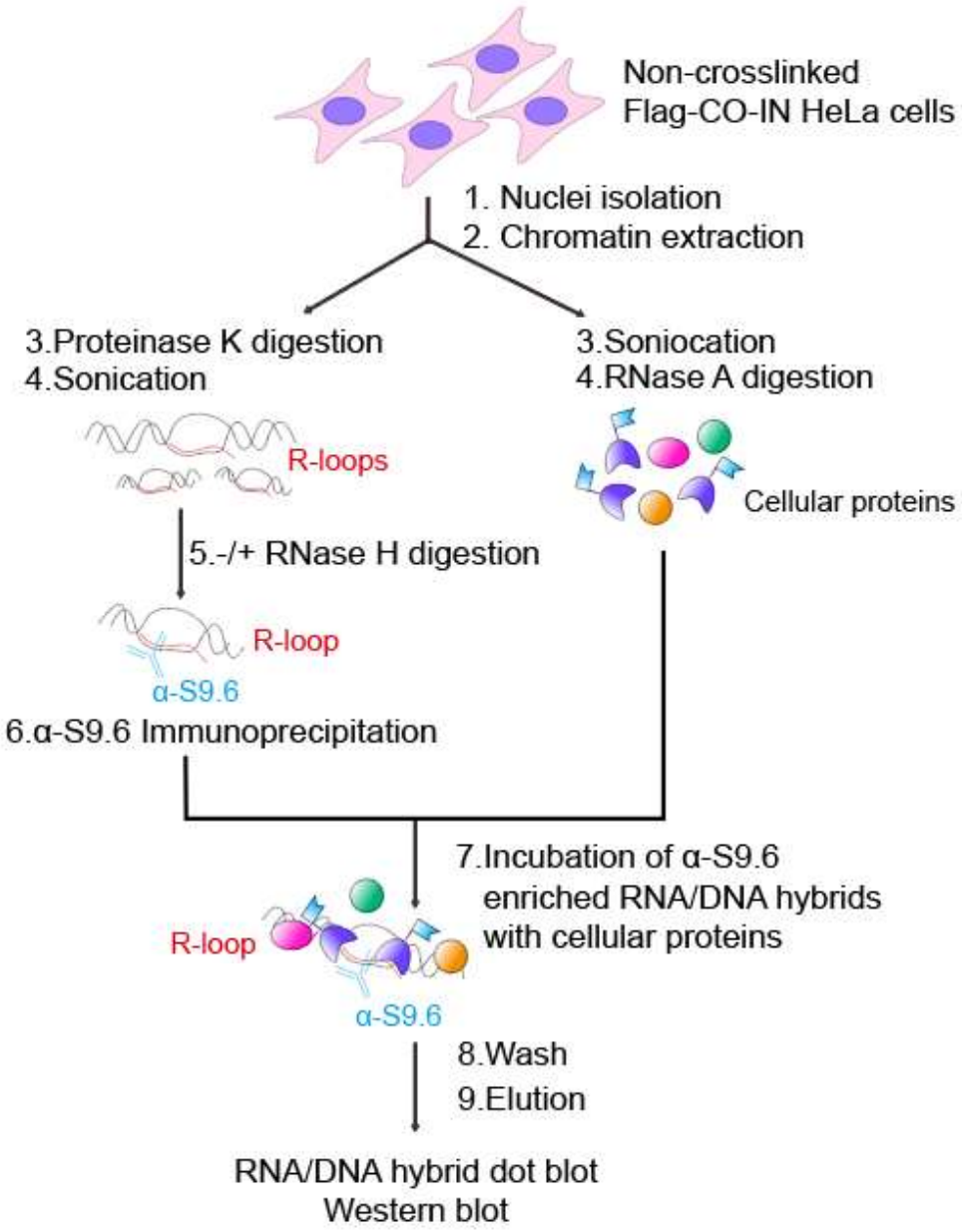
Non-crosslinked
Flag-CO-IN HeLa cells
1. Nuclei isolation
2. Chromatin extraction
3.Proteinase K digestion
4.Sonication
3.Soniocation
4.RNase A digestion
R-loops
Cellular proteins
5.-/+ RNase H digestion
R-loop
α-S9.6
6.α-S9.6 Immunoprecipitation
7.Incubation of α-S9.6
enriched RNA/DNA hybrids
with cellular proteins
R-loop
α-S9.6
8.Wash
9.Elution
RNA/DNA hybrid dot blot
Western blot

Figure 28. Summary of the experimental design for R-loop immunoprecipitation using S9.6 antibody in FLAG-tagged HIV-1 integrase protein-expressing HeLa cells with pre-immunoprecipitation in vitro RNase H treatment.

Nuclei of non-crosslinked HeLa cells ectopically expressing FLAG-tagged HIV-1 integrase proteins were isolated and the chromatin extracts were sonicated. RNase A were treated to the pre-immunoprecipitate materials to remove ssRNA, which non-specifically bind to S9.6 antibodies. Before immunoprecipitation by using S9.6 antibodies, the half of pre-immunoprecipitate materials were treated or left untreated with RNase H1 enzymes. The S9.6 antibody-immunoprecipitated materials were assessed by dot blotting or examined by western blotting.

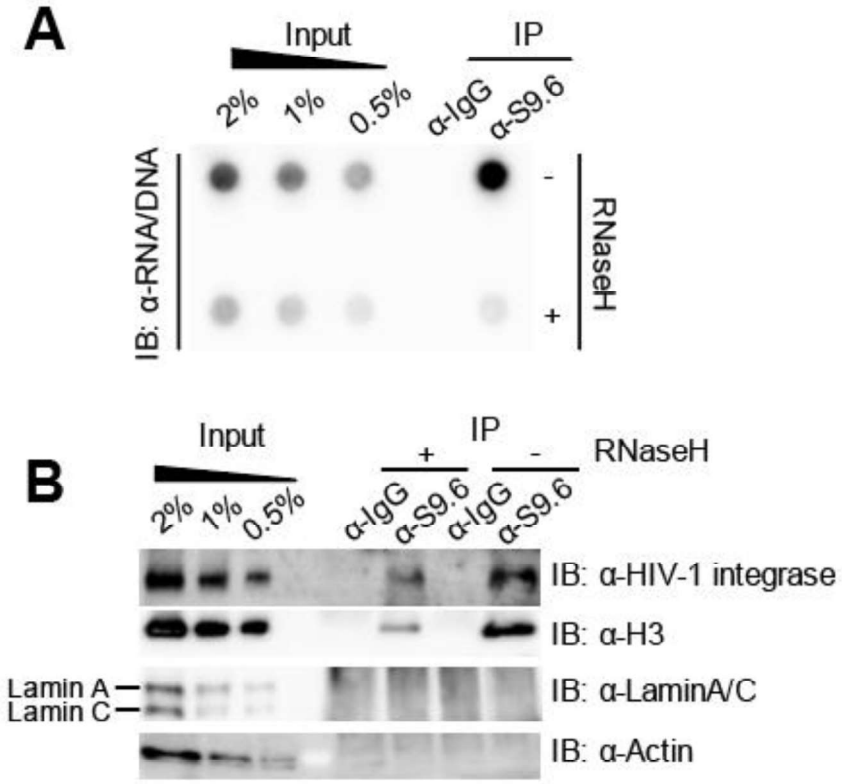
A
Input
IP
2%
1%
0.5%
α-IgG
α-S9.6
IB: α-RNA/DNA
-
+
RNaseH
B
Input
IP
+
-
RNaseH
2%
1%
0.5%
α-IgG
α-S9.6
α-IgG
α-S9.6
IB: α-HIV-1 integrase
IB: α-H3
Lamin A
Lamin C
IB: α-LaminA/C
IB: α-Actin

Figure 29. RNase H1 treatment before immunoprecipitation by S9.6 antibodies reduces immunoprecipitation of FLAG-tagged HIV-1 integrases by R-loop.

(A and B) HeLa gDNA input was either untreated (–) or treated (+) with RNase H before enrichment for DNA–RNA hybrids using the S9.6 antibody. gDNA–RNA hybrids were incubated with nuclear extracts depleted of DNA–RNA hybrids with RNase A followed by S9.6 immunoprecipitation. DNA–RNA hybrid dot blot (A) and western blot of DNA–RNA hybrid immunoprecipitation, probed with the indicated antibodies (B).

I performed reciprocal immunoprecipitation using an anti-FLAG monoclonal antibody and detected immunoprecipitated R-loops using dot blot analysis with anti-S9.6. R-loops were immunoprecipitated by HIV-1 integrase, and the S9.6 signal of immunoprecipitated nucleic acids was highly sensitive to RNase H treatment, while the total DNA content of anti-FLAG immunoprecipites was not affected by RNase H treatment (Figure 30A and 30B). Together, these results demonstrate that HIV-1 integrase proteins and R-loops physically interact in the host cells. I investigated whether HIV-1 integrase proteins and R-loops possess direct physical binding.

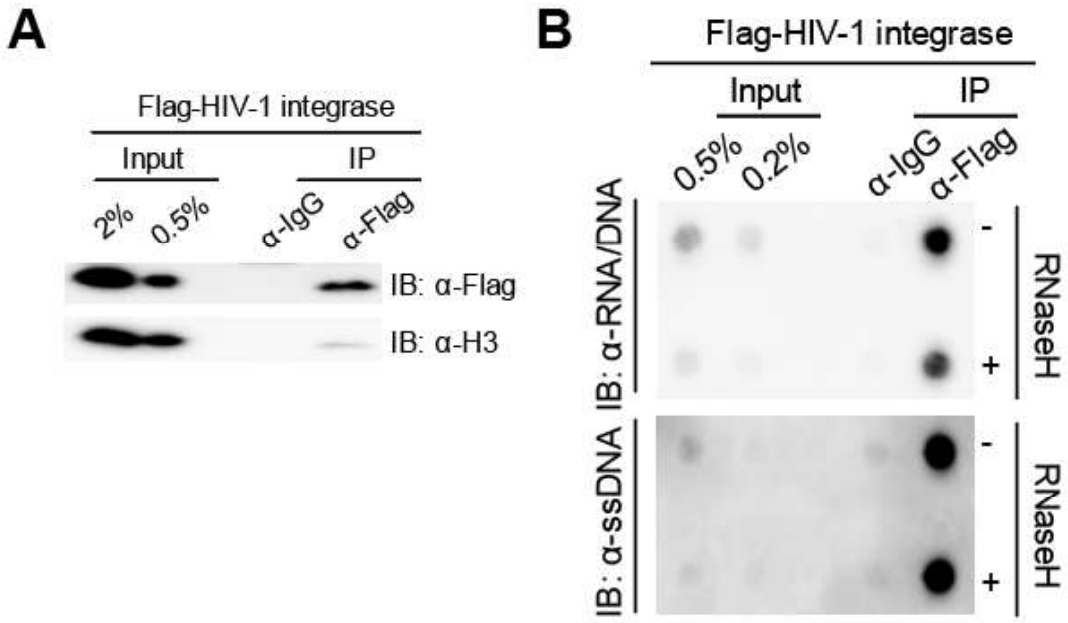
A
Flag-HIV-1 integrase
Input
IP
2%
0.5%
α-IgG
α-Flag
IB: α-Flag
IB: α-H3
B
Flag-HIV-1 integrase
Input
IP
0.5%
0.2%
α-IgG
α-Flag
IB: α-RNA/DNA
IB: α-ssDNA
-
+
RNaseH

Figure 30. Cellular R-loops are immunoprecipitated with FLAG-tagged HIV-1 integrases by using FLAG-tag antibody.

(A) Protein extracts from FLAG-HIV-1-integrase-expressing cells were immunoprecipitated using anti-FLAG antibody. Western blot of FLAG immunoprecipitation was probed with anti-FLAG or anti-H3 antibodies. (B) DNA-RNA hybrid dot blot of FLAG antibody-immunoprecipitated nucleic acid extracts. Where indicated, nucleic acid extracts were untreated (–) or treated (+) with RNase H before probing with the indicated antibodies.

HIV-1 integrases are DNA and RNA binding proteins (Kessl et al., 2016; van Gent et al., 1991) , but its binding ability to such nucleic acid structure like R-loop has not been investigated. I carried in vitro protein-nucleic acid binding assay by electrophoretic mobility shift assay (EMSA) with Sso7d-tagged HIV-1 integrase recombinant proteins and diverse structures of nucleic acid substrates including R-loop and dsDNA. Interestingly, R-loop bound to HIV-1 integrase proteins with greater binding affinity than dsDNA, which is a known target nucleic acids form of HIV-1 integrases (Figure 31A and 31B). Additionally, R-loop composing forms of nucleic acid structures such as RNA-DNA hybrid with exposed ssDNA (R:D+ssDNA), RNA-DNA hybrid (hybrid) bound to HIV-1 integrase protein with similar binding affinity with R-loop (Figure 31A and 31B). This suggests that the HIV-1 integrates viral cDNAs into R-loop regions through direct physical binding of HIV-1 integrase proteins, which prefer to bind the genomic R-loops in the host.

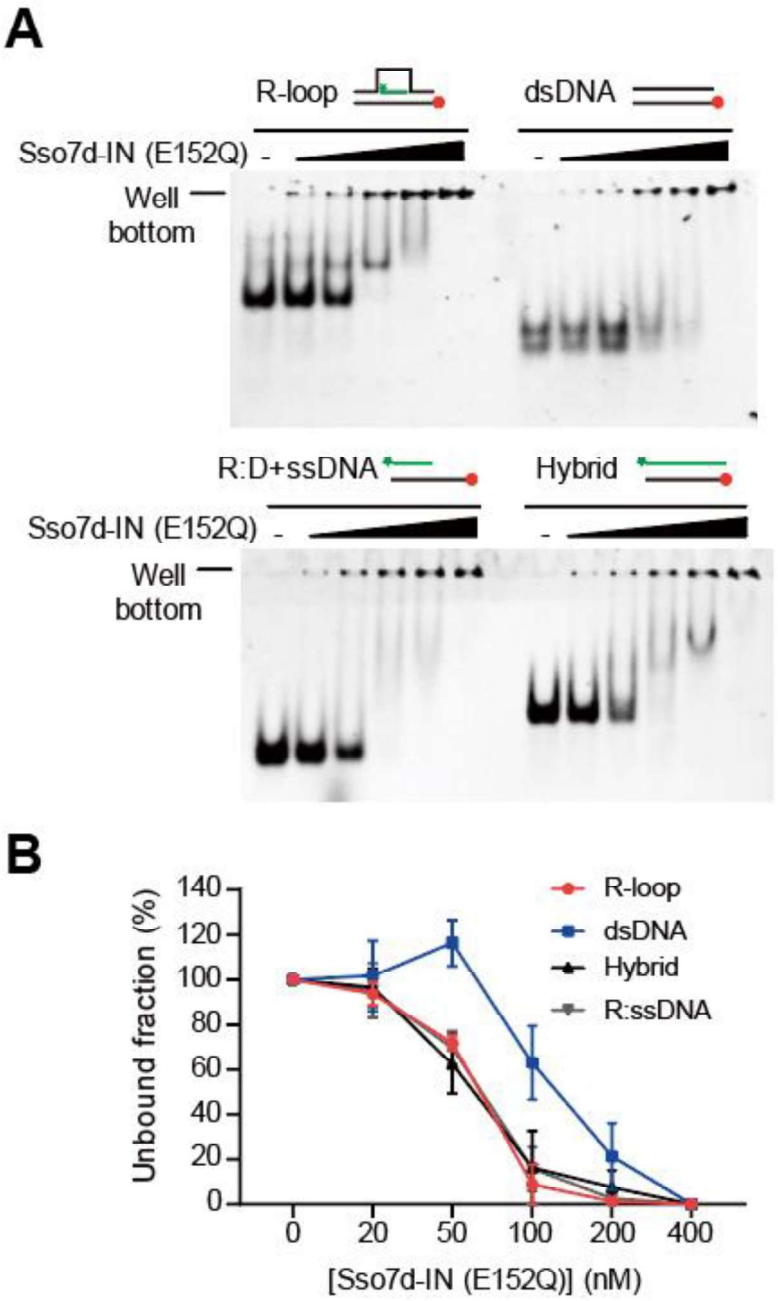
A
R-loop
dsDNA
Sso7d-IN (E152Q)
-
-
Well
bottom
R:D+ssDNA
Hybrid
Sso7d-IN (E152Q)
-
-
Well
bottom
B
Unbound fraction (%)
140
120
100
80
60
40
20
0
0
20
50
100
200
400
[Sso7d-IN (E152Q)] (nM)
R-loop
dsDNA
Hybrid
R:ssDNA

Figure 31. Electrophoretic mobility shift assay with Sso7d-tagged HIV-1 integrase recombinant proteins and nucleic acid substrates.

(A) Representative gel images for EMSA of Sso7d-tagged HIV-1-integrase (E152Q) with different types of nucleic acids substrates (R-loop, dsDNA, R:D+ssDNA and Hybrid). 100 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 (n = 3). (B) Unbound fraction were quantified for EMSA of Sso7d-tagged HIV-1-integrase (E152Q) with different types substrates (R-loop, dsDNA, R:D+ssDNA and Hybrid). Data are presented as the mean ± SEM, n = 3.

Subsequently, I attempted to observe the interaction between the R-loops and HIV-1 integrase using a proximity-ligation assay (PLA), in HIV-1-infected cells. I used two antibodies: one that binds to R-loops (anti-S9.6) and another one that binds to GFP-tagged HIV-1 integrase. I detected PLA signals in cells infected with HIV-IN-EGFP virions and in non-infected control cells. PLA signals in non-infected cells were comparable to those in S9.6-alone and GFP-alone single antibody-negative controls; however, PLA signals significantly increased upon HIV-1 infection (Figure 32A and 32B).

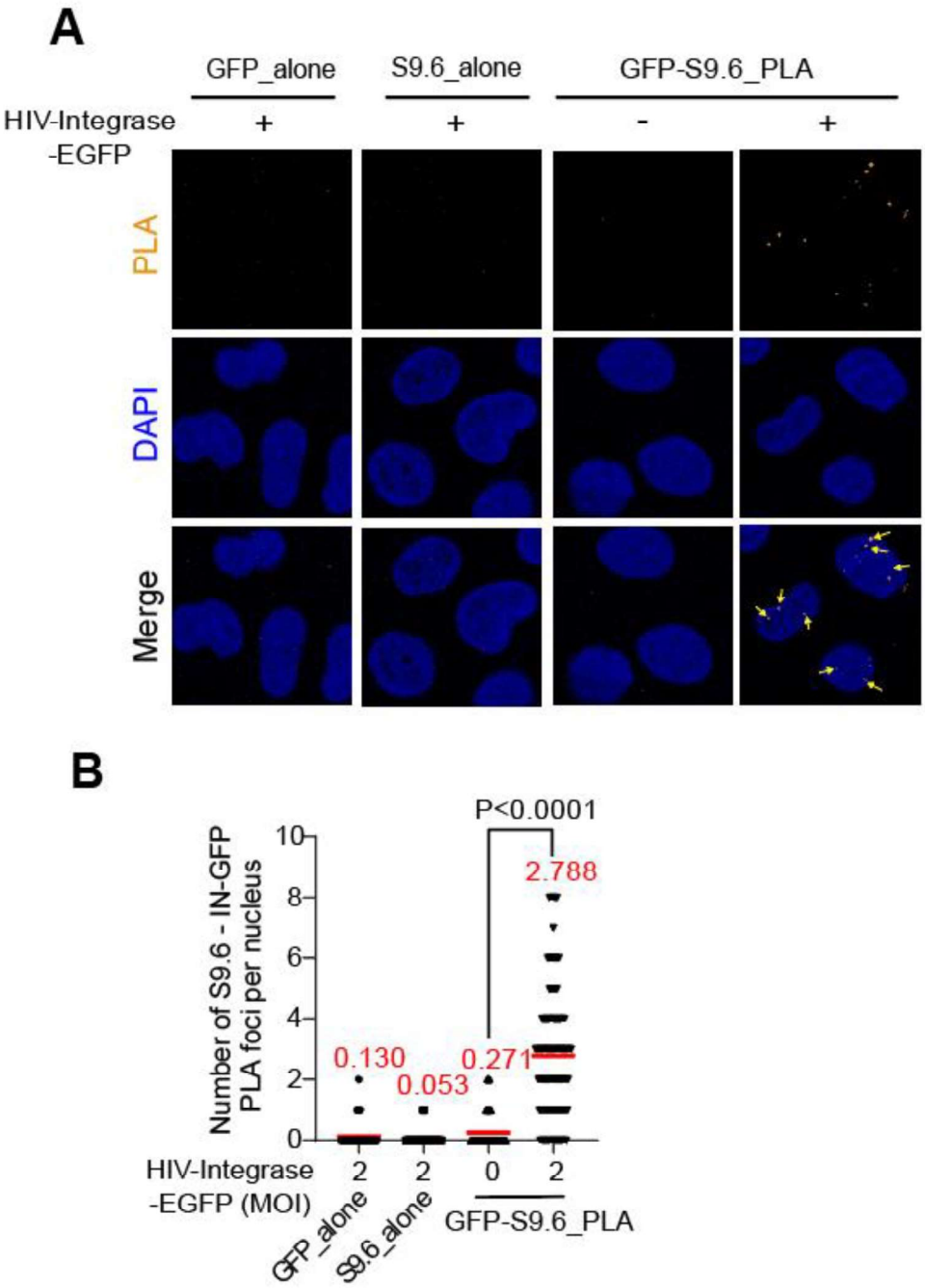
A
GFP_alone
S9.6_alone
GFP-S9.6_PLA
HIV-Integrase
-EGFP
+
+
-
+
PLA
DAPI
Merge
B
P<0.0001
2.788
0.130
0.053
0.271
Number of S9.6 - IN-GFP
PLA foci per nucleus
10
8
6
4
2
0
HIV-Integrase
-EGFP (MOI)
2
2
0
2
GFP_alone
S9.6_alone
GFP-S9.6_PLA

Figure 32. Proximity-ligation assay with anti-S9.6 and anti-EGFP in HeLa cells infected with HIV-IN-EGFP viruses.

(A) Representative images of the proximity-ligation assay (PLA) between GFP and S9,6 antibodies in HIV-IN-EGFP virion-infected HeLa cells at 6 hpi. Cells were subjected to PLA (orange) and co-stained with DAPI (blue). GFP_alone and S9.6_alone were used as single-antibody controls from HIV-IN-EGFP virion-infected HeLa cells. PLA puncta in the nucleus are indicated by the yellow arrows. (B) Quantification analysis of number of PLA foci per nucleus. The mean value for each data point is indicated by the red line. P value was calculated using a two-tailed unpaired t-test ($n > 50$).

Moreover, since I found HIV-1 integration prefers HIV-1-induced R-loops over constitutive R-loop, I attempted to quantify and compare the integrase binding at HIV-1-induced R-loops versus constitutive R-loops by Chromatin Immunoprecipitation (ChIP)-qPCR analysis. To quantify the 'R-loop specific' ChIP signals, I treated the immunoprecipitated materials with DNase I and reverse transcribed them, as previously described in DRIPc-seq library construction (Sanz and Chedin, 2019) (Figure 33). I infected HeLa cells expressing Flag-tagged HIV-1 integrase proteins with VSV-G-pseudotyped HIV-1-EGFP and harvested at 6 hpi, then carried immunoprecipitation with anti-Flag. When Flag-tagged HIV-1 integrase proteins were successfully immunoprecipitated (Figure 34A), the HIV-1 integrase bound R-loops were recovered and quantified using primers targeting 5 constitutive R-loop regions and 5 HIV-1-induced R-loop regions (Table 9). I found that HIV-1-induced R-loops, which drives HIV-1 integration, were more preferentially bound by HIV-1 integrase proteins than the constitutive R-loops (Figure 34B).

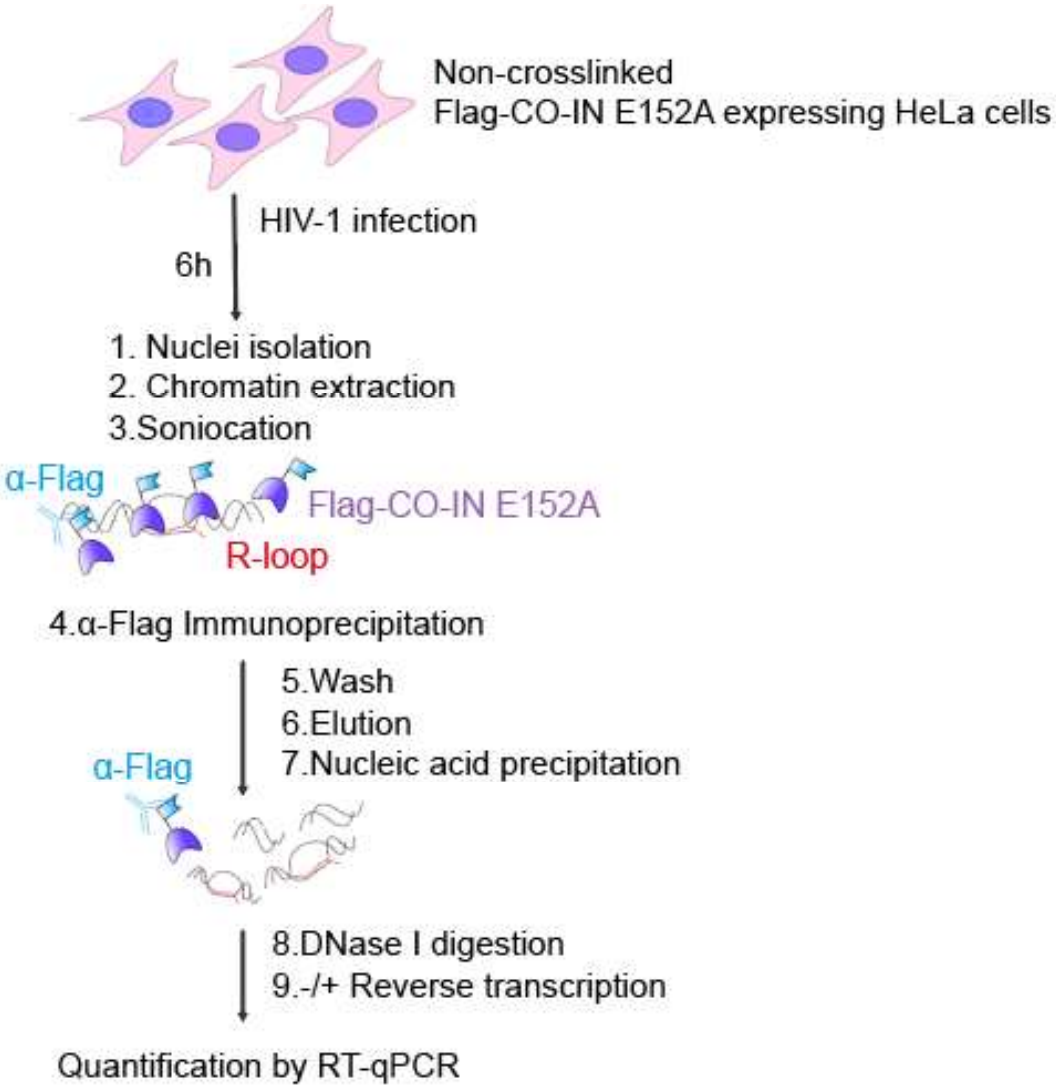
Non-crosslinked
Flag-CO-IN E152A expressing HeLa cells
HIV-1 infection
6h
1. Nuclei isolation
2. Chromatin extraction
3.Soniocation
α-Flag
Flag-CO-IN E152A
R-loop
4.α-Flag Immunoprecipitation
5.Wash
6.Elution
7.Nucleic acid precipitation
α-Flag
8.DNase I digestion
9.-/+ Reverse transcription
Quantification by RT-qPCR

Figure 33. Summary of the experimental design for anti-FLAG ChIP in FLAG-HIV-1-integrase (E152A)-expressing cells infected with HIV-1.

S9.6 antibody-immunoprecipitated nucleic acids were treated with DNase I. Half of DNase I treated immunoprecipitated nucleic acids were reverse transcribed and another half were left as reverse transcription negative control. The R-loop specific FLAG ChIP signal were quantified by using qPCR analysis with genomic R-loop region specific primer pairs.

Table 9. Chromosomal position and DRIPc−seq signal for constitutive and HIV−1−induced R−loop regions in HeLa cells.

Annotate	Chromosom	Position (hg38)	DRIPc−seq signal	Integration site−seq peak number
C1	chr3	183673730−183674488	0hpi=0.85, 3hpi=0.77, 6hpi=0.68, 12hpi=0.95	267
C2	chr7	32460869−32461386	0hpi=0.82, 3hpi=0.78, 6hpi=0.93, 12hpi=0.72	977
C3	chr13	57222544−57223482	0hpi=0.43, 3hpi=0.46, 6hpi=0.29, 12hpi=0.56	313
C4	chr17	17621354−17622095	0hpi=0.85, 3hpi=0.78, 6hpi=0.95, 12hpi=0.96	1457
C5	chr19	1281979−1282812	0hpi=0.69, 3hpi=0.63, 6hpi=0.67, 12hpi=0.65	1207
I1	chr1	113450189−113450774	0hpi=1.55, 3hpi=5.95, 6hpi=6.29, 12hpi=1.07	2158
I2	chr1	160358335−160358936	0hpi=0.08, 3hpi=3.25, 6hpi=0.58, 12hpi=0.05	1002
I3	chr1	74782704−74783339	0hpi=1.83, 3hpi=4.27, 6hpi=8.38, 12hpi=2.03	2637
I4	chr1	121234384−121234896	0hpi=0.09, 3hpi=0.64, 6hpi=2.93, 12hpi=0.10	1919
I5	chr1	144021725−144022703	0hpi=0.23, 3hpi=0.87, 6hpi=3.34, 12hpi=0.23	1849

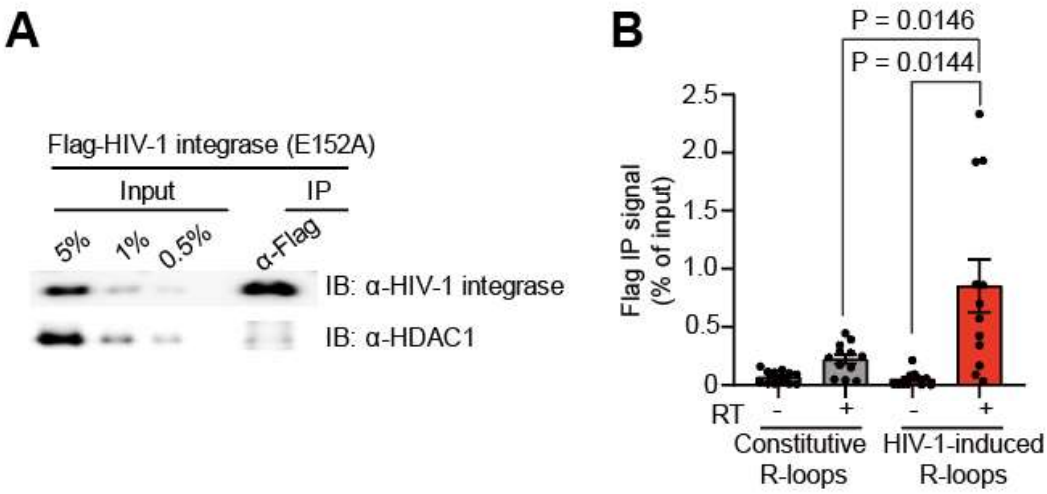
A
Flag-HIV-1 integrase (E152A)
Input
IP
5%
1%
0.5%
α-Flag
IB: α-HIV-1 integrase
IB: α-HDAC1
B
P = 0.0146
P = 0.0144
Flag IP signal
(% of input)
2.5
2.0
1.5
1.0
0.5
0
RT
-
+
-
+
Constitutive
R-loops
HIV-1-induced
R-loops

Figure 34. Anti-FLAG ChIP of constitutive or HIV-1-induced R-loops in FLAG-HIV-1-integrase (E152A)-expressing cells infected with HIV-1.

(A) Protein extracts from FLAG-HIV-1-integrase (E152A)-expressing cells were immunoprecipitated using anti-FLAG antibody. Western blot of FLAG immunoprecipitation was probed with α-FLAG or α-HDAC1 antibodies. (B) Anti-FLAG ChIP in FLAG-HIV-1-integrase (E152A)-expressing cells infected with HIV-1 and harvested at 6 hpi. Immunoprecipitated nucleic acid were subjected to qPCR analysis using 5 pairs of primers targeting constitutive R-loops or HIV-1-induced R-loops. Individual dot indicates an individual biological replicate for FLAG ChIP signal of each constitutive R-loop or HIV-1-induced R-loop regions.

4. DISCUSSION

In this study, I found that HIV-1 preferentially integrates into regions rich in R-loops, suggesting that R-loops are a pivotal host factor governing HIV-1 integration site selection. Using our R-loop-inducible cell models, R-loop formation, rather than transcription activity itself, was found to be the determinant for HIV-1 integration site selection. In addition, HIV-1 integrase proteins physically bind with R-loops in vitro, and they interact with host genomic R-loops in HIV-1-infected cells. These results demonstrated that HIV-1 exploits and selectively targets the host genomic R-loops for successful integration and infection (Figure 35).

One possible explanation for why HIV-1 integration shows a preference towards host genomic R-loops is that the R-loop structure may drive dynamics in the genomic environment and spatial organization of the genome, resulting in increased accessibility for HIV-1 intasome binding to the target host genomic region. R-loops display enhancer and insulator chromatin states, which can act as distal regulatory elements, by recruiting diverse chromatin binding factors (Sanz et al., 2016). This not only allows R-loops to drive dynamics in the genome, but also possibly drives R-loop-mediated integration over long-range genomic regions. R-loop regions are known to exhibit increased chromatin accessibility. In the cellular genome, these structures relieve superhelical stresses and are often associated with open chromatin marks and active enhancers (Chedin,

2016; Sanz et al., 2016), which are also distributed over HIV-1 integration sites (Schroder et al., 2002). R-loops may take a role as an intermediate regulator of HIV-1 integration sites selection by such host factors driving HIV-1 integration by closely interacting with such genomic compartments. For example, LEDGF/p75 and CPSF6 directing integration sites selection by interacting with integrase or trafficking viral preintegration complex (PIC) were recently identified as R-loop binding cellular factors (Cristini et al., 2018; Mosler et al., 2021). A guanine-quadruplex (G4) structure can be generated in the non-template DNA strand of the R-loop, which is another contributor to genome architecture (Lee et al., 2020). A recent study has shown that G4 DNA can influence both productive and latent HIV-1 integration, as well as the potential for reactivation of latent proviruses (Ajoge et al., 2022). Taking into account these previous studies alongside our current findings, I have demonstrated a novel role for host cellular R-loops in the selection of HIV-1 integration sites.

Our data show that HIV-1 integrase proteins physically interact with genomic R-loops in vitro and in cells. Recent advancements in cryogenic electron microscopy (cryo-EM) technology have enabled the disclosure of conformational characteristics of the target DNA during retroviral integration (Ballandras-Colas et al., 2022; Jozwik et al., 2022). During retroviral integration, the target DNA undergoes a transition of its conformation from B-form to A-form. R-loops, which represent intermediates between B-form DNA and A-form RNA conformation (Jozwik et al., 2022), may have intrinsic preferential binding ability to retroviral intasomes over other nucleic acid structures.

Cellular R-loops are recognized and regulated by numerous cellular proteins (Cristini et al., 2018; Mosler et al., 2021). In particular, *in cis* R-loops formed in gene bodies should be tightly regulated by cellular factors such as R-loop resolving factors, DNA damage response proteins, and even DNA replisome and RNA polymerase complexes, because they can cause distinct DNA stalling and damage(Garcia-Muse and Aguilera, 2019; Petermann et al., 2022). Our analysis indicated that constitutive R-loops owns a higher GC skew, a strong predictor of transcriptional R-loop formation, than the HIV-1-induced R-loops. Also the convergent nucleotide skew were only found for constitutive R-loops. Besides, since R-loop induction by HIV-1 does not follows tanscriptome changes upon HIV-1 infection, it is possible that HIV-1-induced R-loops are non-

canonical, and formed independently of transcription activation, perhaps *in-trans* manner. Considering these together with physical interaction between R-loops and integrase proteins and enrichment of integrase binding towards HIV-1-induced R-loops, it is plausible that HIV-1-induced R-loops, which are formed independently of canonical cellular R-loop forming mechanisms, would be less targeted or 'wrapped' by cellular R-loop binding cofactors and thus more likely to be exposed for HIV-1 integrase binding. However, further study of the differences between pathogen-induced and constitutive R-loops is required.

Viruses often take advantage of various host factors, and targeting viral components that manipulate the host cellular environment can be an effective strategy for antiviral therapy. Our study has shown that host genomic R-loops accumulate significantly shortly after HIV-1 infection. Notably, the genomic regions where HIV-1-induced R-loops accumulated did not necessarily co-localize with actively expressed genes. Thus, it is possible that virion-associated HIV-1 proteins are responsible for inducing these R-loops. For instance, the HIV-1 accessory protein Vpr causes genomic damage (Li et al., 2020a) and transcriptomic changes during the early stages post infection (Bauby et al., 2021), both of which can lead to *in cis* and *in trans* R-loop formation (Petermann et al., 2022). Another HIV-1 accessory protein, Vif, counteracts the host antiviral factor, APOBEC3 (Stopak et al., 2003), which regulates cellular R-

loop levels(McCann et al., 2021). Identifying the HIV-1 components responsible for inducing host cellular R-loops, and elucidating the mechanism by which they induce genome-wide R-loop formation and contribute to successful viral integration into selective genomic regions, represents an area for further research.

Although most HIV-1 integration occurs in genic regions (Einkauf et al., 2022), HIV-1 proviruses are also found in non-genic regions (Yukl et al., 2018) and understanding these "transcriptionally silent" proviruses is critical for developing strategies to completely eliminate HIV-1. In HIV-1 elite controllers, who suppress viral gene expression to undetectable levels, HIV-1 proviruses accumulate in heterochromatic regions (Jiang et al., 2020). Moreover, proviruses with lower expression level can persist in the host genome even during antiretroviral therapy (Einkauf et al., 2022). However, the mechanism by which HIV-1 targets gene-silent regions for "invisible" integration remains unclear. Our study has revealed that R-loops are enriched in both genic and non-genic regions during HIV-1 infection, and that the virus preferentially targets these R-loops for integration. I propose that HIV-1-induced R-loops, particularly those enriched in non-genic regions, may represent the mechanism by which the virus achieves "invisible" and permanent infection.

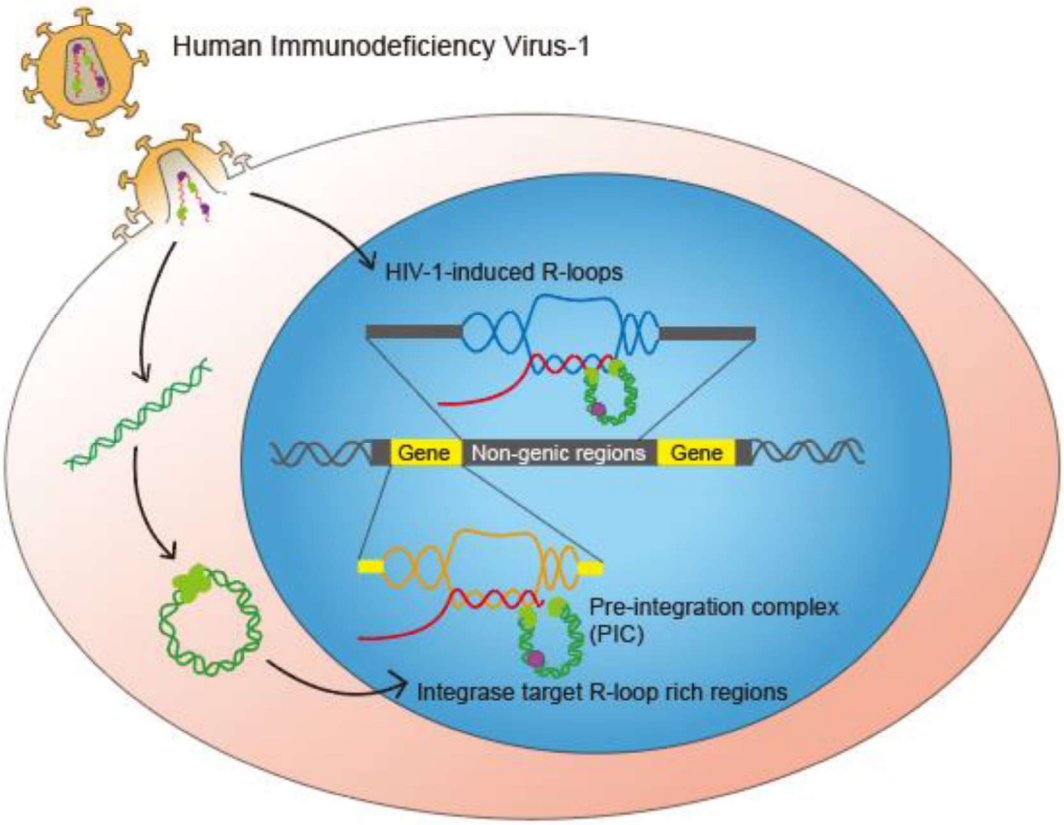

Figure 35. Model of the HIV-1 integration targeting host genomic R-loop induced upon infection.

HIV-1 infection induces genome-wide host genomic R-loops. HIV-1 PICs consisting of HIV-1-encoded integrase proteins binds the host genomic R-loops and HIV-1 favorably targets R-loop rich regions for its viral genome integration.

5. REFERENCES

Achuthan, V., Perreira, J.M., Sowd, G.A., Puray-Chavez, M., McDougall, W.M., Paulucci-Holthauzen, A., Wu, X., Fadel, H.J., Poeschla, E.M., Multani, A.S., *et al.* (2018). Capsid-CPSF6 Interaction Licenses Nuclear HIV-1 Trafficking to Sites of Viral DNA Integration. Cell Host Microbe *24*, 392-404 e398.

Ajoge, H.O., Kohio, H.P., Paparisto, E., Coleman, M.D., Wong, K., Tom, S.K., Bain, K.L., Berry, C.C., Arts, E.J., and Barr, S.D. (2022). G-Quadruplex DNA and Other Non-Canonical B-Form DNA Motifs Influence Productive and Latent HIV-1 Integration and Reactivation Potential. Viruses *14*.

Albanese, A., Arosio, D., Terreni, M., and Cereseto, A. (2008). HIV-1 pre-integration complexes selectively target decondensed chromatin in the nuclear periphery. PLoS One *3*, e2413.

Amemiya, H.M., Kundaje, A., and Boyle, A.P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci Rep *9*, 9354.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

Arora, R., Lee, Y., Wischnewski, H., Brun, C.M., Schwarz, T., and Azzalin, C.M. (2014). RNaseH1 regulates TERRA-telomeric DNA hybrids and telomere maintenance in ALT tumour cells. Nat Commun *5*, 5220.

Ballandras-Colas, A., Chivukula, V., Gruszka, D.T., Shan, Z., Singh, P.K., Pye, V.E., McLean, R.K., Bedwell, G.J., Li, W., Nans, A., *et al.*

(2022). Multivalent interactions essential for lentiviral integrase function. Nat Commun *13*, 2416.

Bauby, H., Ward, C.C., Hugh-White, R., Swanson, C.M., Schulz, R., Goujon, C., and Malim, M.H. (2021). HIV-1 Vpr Induces Widespread Transcriptomic Changes in CD4(+) T Cells Early Postinfection. mBio *12*, e0136921.

Brussel, A., and Sonigo, P. (2003). Analysis of early human immunodeficiency virus type 1 DNA synthesis by use of a new sensitive assay for quantifying integrated provirus. J Virol *77*, 10119-10124.

Chedin, F. (2016). Nascent Connections: R-Loops and Chromatin Patterning. Trends Genet *32*, 828-838.

Chedin, F., and Benham, C.J. (2020). Emerging roles for R-loop structures in the management of topological stress. J Biol Chem *295*, 4684-4695.

Chen, H.C., Martinez, J.P., Zorita, E., Meyerhans, A., and Filion, G.J. (2017). Position effects influence HIV latency reversal. Nat Struct Mol Biol *24*, 47-54.

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

Cristini, A., Groh, M., Kristiansen, M.S., and Gromak, N. (2018). RNA/DNA Hybrid Interactome Identifies DXH9 as a Molecular Player in Transcriptional Termination and R-Loop-Associated DNA

Damage. Cell Rep *23*, 1891-1905.

Crossley, M.P., Song, C., Bocek, M.J., Choi, J.H., Kousorous, J., Sathirachinda, A., Lin, C., Brickner, J.R., Bai, G., Lans, H., *et al.* (2023). R-loop-derived cytoplasmic RNA-DNA hybrids activate an immune response. Nature *613*, 187-194.

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

Einkauf, K.B., Osborn, M.R., Gao, C., Sun, W., Sun, X., Lian, X., Parsons, E.M., Gladkov, G.T., Seiger, K.W., Blackmer, J.E., *et al.* (2022). Parallel analysis of transcription, integration, and sequence of single HIV-1 proviruses. Cell *185*, 266-282 e215.

Felix Krueger, F.J., Phil Ewels, Ebrahim Afyounian, & Benjamin Schuster-Boeckler (2021). FelixKrueger/TrimGalore: v0.6.7 - DOI via Zenodo (0.6.7). Zenodo.

Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J.E., Mudge, J.M., Sisu, C., Wright, J.C., Armstrong, J., Barnes, I., *et al.* (2021). Gencode 2021. Nucleic Acids Res *49*, D916-D923.

Fu, S., Phan, A.T., Mao, D., Wang, X., Gao, G., Goff, S.P., and Zhu, Y. (2022). HIV-1 exploits the Fanconi anemia pathway for viral DNA integration. Cell Rep *39*, 110840.

Garcia-Muse, T., and Aguilera, A. (2019). R Loops: From Physiological to Pathological Roles. Cell *179*, 604-618.

Garcia-Rubio, M.L., Perez-Calero, C., Barroso, S.I., Tumini, E., Herrera-Moyano, E., Rosado, I.V., and Aguilera, A. (2015). The

Fanconi Anemia Pathway Protects Genome Integrity from R-loops. PLoS Genet *11*, e1005674.

Giannini, M., Bayona-Feliu, A., Sproviero, D., Barroso, S.I., Cereda, C., and Aguilera, A. (2020). TDP-43 mutations link Amyotrophic Lateral Sclerosis with R-loop homeostasis and R loop-mediated DNA damage. PLoS Genet *16*, e1009260.

Ginno, P.A., Lim, Y.W., Lott, P.L., Korf, I., and Chedin, F. (2013). GC skew at the 5' and 3' ends of human genes links R-loop formation to epigenetic regulation and transcription termination. Genome Res *23*, 1590-1600.

Ginno, P.A., Lott, P.L., Christensen, H.C., Korf, I., and Chedin, F. (2012). R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. Mol Cell *45*, 814-825.

Hamperl, S., Bocek, M.J., Saldivar, J.C., Swigut, T., and Cimprich, K.A. (2017). Transcription-Replication Conflict Orientation Modulates R-Loop Levels and Activates Distinct DNA Damage Responses. Cell *170*, 774-786 e719.

Jiang, C., Lian, X., Gao, C., Sun, X., Einkauf, K.B., Chevalier, J.M., Chen, S.M.Y., Hua, S., Rhee, B., Chang, K., *et al.* (2020). Distinct viral reservoirs in individuals with spontaneous control of HIV-1. Nature *585*, 261-267.

Jin, Y., Tam, O.H., Paniagua, E., and Hammell, M. (2015). TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. Bioinformatics *31*, 3593-3599.

Johnson, W.E. (2019). Origins and evolutionary consequences of ancient endogenous retroviruses. Nat Rev Microbiol *17*, 355–370.

Jones, R.B., Song, H., Xu, Y., Garrison, K.E., Buzdin, A.A., Anwar, N., Hunter, D.V., Mujib, S., Mihajlovic, V., Martin, E., *et al.* (2013). LINE-1 retrotransposable element DNA accumulates in HIV-1-infected cells. J Virol *87*, 13307–13320.

Jozwik, I.K., Li, W., Zhang, D.W., Wong, D., Grawenhoff, J., Ballandras-Colas, A., Aiyer, S., Cherepanov, P., Engelman, A.N., and Lyumkis, D. (2022). B-to-A transition in target DNA during retroviral integration. Nucleic Acids Res *50*, 8898–8918.

Kessl, J.J., Kutluay, S.B., Townsend, D., Rebensburg, S., Slaughter, A., Larue, R.C., Shkriabai, N., Bakouche, N., Fuchs, J.R., Bieniasz, P.D., *et al.* (2016). HIV-1 Integrase Binds the Viral RNA Genome and Is Essential during Virion Morphogenesis. Cell *166*, 1257–1268 e1212.

Lee, C.Y., McNerney, C., Ma, K., Zhao, W., Wang, A., and Myong, S. (2020). R-loop induced G-quadruplex in non-template promotes transcription by successive R-loop formation. Nat Commun *11*, 3392.

Li, D., Lopez, A., Sandoval, C., Nichols Doyle, R., and Fregoso, O.I. (2020a). HIV Vpr Modulates the Host DNA Damage Response at Two Independent Steps to Damage DNA and Repress Double-Strand DNA Break Repair. mBio *11*.

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

Li, W., Singh, P.K., Sowd, G.A., Bedwell, G.J., Jang, S., Achuthan, V., Oleru, A.V., Wong, D., Fadel, H.J., Lee, K., *et al.* (2020b). CPSF6-

Dependent Targeting of Speckle-Associated Domains Distinguishes Primate from Nonprimate Lentiviral Integration. mBio *11*.

Lim, Y.W., Sanz, L.A., Xu, X., Hartono, S.R., and Chedin, F. (2015). Genome-wide DNA hypomethylation and RNA:DNA hybrid accumulation in Aicardi-Goutieres syndrome. Elife *4*.

Lucic, B., Chen, H.C., Kuzman, M., Zorita, E., Wegner, J., Minneker, V., Wang, W., Fronza, R., Laufs, S., Schmidt, M., *et al.* (2019). Spatially clustered loci with multiple enhancers are frequent targets of HIV-1 integration. Nat Commun *10*, 4059.

Lusic, M., and Siliciano, R.F. (2017). Nuclear landscape of HIV-1 infection and integration. Nat Rev Microbiol *15*, 69-82.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011 *17*, 3 %J EMBnet.journal.

McAllister, R.G., Liu, J., Woods, M.W., Tom, S.K., Rupar, C.A., and Barr, S.D. (2014). Lentivector integration sites in ependymal cells from a model of metachromatic leukodystrophy: non-B DNA as a new factor influencing integration. Mol Ther Nucleic Acids *3*, e187.

McCann, J.L., Cristini, A., Law, E.K., Lee, S.Y., Tellier, M., Carpenter, M.A., Beghè, C., Kim, J.J., Jarvis, M.C., Stefanovska, B., *et al.* (2021). R-loop homeostasis and cancer mutagenesis promoted by the DNA cytosine deaminase APOBEC3B. 2021.2008.2030.458235.

Mosler, T., Conte, F., Longo, G.M.C., Mikicic, I., Kreim, N., Mockel, M.M., Petrosino, G., Flach, J., Barau, J., Luke, B., *et al.* (2021). R-loop proximity proteomics identifies a role of DDX41 in transcription-associated genomic instability. Nat Commun *12*, 7314.

Nguyen, H.D., Yadav, T., Giri, S., Saez, B., Graubert, T.A., and Zou, L. (2017). Functions of Replication Protein A as a Sensor of R Loops and a Regulator of RNaseH1. Mol Cell *65*, 832-847 e834.

Niehrs, C., and Luke, B. (2020). Regulatory R-loops as facilitators of gene expression and genome stability. Nat Rev Mol Cell Biol *21*, 167-178.

Passos, D.O., Li, M., Yang, R., Rebensburg, S.V., Ghirlando, R., Jeon, Y., Shkriabai, N., Kvaratskhelia, M., Craigie, R., and Lyumkis, D. (2017). Cryo-EM structures and atomic model of the HIV-1 strand transfer complex intasome. Science *355*, 89-92.

Petermann, E., Lan, L., and Zou, L. (2022). Sources, resolution and physiological relevance of R-loops and RNA-DNA hybrids. Nat Rev Mol Cell Biol *23*, 521-540.

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

Sanz, L.A., and Chedin, F. (2019). High-resolution, strand-specific R-loop mapping via S9.6-based DNA-RNA immunoprecipitation and high-throughput sequencing. Nat Protoc *14*, 1734-1755.

Sanz, L.A., Hartono, S.R., Lim, Y.W., Steyaert, S., Rajpurkar, A., Ginno, P.A., Xu, X., and Chedin, F. (2016). Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals. Mol Cell *63*, 167-178.

Schrijvers, R., De Rijck, J., Demeulemeester, J., Adachi, N., Vets, S., Ronen, K., Christ, F., Bushman, F.D., Debyser, Z., and Gijsbers, R. (2012). LEDGF/p75-independent HIV-1 replication demonstrates a

role for HRP-2 and remains sensitive to inhibition by LEDGINs. PLoS Pathog *8*, e1002558.

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

Sowd, G.A., Serrao, E., Wang, H., Wang, W., Fadel, H.J., Poeschla, E.M., and Engelman, A.N. (2016). A critical role for alternative polyadenylation factor CPSF6 in targeting HIV-1 integration to transcriptionally active chromatin. Proc Natl Acad Sci U S A *113*, E1054-1063.

Srinivasachar Badarinarayan, S., Shcherbakova, I., Langer, S., Koepke, L., Preising, A., Hotter, D., Kirchhoff, F., Sparrer, K.M.J., Schotta, G., and Sauter, D. (2020). HIV-1 infection activates endogenous retroviral promoters regulating antiviral gene expression. Nucleic Acids Res *48*, 10890-10908.

Stopak, K., de Noronha, C., Yonemoto, W., and Greene, W.C. (2003). HIV-1 Vif blocks the antiviral activity of APOBEC3G by impairing both its translation and intracellular stability. Mol Cell *12*, 591-601.

van Gent, D.C., Elgersma, Y., Bolk, M.W., Vink, C., and Plasterk, R.H. (1991). DNA binding properties of the integrase proteins of human immunodeficiency viruses types 1 and 2. Nucleic Acids Res *19*, 3821-3827.

Yukl, S.A., Kaiser, P., Kim, P., Telwatte, S., Joshi, S.K., Vu, M., Lampiris, H., and Wong, J.K. (2018). HIV latency in isolated patient CD4(+) T cells may be due to blocks in HIV transcriptional

elongation, completion, and splicing. Sci Transl Med *10*.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., *et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol *9*, R137.

6. ABSTRACT IN KOREAN

인간면역결핍 바이러스 (HIV-1)는 RNA 게놈을 DNA로 역전사하고 숙주 게놈에 통합함으로써 숙주세포를 감염시킨다. 그러나 HIV-1 게놈이 숙주 게놈에 통합되는 위치를 결정하는 요인은 명확히 밝혀지지 않았다. HIV-1 통합 위치는 인간 게놈에서 유전자가 활발히 전사되는 부위를 선호하는 것으로 간주되었지만, 최근 개별 HIV-1 통합 위치에 대한 고해상도 분석 결과에 따르면 HIV-1 바이러스 게놈은 비유전자 영역을 포함한 다양한 숙주 게놈 영역에 통합될 수 있음이 밝혀졌다. 본 연구에서는 HIV-1이 숙주 게놈의 DNA-RNA 혼합체로 구성된 게놈 구조인 R-루프에 우선적으로 바이러스 게놈을 통합시킨다는 것을 밝혔다. HIV-1 바이러스 감염은 숙주 게놈의 유전자 및 비유전자 영역에서 모두 R-루프 형성을 유도하고, HIV-1 감염에 의해 유도된 R-루프 영역에 바이러스 게놈을 통합시킨다는 것을 발견했다. 특히, 유전자 전사 활동과 R-루프 형성을 독립적으로 제어할 수 있는 새로운 세포 모델을 사용하여, 전사 활동에 관계없이 R-루프의 존재가 HIV-1 통합 위치를 결정하는 주요 요인임을 입증했다. 또한 HIV-1 통합 효소 단백질이 R-루프에 물리적으로 결합한다는 것을 밝혀냄으로써 새로운 레트로바이러스 게놈 통합 기작을 제시하는 것뿐만 아니라 항레트로바이러스 치료제 개발을 위한 통찰력을 제공한다.

핵심어: 인간면역결핍 바이러스, 레트로바이러스, R-루프, DNA-RNA 혼합체, 바이러스 게놈 통합

학번: 2017-26763