

Mapping HIV-1 RNA Structure, Homodimers, Long-Range Interactions and persistent domains by HiCapR

Reviewed Preprint

v2 • February 7, 2025

Revised by authors

Reviewed Preprint

v1 • November 21, 2024

Yan Zhang, Jingwan Han, Dejian Xie, Wenlong Shen, Ping Li, Jian You Lau, Jingyun Li , Lin Li , Grzegorz Kudla , Zhihu Zhao 

Laboratory of Advanced Biotechnology, Beijing Institute of Biotechnology, Beijing, China • Institute of Microbiology and Epidemiology, Beijing, China • MRC Human Genetics Unit, University of Edinburgh, Edinburgh, United Kingdom

 https://en.wikipedia.org/wiki/Open_access

 Copyright information

eLife Assessment

This manuscript focuses on the identification of RNA crosslinks within the HIV RNA genome under different conditions i.e. in infected cells and in virions using a new method called HiCapR. These cross-links reveal long-range interactions that can be used to determine the structural arrangement of the viral RNA, providing **valuable** data that show differences in the genomic organization in different conditions. The data analysis, however, is **incomplete** and based on extensive computational analysis from a limited number of datasets.

<https://doi.org/10.7554/eLife.102550.2.sa3>

Abstract

Human Immunodeficiency Virus (HIV) persists as a leading global health issue. A significant gap in knowledge exists in our understanding of long-range interactions within the HIV-1 RNA genome. To bridge this gap, we introduce HiCapR, incorporating a psoralen crosslinking RNA proximity ligation and post-library hybridization for capturing HIV RNA-RNA interactions.

Leveraging HiCapR, we confirm the presence of stem structures in the key regions, such as the 5'-untranslated region(5'-UTR) and Rev Response Element (RRE) stems, and dimer sites in 5'-UTR region, which is responsible for HIV packaging. Importantly, we reveal multiple previously unknown homodimers along the HIV genome, which may have important implications for viral RNA splicing and packaging processes. Also, we uncover a wealth of unprecedented long-range interactions, particularly within the 5'-UTR of infected cells.

Intriguingly, our findings indicate a pronounced reduction in long-range RNA-RNA interactions, signifying a transition from a state of abundant interactions, hence a relative loose state within infected cells to a condensed structure within virions. Concurrently, we have demonstrated the presence of stable genomic domains within virions that are

instrumental in the dimerization process. These domains are preserved throughout the packaging process.

Our findings shed light on the functional significance of RNA organization, including stable and persistent genomic domains, homodimerization, and long-range RNA-RNA interactions, in the splicing, packaging as well as assembly of HIV.

Highlights

HiCapR is a new proximity ligation method for mapping RNA structures and homodimers in the HIV genome with sufficient reliability and efficiency.

Multiple homodimers were discovered along the genome, with potential implications for splicing and packaging processes.

Long-range RNA-RNA interactions are abundant in infected cells but significantly reduced in virions.

Stable genomic domains including homodimer sites are persistent in virions and are involved in dimerization.

Introduction

Human Immunodeficiency Virus (HIV) remains a significant global health concern despite advancements in antiretroviral therapy even with the newest report of total seven AIDS patients been completely “cured”^{1,2,3}. Research on HIV-1 RNA structure and function has intensified, focusing on assembly, release, and maturation processes, using NMR⁴, cryo-EM⁵ and chemical probing⁶⁻⁸. However, our knowledge of the global architecture and the long-range interactions within the HIV-1 RNA genome remains limited, primarily due to the complexity and dynamic nature of these structures.

The HIV genome, comprising two copies of positive-sense RNA, has been extensively studied to understand its regulatory elements, such as the 5'-UTR, crucial for dimerization and selective packaging into viral particles⁹. Additionally, interactions involving the nucleocapsid protein (NC) and the dimerization signal (Ψ) play pivotal roles in the assembly and maturation of HIV virions¹⁰. Furthermore, the RRE in the HIV genome, responsible for nucleocytoplasmic transport of viral RNA, exhibits structural variability impacting virus replication rates^{11,12}. While current techniques like SHAPE began to provide insights into local RNA structures^{7,8,13}, studying the intricate long-range RNA interactions and homodimers within the HIV genome under physiological conditions remains a significant challenge, primarily due to the limitations of these methods in capturing the full spectrum of RNA interactions and the dynamic nature of the viral genome.

Proximity ligation-based methods have been shown to be particularly useful for studying long-range RNA interactions, as they can overcome the RNA length limitations of traditional techniques. These methods have been applied to the study of various viruses, including influenza^{14,15}, Zika¹⁶⁻¹⁸, and SARS-CoV-2¹⁹⁻²², and have led to the identification of numerous RNA structures that are closely related to the virus lifecycle. We and others have recently shown that RNA proximity ligation data also contains information about RNA homodimerization events^{23,24}. Nevertheless, as of now, there are no proximity ligation studies of long-range interactions or dimeric interactions of the HIV RNA genome. Probably due to high requirements of starting material, making it very difficult for low abundant HIV sample treatment.

In this study, we modified our previous protocol²² by integrating a hybridization-based capture of HIV-1 sequence after library construction, resulting in a method we call High throughput Capture of RNA interactions (HiCapR). This advancement enabled the comprehensive capture and analysis of the complete HIV RNA genome. By applying HiCapR to both infected cells and virions, we uncover a distinct genomic compression pattern in virions, highlighting the critical role of global genome folding in HIV packaging and assembly. Our findings not only reveal the presence of persistent genomic domains within virions that facilitate whole genome dimerization, even the transgenerational inheritance but also a significant reduction of long-range RNA-RNA interactions in virions compared to infected cells. These insights provide a new perspective on the functional implications of HIV RNA structure in splicing and packaging processes and may lead to the identification of new targets for antiviral intervention.

Results

Capturing in vivo HIV genome RNA structure in infected cells and in virion

To effectively capture and analyze full-length low abundant HIV RNA-RNA interactions in infected cells, we developed an improved version of the protocol, which we call High throughput Capture of RNA interactions (HiCapR). This method incorporates hybridization-based capture with next-generation sequencing (NGS) library construction, a principle well established in capture Hi-C methods^{25–28} and in RNA proximity ligation^{17,19}. Applying this strategy, we compared the NL4-3 and GX2005002 strains²⁹, which are two distinct HIV-1 strains that exhibit different prevalence patterns in various geographical regions. The NL4-3 strain is a well-characterized laboratory strain that is widely used in HIV research and is representative of the HIV-1 subtype B, which is highly prevalent in Europe and the Americas. On the other hand, GX2005002 is a primary isolate of the CRF01_AE subtype, which is one of the most prevalent strains in Southeast Asia, particularly in China. The overall design of this study is presented in **Figure 1A**, with detailed methods provided in the methods section. Supplementary Figure S1 illustrates the bioinformatic pipeline used to detect RNA-RNA interactions and homodimers. Approximately 16 million raw reads were obtained per sequencing library, with over 95% of the reads containing HIV sequences, indicating efficient capture of HIV RNA transcripts. The alignment rate exceeded 90%, and we achieved an average depth of coverage of approximately 11,500X across the genome (Supplementary Table 1). The uniform distribution of reads post-capture (Supplementary Figures S2) demonstrates HiCapR's effectiveness in capturing the entire HIV RNA genome with minimal bias. The overall duplication rate remained below 50%, with a lower rate in the ligation group, suggesting increased library complexity due to RNA proximity ligation.

Chimeric reads were identified using the hyb technique as previously described²², with a database comprising human spliced mRNAs, noncoding RNAs³⁰ and the HIV genome sequence as reference. The chimeric rate, a crucial indicator of ligation efficiency, was approximately 1% in non-ligation controls and around 9% in the ligation group, aligning with the widely reported range. The proportion of 5'-3' and 3'-5' chimeras was nearly equal in long-range interactions, while proximal interactions were enriched for 3'-5' chimeras. Pearson correlation analysis revealed a high degree of correlation ($r > 0.99$) between biological replicates, indicating high reproducibility. Additionally, a significant similarity was observed between virions in the supernatant and infected cells from the same viral strain (Supplementary Figure S3).

Using a similar approach as described in our previous paper²², we generated a contact matrix for the global HIV-1 genome, revealing clear local and long-range interactions (**Figure 1B** and Supplementary Figure S4). In summary, HiCapR demonstrated high reproducibility, low bias, reliability, and efficiency in capturing HIV transcripts, making it a robust tool for analyzing the RNA structure of the HIV genome and potentially other similar RNA viruses.

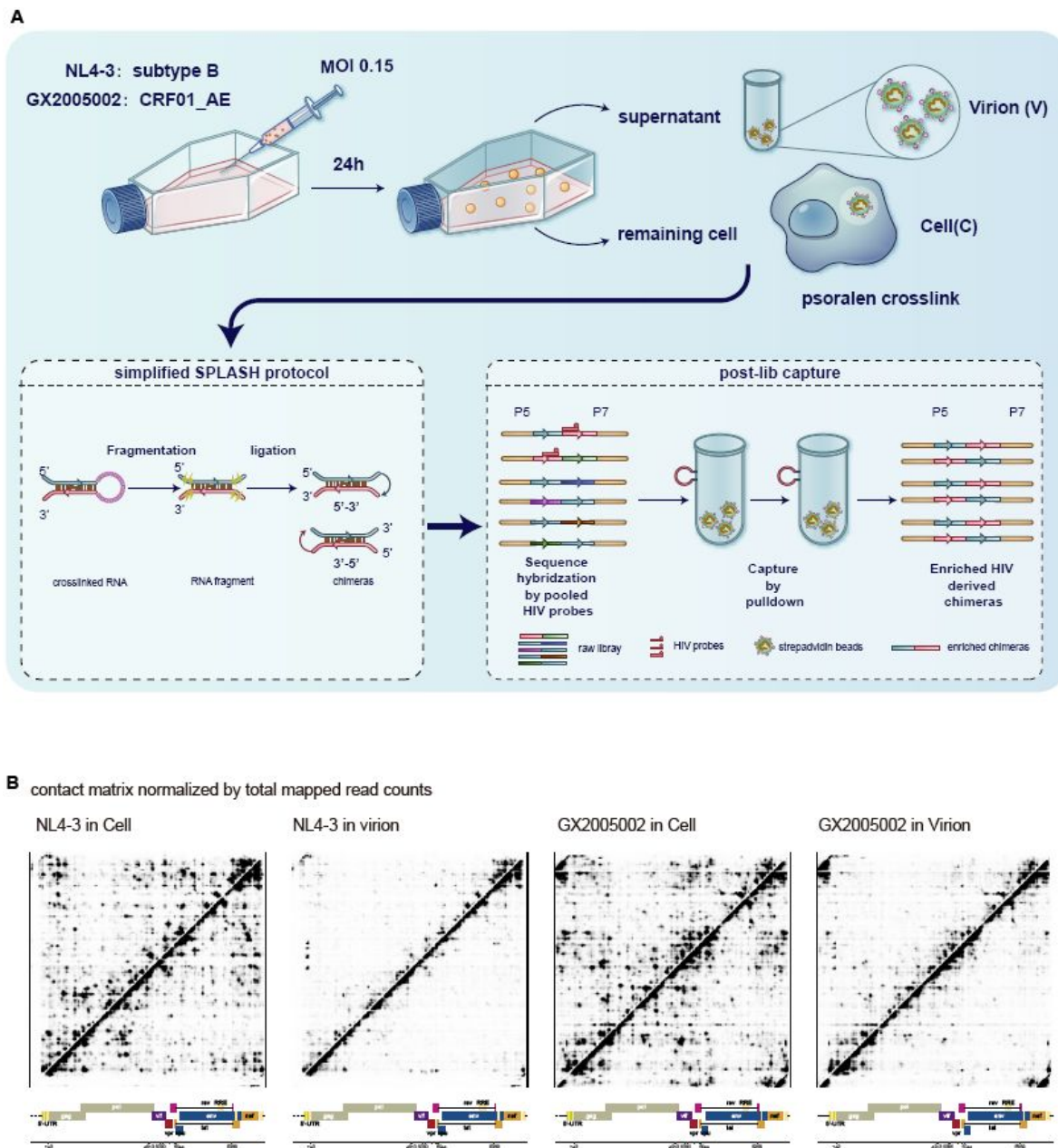


Figure 1

Experimental design of HiCapR and overall HIV-1 contact matrix

A) Experimental design of HiCapR for profiling RNA-RNA interactions and dynamics of the HIV genome. Based on the simplified SPLASH protocol, HiCapR incorporates a post-library probe-based hybridization and streptavidin pulldown method to enrich HIV RNA chimeras from the SPLASH library. The HiCapR method has been applied to two strains of HIV, NL4-3 and GX2005002, in both infected cells and their corresponding virus particles. B) Comprehensive contact matrix derived from infected cells and virions of both NL4-3 and GX2005002 HIV-1 strains in the HiCapR experimental groups. The heatmap displays the average count of chimeras per 1 million mapped reads, combining data from two biological replicates for each sample.

Local RNA structure heterogeneity, dynamics and robustness in HIV 5'-UTR

We aimed to characterize specific local structures and their dynamics within the HIV genome. One of the extensively investigated structures is 5'-UTR, as these structures are closely associated with crucial life processes of HIV.

Chimeras analysis support the presence of key structural elements, including TAR, polyA, SL1, SL2, and SL3, as well as polyA-SL1 in the monomeric conformation in the HIV genome (**Figure 2A** [↗](#) and Supplementary Figure S5). Despite the pivotal role of the 5'-UTR in replication, packaging, transcription and translation, we noted that the 5'-UTR sequence of HIV is not highly conserved across different HIV strains. A comparison between the NL4-3 and GX2005002 strains revealed notable insertions and deletions in the U5 region, along with point mutations in the loop of the SL1 core region, which is critical for HIV dimerization (**Figure 2B** [↗](#)). These findings raise questions about the conservation of the structural integrity of this region. We therefore applied the comradsFold algorithm ¹⁷[↗](#) to our HiCapR data to generate 1000 structure predictions of the 5'-UTR (extending 100 nt downstream of the AUG start codon) for each sample, and then performed MDS analysis after aligning the viral genome coordinates of both strains (**Figure 2C** [↗](#)). This analysis revealed that the reported dimer and monomer structures clustered together in NL4-3 but form distinct clusters in GX2005002 (**Figure 2C** [↗](#)). Based on these folded structures, we calculated the base pairing probability for each base pair, which involves determining the number of folded structures supporting a specific base pair divided by the total number of structures. Visualizing this base pairing probability as a heatmap identifies the most stable base pairs in the 5' UTR of HIV. We observed a consistent presence of key structural elements such as polyA, TAR, SL1, SL2, and SL3 in both NL4-3 and GX2005002 strains (**Figure 2D** [↗](#)), suggesting robustness in the overall structure despite sequence variations and alternative RNA conformations.

Novel homodimerization sites and their implications in RNA splicing and viral assembly

Proximity ligation techniques enable unambiguous detection of RNA homodimers from “overlapping” chimeric reads^{23,24} [↗](#) (**Figure 3A** [↗](#)). To ensure the accuracy of our results, we implemented a rigorous data filtering process to select chimeras formed exclusively through dimerization, minimizing background interference caused by RNA interactions or self-ligation (Supplementary Figure S1 and Methods). We quantified the total number of reads for dimeric chimeras, 3'-5' chimeras, and 5'-3' chimeras respectively in each sample (Supplementary Figure S6) and plotted contact matrices made from non-overlapping chimeras and dimeric chimeras (Supplementary Figure S7). The non-overlapping chimeras reveal short-range intramolecular RNA-RNA interactions, while dimeric chimeras capture homodimer formation in the HIV genome.

The strongest homodimerization signal was found in the 5'-UTR region, which aligns with previous studies. The 5'-UTR region is well-known for its role in triggering HIV dimerization. Previous literature has highlighted the importance of SL1 (DIS) and SL3 (Ψ) in 5'-UTR^{8,31,32} [↗](#). Our data shows that the SL1, SL2, and SL3 regions all have supportive dimeric chimeras (Supplementary Figure S7). We derived a Dimerization Score by calculating the reads that support base pairing of homodimeric chimeras (Supplementary Figure S8), analogous to the COMRADES Score described previously^{17,19,22} [↗](#). The base pairing of homodimers in the SL1 region of the HIV genome is consistent with previous data from NMR and cryo-EM studies^{33,34} [↗](#). Additionally, we observed that variations in nucleotides within the SL1 region do not alter the base pairing pattern of dimers between the NL4-3 and GX2005002 strains (**Figure 3B** [↗](#)) which is egret with previous structure robustness revealed by permutation and structure prediction results. Strikingly, in addition to the known homodimer within the SL1 and SL3 regions, we observed homodimerization distributed

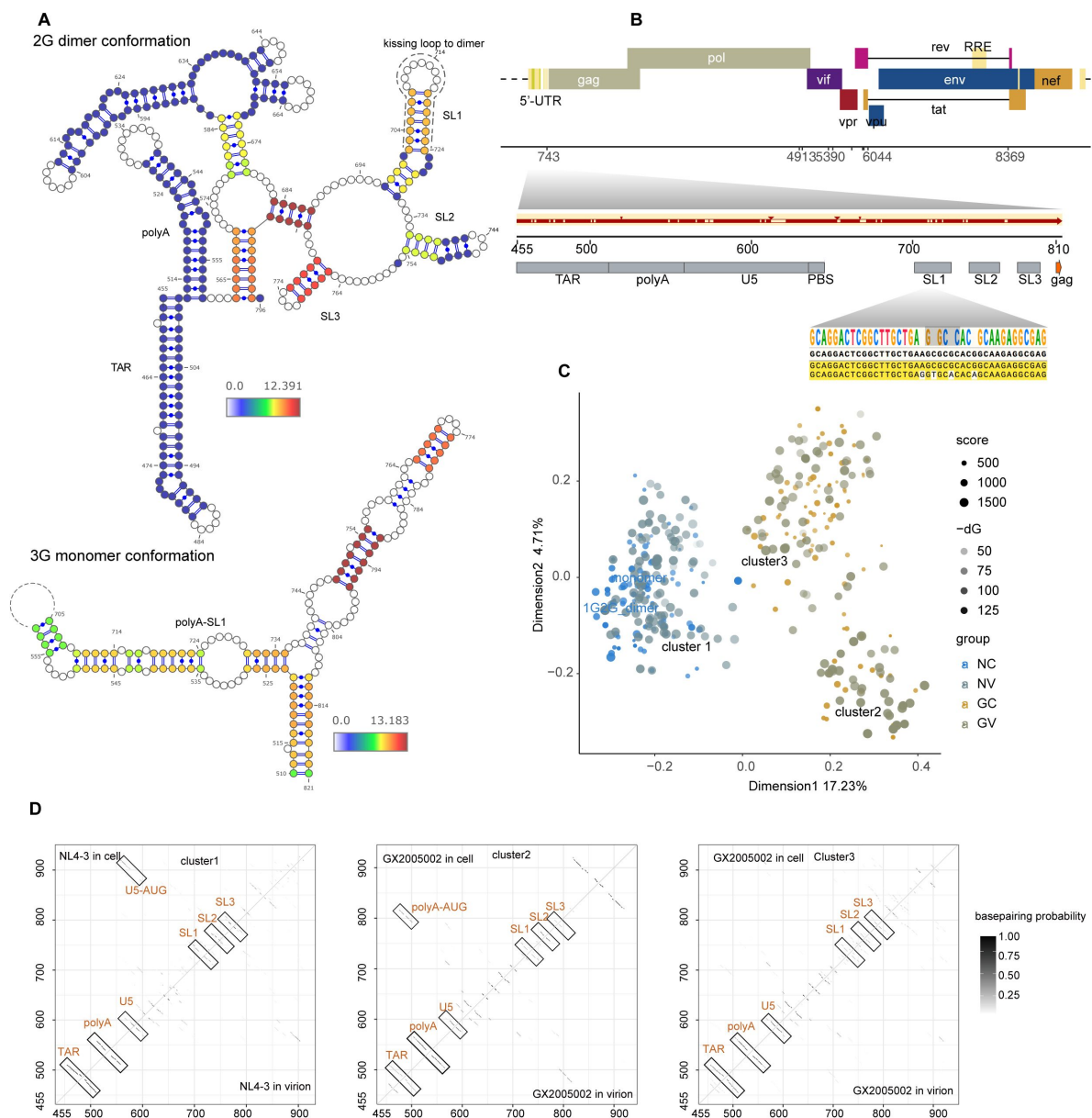


Figure 2
conformations of HIV-1 5'-UTR

A) Known conformation of HIV-1 5'-UTR supported by chimeras. Previous reported stems in 5'-UTR are supported by chimeras from HiCapR. Colors of the nucleotides indicate the log2 transformed base pairing scores. B) Representation of the one-dimensional structure of the HIV-1 5'-UTR, highlighting the conservation between the GX2005002 and NL4-3 strains in this region. The diagram includes rectangular boxes denoting the locations of key structural elements, with numerical coordinates referencing NCBI DNA genome coordinates. Dashed boxes indicate regions that are either absent or distinct in GX2005002 compared to NL4-3. Small triangle arrows indicate insertions in the GX2005002 5'-UTR. At the top, a seqlogo displays the consensus nucleotides in the SL1 region. C) Multidimensional scaling (MDS) plot clustering each of the 1,000 computationally predicted structures of the 5'-UTR in two strains under two conditions. D) Basepairing probability matrices were calculated from 1000 computationally predicted structures, with color indicating the percentage of structures supporting the specific basepair.

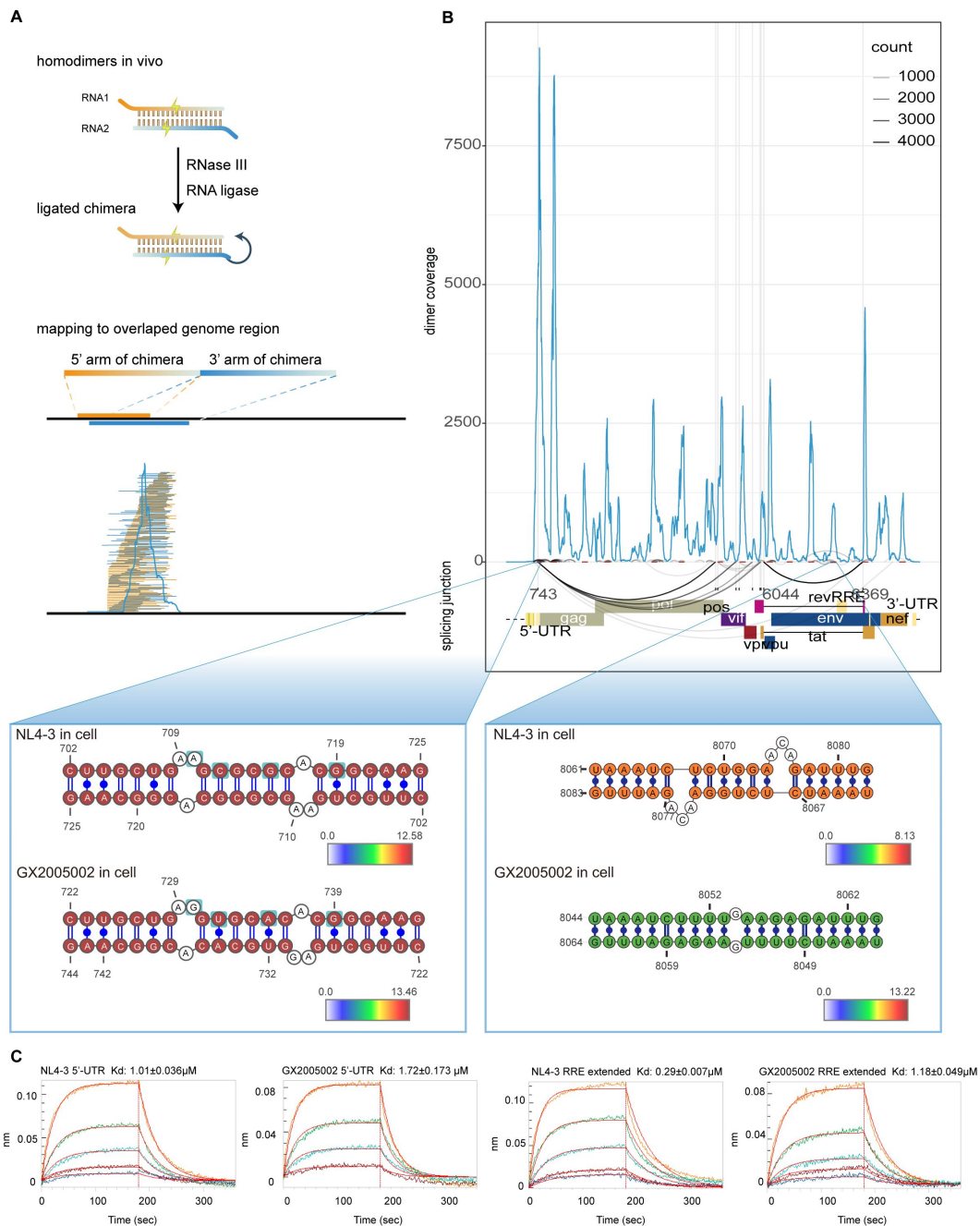


Figure 3

Identification and validation of homodimers in HIV-1 genome

A) Visual depiction of the RNA homodimer formation process: Inter-ligated fragments (arms) originating from homodimeric RNA molecules generate chimeras, where each arm aligns to overlapping coordinates on the HIV-1 genome. This process enables the plotting of coverage and specific details of dimers by utilizing the positions and counts of these overlapping chimeras. B) distribution of homodimers throughout the HIV-1 genome. The blue line plot showcases the homodimer coverage derived from ligated samples of NL4-3 infected cells along the HIV-1 genome. Arc plots exhibit discontinuous reads in non-ligated samples of NL4-3 infected cells, with dark red segments indicating peaks of homodimers. The lower panels depict the base pairing of homodimers in the SL1 region of the 5'-UTR and downstream of the RRE region. Color mapping indicates the log2 transformed dimer score. C) Assessment of dimer self-binding using Bio-Layer Interferometry (BLI). The data presented here are from three separate experiments, offering insights into how dimers interact and bind to themselves. The dissociation constant (Kd), indicated by the mean $\pm$ standard deviation, was determined from these experiments.

along the entire length of the HIV genome (Supplementary Figure 9). Homodimers were present in both infected cells and virions of both strains, with approximately 20 peaks of dimerization that are conserved between the NL4-3 and GX2005002 strains along the genome (Supplementary table2).

We investigated whether dimers could be predicted by the strength of base pairing of intermolecular loop-loop interactions by performing a systematic molecular hybridization analysis on the genomes of two HIV-1 strains, NL4-3 and GX2005002. Our findings showed no correlation between the predicted folding energy and the abundance of measured dimeric chimeras (Supplementary Figure S9A), suggesting that the formation of HIV homodimers is not solely a consequence of local base pairing propensities. Instead, it implies that additional factors, such as the binding of specific proteins, may significantly influence the dimerization process, potentially playing a more decisive role in stabilizing or facilitating the formation of these homodimers.

The identification of multiple dimers in HIV RNA beyond the 5'-UTR region raises intriguing questions regarding their potential roles serving within the HIV genome. To this end, we analyzed the sequence within dimer peaks, identifying a prevalent AG-rich motif rich, present in nearly every peak. This motif highly resembles the RNA binding motif of Serine/Arginine-Rich Splicing Factor (SRSF) (Supplementary Figure S9B,S9C and S10), which are essential RNA-binding proteins involved in pre-mRNA splicing and alternative splicing regulation³⁵, serving as the main type of HIV-1 splicing factors³⁶. This result suggests a potential role of dimerization in RNA splicing processes within the HIV genome.

5'-3' discontinuous reads were identified in non-ligation control data, enabling the inference of splicing junction sites. These inferred sites showed strong concordance with canonical splicing sites and recent nanopore sequencing data³⁷ (Figure 3B). Notably, almost every junction site, including splicing donor and acceptor sites, exhibited dimer peaks around them. Additionally, the majority of dimeric chimeras covered these junction sites and enriched downstream of acceptor sites (Supplementary Figure S10), suggesting that dimerization process/events takes place on unspliced genomic RNA. Interestingly, dimers surrounding splicing regulatory elements (as summarized in³⁶) form stable base pairing supported by numerous chimeras (Supplementary Figure S10). This observation underscores the close proximity and potential functional relationship between dimerization events and RNA splicing processes at these critical sites within the HIV genome, suggesting a potential role of dimerization in splicing processes or assisting in the transport of unspliced RNA out of the nucleus.

According to a previous report that utilized CLIP-seq to analyze the RNA binding specificity of the Gag protein in both cells and virions/From a previous CLIP-seq analyzing of Gag protein's RNA binding specificity, it was observed that Gag exhibits a preference for binding to multiple elements, such as psi and RRE, in HIV infected cells³⁸. Interestingly, significant dimer peaks were also identified in these regions (Supplementary Figure S9A). This suggests a potential connection between Gag binding and dimerization of the HIV genome.By plotting contact matrices derived from non-overlapping and dimeric chimeras around the RRE region, we were able to identify two distinct dimer peaks flanking the RRE region (Supplementary Figure S11A).

To validate these novel homodimers, we synthesized extended RRE RNA fragments with two dimer sites using in vitro transcription and confirmed their ability to form a dimeric conformation through annealing and non-denaturing gel electrophoresis (Supplementary Figure S12A), as well as Agilent Tapestation 4200 capillary electrophoresis (Supplementary Figure S12B) and Bio-Layer Interferometry (BLI) technology (Figure 3C).

Interplay between dimeric peaks and 3D genome organization

The local interaction surrounding the RRE forms an extended "arch" structure that encompasses this element, as illustrated in Supplementary Figure S13. This unique architecture may contribute to stabilizing the core RRE conformation, providing a structural basis for the rev-RRE interaction and HIV RNA transport. Additionally, we observed substantial local interaction signals around dimer sites in the 5'-UTR, as depicted in Supplementary Figure S7. To provide a comprehensive overview of the local interaction landscape around dimer sites, we generated meta matrices by computing local interactions (from non-overlapping chimeras) at dimer sites and the flanking regions located 1/2 upstream and 1/2 downstream (Supplementary Figure S12). This analysis unveiled significant local interactions around dimer sites, suggesting the involvement of stable local structures in the formation of dimers between two copies of the HIV genomes.

We observed the meta matrices around dimer sites are similar to genomic domains, which have been previously described in SARS-CoV-2²². This observation prompted us to investigate the potential existence of genomic domains in the HIV RNA genome. Using similar methods³⁹, we calculated insulation scores for two strains of the virus in both cells and virions, as depicted in **Figure 4A**.

In this way, we identified 31, 33 HIV genomic domains in infected cells and 36, 32 domains in virions for NL4-3 and GX2005002 strains, respectively. The high correlation of insulation scores between cells and virions and consistent domain boundaries are observed (**Figure 4B**), and more importantly, the boundary strengths did not show any significant differences between cells and virions (**Figure 4C**), suggesting conserved, stable and persistent genomic domains during the analyzed stages.

Interestingly, we found that dimer sites are often located within genomic domains, and overlaying dimer coverage onto domains and insulation score curves revealed a striking concordance between dimer sites and genomic domains (Supplementary Figure S14). These findings suggest an interplay between the dimerization signals and global 3D organization of the HIV genome, providing insights into the complex mechanisms of HIV-1 replication.

Dynamic changes in global and long-range interactions throughout HIV-1 life cycle

We noticed extensive long-range interactions between the HIV-1 genome of two strains (NL4-3 and GX2005002) (**Figure 1B** and Supplementary Figure S4).

To quantitatively analyze these interactions, we utilized DESeq2, following a similar approach as previous studies. Our results showed that both strains exhibit a substantial number of long-range interactions in infected cells, while these interactions are significantly reduced in virions (**Figure 5A**). Furthermore, histograms of enriched interaction pairs revealed a significant decrease in long-range interactions exceeding 2500nt in virions compared to cells (Supplementary Figure S15A). The Contact probability decay curves (PS curve) also demonstrated a faster decay of genomic interactions in virions compared to cells, for both strains of the HIV-1 virus (Supplementary Figure S15B).

Specifically, we employed viewpoint analysis to examine interactions involving the 5'-UTR. The results, depicted in **Figure 5B**, reveal multiple peaks of 5'-UTR interactions across the HIV-1 genome. Interestingly, for the majority of these peaks, the signals are lower in virions compared to cells. However, cyclization interactions (5'-UTR-3'-UTR) at the ends are still maintained in virions (**Figure 5C**). Intriguingly, we also identified multiple interaction peaks located near crucial elements, such as the frameshift element (Supplementary Figure S16A), rev start codon (Supplementary Figure S16B), RRE, and nuclear exporting signal (NES) (Supplementary Figure

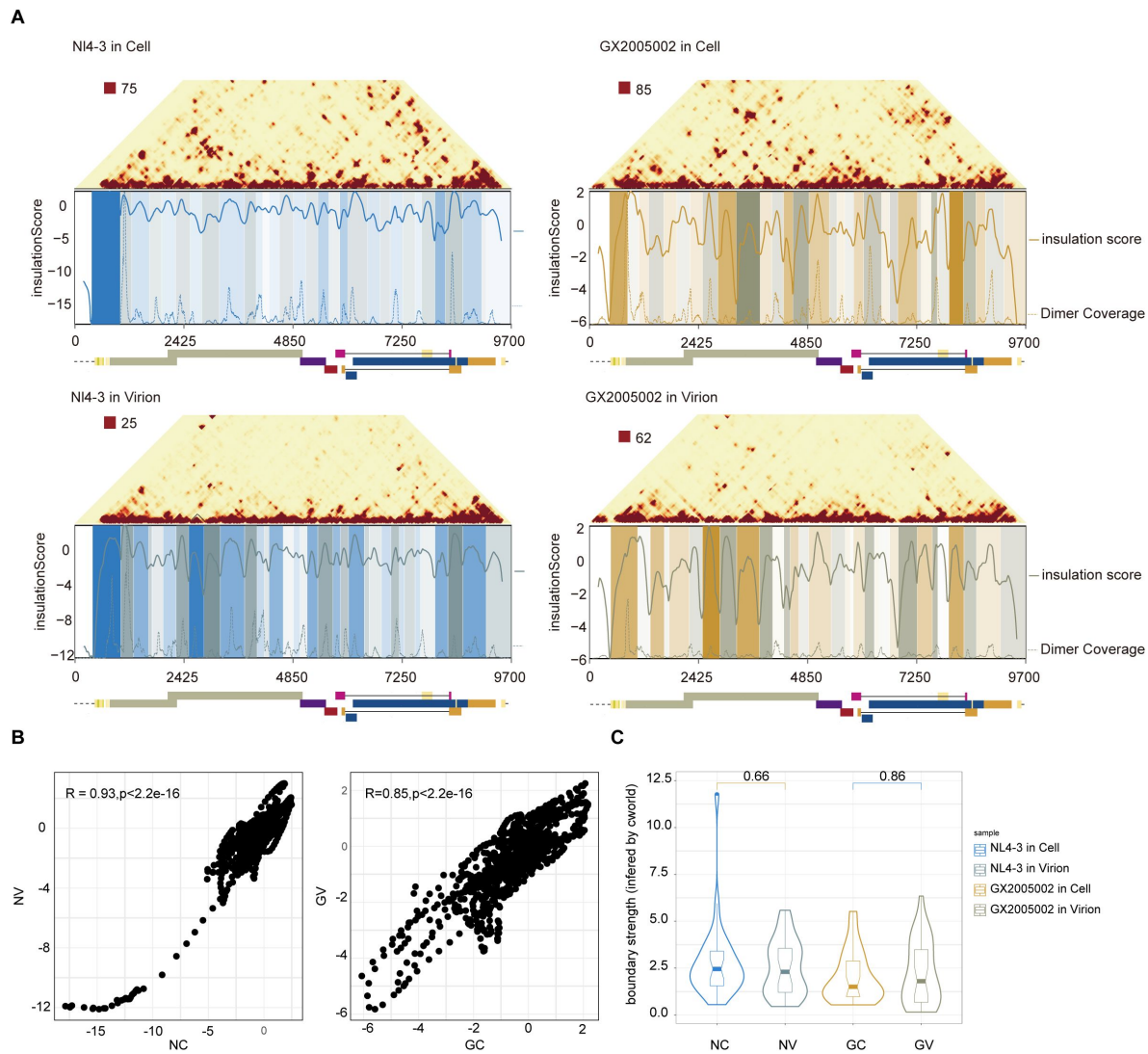


Figure 4
genome domains along HIV-1 genome

A) Each panel shows triangle matrix and genome domains of two strains of the virus in both cell and virion states, as calculated using C-world. The x-axis represents the genomic position, while the y-axis represents the insulation score (solid lines) and dimer coverage (dashed lines). B) Correlation of insulation score in infected cells and virion for NL4-3 and GX2005002 strain. C) Violin and boxplot comparing the boundary strength of the genomic domains of two strains between infected cells and virions. The boxplot displays the median, quartiles, and range of the boundary strength values for each strain in each state.

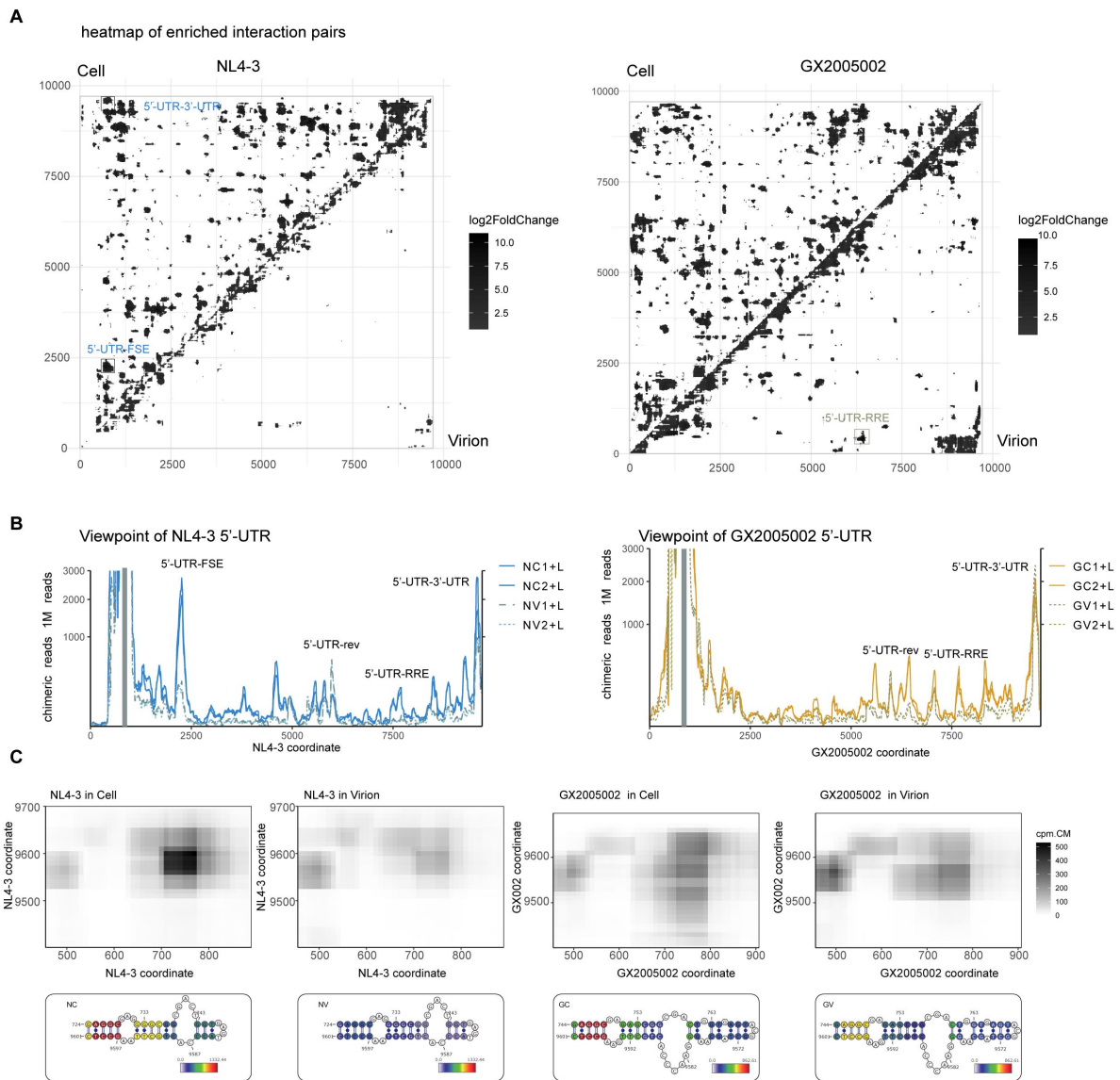


Figure 5

long range interaction in HIV-1 genome

A) heatmaps of enriched interactions obtained from NL4-3 and GX2005002 infected cells and virions. The upper diagonal shows interactions from infected cells, while the lower diagonal region displays interactions from virions B) Viewpoint lines depicts the binding positions of the 5'-UTR along the HIV-1 genome. The gray rectangles indicate the viewpoint regions. The colors of the lines represent specific samples, with samples from virions shown as dashed lines. C) Contact matrices and base pairing details between 5'-UTR and 3'-UTR. The top panels display heatmaps indicating contact probability, with the color bar indicating chimeric reads per 1M reads in each specific sample. The bottom panels show base pairing colored by base pairing scores.

S16C), indicating complex roles of 5'-UTR in HIV life cycle. Among these interactions, some are strain specific, such as the interaction between 5'-UTR and FSE is specific to the NL4-3 strain, while interaction between 5'-UTR and 8.5K region (near NES) are highly enriched in GX2005002 strain (Supplementary Figure S16C). As mentioned earlier, the majority of interactions decrease in the virion, with the notable exception of interactions in the 5'-UTR and 6K region (near D4 splicing donor site), which significantly increase in the virion. (Supplementary Figure S16B). The enhanced interactions in virions are consistent across both strains.

These results indicate that the HIV genome undergoes systematic remodeling during virus packaging, with a general loss of long-range interactions but maintenance of specific interactions, including genome cyclization. The exploration of long-range interactions and their dynamics has provided us with an unprecedented understanding of the structural organization of the HIV genome. The significance of these interactions and their impact on the HIV life cycle and pathogenesis warrants further investigation in future research.

Discussion

HicapR : a reliable and efficient method for exploring structures of viral RNA genomes

In this study, we have introduced the HiCapR protocol, which combines the SPLASH method with the capture of HIV RNA-derived libraries. This integration of RNA extraction, fragmentation, proximity ligation, library construction, and subsequent capture of target RNA involved in interactions enables the investigation of various viral genomic structures. This protocol offers a streamlined and efficient approach for studying the structure and dynamics of complete HIV RNA genomes.

HiCapR uses a similar hybridization-based capture strategy as COMRADES^{17,19} except that in HiCapR, this capture step occurs after library construction, whereas in COMRADES, it takes place prior to proximity ligation. The post-library-capture principle employed in HiCapR is a well-established approach utilized in other methods such as capture Hi-C or capture-C⁴⁰⁻⁴³, and it offers additional flexibility relative to previous protocols. Our results demonstrate remarkable sensitivity and reliability of the HiCapR technique in capturing low-abundance HIV RNA structure, providing an unprecedented resolution and comprehensive approach to elucidate the intricate RNA structures within the HIV-1 genome. This novel proximity ligation method transcends the limitations of traditional techniques, offering a robust framework for dissecting the complex architecture of viral RNA.

New insight of HIV RNA local and long-range interactions

The 5'-UTR is crucial for various stages in the HIV life cycle. Our analysis of its structures across different strains and conditions revealed consistent canonical stem-loops and interactions, despite sequence variations and alternative conformations, highlighting the structural robustness, stability and functional significance of the HIV 5'-UTR.

In addition to its local folding, the 5'-UTR engages in extensive and dynamic long-range interactions, notably with the 3'-UTR, suggesting genome cyclization akin to other viruses (Figure 4). This phenomenon, also observed in Zika, influenza, and SARS-CoV-2^{15,17,19,22}, hints at broader biological implications warranting further exploration. Additionally, the 5'-UTR interacts with key import elements in the HIV genome, such as NES, RRE, and the rev start codon. Given the Rev protein's roles in RNA transport and nuclear export⁴⁴, the 5'-UTR may also contribute to these processes.

Previous studies have highlighted folding principles of Zika, SARS-COV-2 coronavirus, and influenza virus, unraveling the intricate mechanisms underlying viral replication, pathogenesis, and host interactions [15](#), [17](#), [20](#), [22](#), [45](#). One of the key findings from this study is that the HIV-1 genome is organized in a complex three-dimensional structure that facilitates long-range interactions between distant regions of the genome. This study shed light on the dynamics of long-range interactions between the HIV-1 genome of two strains, NL4-3 and GX2005002. These interactions are largely reduced in virions compared to cells, suggesting a critical role in virus assembly and release.

Previous studies has elucidated the folding principles of various viruses like Zika, SARS-CoV-2, and influenza, shedding light on viral replication, pathogenesis, and host interactions [15](#), [17](#), [20](#), [22](#), [45](#). Our study revealed that the HIV-1 genome's intricate three-dimensional organization enables long-range interactions.

The contrasting loss of long-range interactions in the HIV-1 genome compared to SARS-CoV-2 [22](#) suggests distinct folding processes between these viruses. This disparity underscores the unique characteristics of each virus and their genomic structures. Speculations indicate that the HIV-1 genome may adopt a rod-like structure within viral particles, unlike the spherical compression seen in SARS-CoV-2 [46](#). Further investigation is needed to unravel the mechanisms and biological significance of this compression in the HIV-1 genome.

Emerging Insights into HIV-1 Dimerization through newly identified dimer sites

Our detailed analysis has consistently identified approximately 20 candidate dimer peaks both within and beyond the 5'-untranslated region (5'-UTR) across the NL4-3 and GX2005002 strains of HIV. Intriguingly, these peaks show a notable enrichment in the vicinity of splicing sites, frequently featuring a sequence motif that bears a strong resemblance to the RNA binding signature of the Serine/Arginine-Rich Splicing Factors (SRSF). This enrichment pattern and the presence of the SRSF-like motif at these sites suggest a potential regulatory role of these dimers in the complex landscape of HIV RNA splicing. The pervasive presence of homodimers flanking almost every splicing site introduces the intriguing hypothesis that these dimers could be actively participating in the regulation of alternative splicing or facilitating the export of unspliced genomic RNA from the nucleus to the cytoplasm. This hypothesis warrants rigorous investigation in subsequent studies to elucidate the mechanistic underpinnings of these observations. Collectively, these findings hint at a significant interplay between RNA dimerization and splicing processes within the HIV genome, which could have profound implications for our understanding of viral gene expression and replication strategies.

Moreover, significant dimer signals were observed in the RRE flanking sequences, adding complexity to its functional role, potentially impacting nucleocytoplasmic transport, Gag binding, and virus packaging processes. Given previous studies highlighting Gag binding to RRE and the crucial role of RRE in anchoring Gag synthesis [3](#), [38](#), [47](#), it is further hypothesized that HIV-1 homodimers may play a role in HIV splicing or assist in transporting unspliced full-length RNA out of the nucleus for genome packaging.

Interestingly, the observation that candidate dimer sites are located within genomic domains suggests that dimerization may be influenced by the local RNA folding environment (Supplementary Figure S12). It is possible that the dense local interaction around these sites may facilitate dimerization by bringing the two regions of the genome into close proximity. Alternatively, dimerization may play a role in shaping the local chromatin environment by promoting even initiating the formation of genomic domains. Further research is needed to fully understand the relationship between RNA genomic domains and dimer sites in the HIV-1 genome. However, the identification of these candidate dimer sites within genomic domains provides a

starting point for investigating the role of local genome RNA interactions in HIV-1 replication and dimerization. These findings may have implications for the development of novel antiviral therapies that target the dimerization process and may provide new avenues for future research in this field.

In summary, our study provides a comprehensive analysis of the HIV-1 genome using reliable HicapR, revealing potential dimer sites and long-range interactions, particularly within the 5'-UTR as well as persistent structure domains? These findings significantly advance our understanding of the structure, dynamics and robustness of HIV RNA and their involvement in splicing, assembly and packaging processes, potentially leading to the development of novel antiviral therapies.

Methods

Cell culture and virus infection

The MT4 cell lines (RRID:CVCL_2632) were cultured under specific conditions: RPMI 1640 medium (Gibco, USA) supplemented with 10% fetal calf serum (Gibco, USA), 100 U/ml penicillin, and 100 µg/ml streptomycin. The cells were maintained at a temperature of 37°C with 5% CO₂ and saturated humidity. To initiate infection, cells were exposed to HIV-1 NL4-3 (HIV-1 strain of subtype B) or HIV-1 GX2005002 (HIV-1 primary strain of CRF01_AE which is one of the main strain in China, accession: GU564222) at a multiplicity of infection (MOI) of 0.15.. Concurrently, parallel control groups consisting of uninfected cells were also established. Both the cells and the cell supernatant were collected after 48hours post infection for subsequent experiments.

HicapR method

First, extracted crosslinked RNA was treated as in simplified SPLASH protocol ²²[\[22\]](#).

Briefly, 500 ng of each sample was fragmented using RNase III (Ambion) in a 20 µl mixture for 10 minutes at 37°C. The fragmented RNA was then purified using 40 µl of MagicPure RNA Beads (TransGen). Each RNA sample was subsequently divided into two halves, with one half used for proximity ligation and crosslink reversal (C, V samples). The proximity ligation process was then carried out using the following conditions: 200 ng of fragmented RNA, 1 unit/µl RNA ligase 1 (New England Biolabs), 1× RNA ligase buffer, 1mM ATP, 1 unit/µl Superase-in (Invitrogen), with a final volume of 200 µl. The reactions were incubated for 16 hours at 16°C and were stopped by cleaning with the miRNeasy kit (Qiagen). To reverse the crosslinking, the RNA was irradiated on ice with 254 nm UltraViolet C radiation for 5 minutes using a CL-1000 crosslinker (UVP). For the non-ligated controls, crosslink reversal was performed immediately after crosslinking, without proximity ligation (these controls were labeled as "-L").

The proximity ligated RNA was then subjected to library construction using SMARTer Stranded Total RNA-Seq Kit v2 (Clonetech).

The cDNA libraries were enriched for HIV fragment using the TargetSeq One Hyb & Wash kit (igenetech) with the T548XV1 probe panel, designed based on HIV genome. The probe sequences are provided as supplementary data

Data preprocessing and chimeric reads identification

The data preprocessing and identification of chimeric reads were conducted as previously described. The reference genome used for NL4-3 strain was a combination of NL4-3 genome sequence (<https://www.ncbi.nlm.nih.gov/nuccore/AF003887> [\[23\]](#)) and human spliced mRNAs and

noncoding RNAs described in [30](#), while for GX2005002 strain, the reference genome used was a combination of GX2005002 genome sequence (<https://www.ncbi.nlm.nih.gov/nucore/KP178420>) and human spliced mRNAs and noncoding RNAs as above.

First, we used pear to merge the overlapped reads:

```
pear -e -j 32 -f sample_R2.fastq.gz -r sample_R1.fastq.gz -o Sample.PEAR
```

note that for Clonetech 634413, the sense strand of RNAs is in R2 read.

Then, we used fastp to filter reads with low quality and cut adapters:

```
fastp -i sample.PEAR.fastq -o sample.PEAR.fastp.fastq -h fastp.html -w 4 -a AGATCGGAAGAGCGTCG
```

And then, chimeric reads are detected using hyb pipeline with default setting:

```
hyb detect in= sample .PEAR.fastp.fastq db=$DB qc=none
```

Chimeras for RNA-RNA interaction and dimer identification

Homodimers were identified in accordance with the methods previously reported in the literature, with a more stringent filter²³. Specifically, we utilized the hub pipeline and filtered chimeras in .hyb files. To calculate the degree of overlap between the two arms of chimeric reads, we used the formula $L=1+\min(e1,e2)-\max(s1,s2)$, where e1 and e2 represent the ends of each arm of chimeric reads, while s1 and s2 represent their starts. To ensure that possible circular RNAs, which may be produced by single RNA end-to-end ligation, are filtered out, we applied the following condition: $(e1 < e2)$ OR $(s1 < s2)$. A schematic diagram of dimer chimeras is provided in Supplementary Figure S1.

We defined the dimer range as the maximum value between s1 and s2, and the minimum value between e1 and e2. Based on this range, we further calculated the coverage of dimer chimeras.

On the other hand, only non-overlapping chimeras (with $L < 0$) were included for RNA-RNA interaction contact matrices to reduce the possibility that the chimeras called for RNA-RNA interactions come from homodimers.

Dimer score and dimer base pairing

The dimer score (DS) was defined as the number of chimeric reads that supported a potential dimer base-pairing event. To visualize each particular candidate dimer, we used the hybrid-min command in the RNA fold package to in-silico hybridize two homodimers. The paired bases were then colored using the aforementioned dimer score.

Local structure folding and MDS plot

Initial local structure folding was performed similarly to a previous report¹⁷. Base pairing scores were first calculated using comradesMakeConstraints function in COMRADES package(<https://github.com/gkudla/comrades>):

```
comradesMakeConstraints -i sample_rm_overlap.hyb -f Genome_fasta -b 1 -e genome_length
```

we folded the RNA structures 1000 times using the COMRADES manual. Subsequently, they utilized multidimensional scaling to calculate the distances between RNA structures.

Calculation of base pairing probability for local structure

For a particular cluster in the MDS plot, we construct a base pairing probability matrix to reveal consensus stems in these structures, where the probability of base pairing for each base pair is determined by calculating the proportion of structures that support the base pair among all folded structures.

Dimer validation

RNA preparation. 5'-UTR and RRE DNA fragments were amplified by RT-PCR using RNA from infected cells, with primers listed in Supplementary table3. Purified NL4-3 and GX2005002 5'-UTR and extended RRE PCR products (200 ng) were utilized as templates for RNA in vitro transcription with T7 RNA polymerase (Vazyme), following manufacturer's protocol. The reaction was then incubated at 37 °C for 16 hours, followed by DNase I treatment for 30 minutes at 37 °C. The RNA was subsequently purified using MagicPure RNA Beads (TransGen) through gel purification.

Native agarose gel electrophoresis

RNA (600 ng) was heated to 95°C and then slowly cooled to room temperature in either high salt buffer (50mM Tris-HCl pH 7.5, 140mM KCl, 10mM NaCl) or low salt buffer (high salt buffer diluted 1:10 with water). Samples were loaded with native loading dye (0.17% Bromophenol Blue and 40% (vol/vol) sucrose) on 2% agarose gel prepared with 1× tris-borate magnesium (TBM) buffer (89 mM Tris base, 89 mM boric acid and 2 mM MgCl₂) and fractionated at 100 V for 85 min at room temperature. RNA in the PAGE gel was visualized using GelRed nucleic acid gel stain(Thomas scientific).

Agilent 4200 TapeStation Capillary electrophoresis

We prepared the RNA samples as described above and loaded them onto the TapeStation RNA screentape without the heating and denaturing step. Subsequently, we initiated the device for electrophoresis analysis, which ran automatically without the need for manual intervention. After the electrophoresis analysis was completed, the system generated electropherograms and relevant data regarding the RNA samples, including their size distribution and concentration. This method allowed us to quickly and accurately evaluate the conformational features of the RNA.

Biomolecular Binding Kinetics Assays

Biomolecular Binding Kinetics Assays were performed using the Octet R8 Platform (sartorius) following a standardized protocol. The assay involved the preparation of analytes at varying concentrations, real-time monitoring of association and dissociation phases, and subsequent data analysis to determine kinetic parameters k_d . Quality control measures were implemented to ensure the reliability and reproducibility of the binding kinetics measurements.

Data availability

The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive ⁴⁸[in](#) National Genomics Data Center ⁴⁹[China National Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of Sciences \(GSA: CRA016024\)](#) that are publicly accessible at <https://ngdc.cncb.ac.cn/gsa>.

Additional information

Code availability

Custom code for the analysis performed in this study is publicly available via biocode at <https://ngdc.cncb.ac.cn/biocode/tools/BT007456>

Additional files

Supplementary Figures. Supplementary Figure S1. This figure illustrates the workflow for identifying chimeric reads supporting RNA-RNA interactions as well as homodimers. The rounded rectangle in the upper right corner of the figure displays the start and end positions of the 5'- and 3'-arms of chimeric reads. These arms can be ligated in multiple ways to form chimeras, some of which are thought to be produced exclusively by homodimers. First, sequencing reads are mapped to the HIV RNA genome, and chimeric reads are identified using hyb pipeline. Dimeric chimeras are then filtered to select only those that can be formed exclusively by dimerization. To reduce background noise, especially contaminated from homodimers, we selected non-overlapping chimeras for our analysis of RNA interactions. This approach allowed us to identify both RNA-RNA interactions and candidate dimer sites and predict secondary structures with high coverage chimeras. **Supplementary Figure S2.** HIV Genome coverage of HiCapR data. (A) The organization of the HIV NL4-3 genome follows NCBI annotation and is specified. (B) and (C) The sequencing data from each group was mapped to the HIV genome and then genome coverage was calculated. The results showed that both NL4-3 (B) and GX2005002 (C) strains of the virus, whether in infected cells or virions, and whether RNAs are ligated or not, had uniformly complete genome coverage. This suggests that the HiCapR method is able to capture the entire HIV genome with low bias. **Supplementary Figure S3.** The heatmap displays the correlation between samples, which was measured using Pearson's coefficients. **Supplementary Figure S4.** Related to Figure 1B, but with the inclusion of information from individual replicates and data from non-ligation controls, offering a more detailed view of the interactions and distributions presented in Figure 1B. **Supplementary Figure S5.** The established conformation of the HIV-1 5'-UTR, supported by chimeras in NL4-3 and GX2005002, both in cells and in virions. Colors of the nucleotides indicate the log2 transformed base pairing scores. **Supplementary Figure S6.** The stacked bar chart displays the number of 3'-5', 5'-3', and dimer chimeras in various ranges across the sample groups. The panel on the left pertains to NL4-3, while the one on the right pertains to GX2005002. **Supplementary Figure S7.** (A) Heatmaps simultaneously plot contact matrices calculated from non-overlapped chimeras, which is informative for interactions and dimer chimeras which is derived for homodimers. The coordination in heatplots is RNA-based, with the numbers in brackets indicating the corresponding coordinates in the DNA genome. (B) Homodimer chimeras details around splicing sites show the 5'-arm represented by blue lines and the 3'-arm represented in yellow, providing a visual representation of the interactions in these regions. **Supplementary Figure S8.** Dimer score matrices around 5'-UTR of NL4-3 and GX2005002 strain in different stages. **Supplementary Figure S9.** (A) Homodimer coverage along HIV-1 genome in NL 4-3 and GX2005002 both in cells and in virion samples. Gag-CLIP data from Kutluay, S. B., et al. (2014) are also shown. Bottom panel depict delta G of each 100-nt window in NL4-3 genome. (B) Distribution of enriched motif in homodimer peaks of NL4-3, region similar to SRSF motif was underlined in the seqlog. (C) Distribution of enriched motif in homodimer peaks of GX2005002, region similar to SRSF motif was underlined in the seqlog. **Supplementary Figure S10.** Top: Homodimer chimeras details around splicing sites show the 5'-arm represented by blue lines and the 3'-arm represented in yellow, providing a visual representation of the interactions in these regions. Known splicing regulatory elements are also indicated. Bottom: basepairing of homodimers around splicing sites as indicated. Color code indicates log2 transformed base pairing score supporting each base-pair. **Supplementary Figure S11.** (A) Contact matrix and homodimeric chimeras surrounding the RRE region in the HIV genome. The left panel displays the contact matrix derived from non-overlapping chimeras and

the contact matrix derived from dimeric chimeras. The right panel presents specific details of the dimeric chimeras around the RRE elements. (B) Base pairing of homodimers in the extended RRE region in specific samples. Color mapping indicates the log₂ transformed dimer score. **Supplementary Figure S12.** (A) Native agarose gel confirming dimeric conformation of RRE (extended to 620nt). (B) Agilent tapestation 4200 capillary electrophoresis confirming dimeric conformation of 5'-UTR and RRE (extended to 620nt) **Supplementary Figure S13.** Predicted structure of extended RRE. Color code indicates log₂ transformed base pairing score supporting each base-pair. **Supplementary Figure S14.** Average contact matrices around peak centers. Heatmaps showing the normalized average interaction frequencies for all peak centers as well as their flanking regions (± 0.5 peak width). The heatmaps were binned at 10 nt resolution. **Supplementary Figure S15.** (A) histograms show distributions of different types of interactions across the HIV genome. The x-axis represents the span of the interactions, while the y-axis represents the frequency of each interaction type. (B) Contact probability decay curves (Ps curves) showing progressive reconfiguration HIV genomes. Both two strains of the virus exhibit faster decay rates within virions than cells. **Supplementary Figure S16.** Contact matrices and base pairing details between 5'-UTR and 2.2K (A), 6K (B) and 7.6~8.5K (C). [↗](#)

Supplementary Table 1. Mapping statistics. [↗](#)

Supplementary Table 2. Homodimer peaks information. [↗](#)

Supplementary Table 3. Primers used for amplifying 5'-UTR and extended RRE. [↗](#)

Supplementary Data. Designed probes for enriching HIV chimeras in fasta format. [↗](#)

References

- 1 Mallapaty S (2024) **Seventh patient ‘cured’ of HIV: why scientists are excited** *Nature* **632**:235–236
- 2 Freed E. O (2015) **HIV-1 assembly, release and maturation** *Nat Rev Microbiol* **13**:484–496 <https://doi.org/10.1038/nrmicro3490>
- 3 Emery A., Swanstrom R (2021) **HIV-1: To Splice or Not to Splice, That Is the Question** *Viruses* **13**
- 4 Keane S. C., et al. (2015) **RNA structure. Structure of the HIV-1 RNA packaging signal** *Science* **348**:917–921 <https://doi.org/10.1126/science.aaa9266>
- 5 Zhang K., et al. (2018) **Structure of the 30 kDa HIV-1 RNA dimerization signal by a hybrid Cryo-EM, NMR, and molecular dynamics approach** *Structure* **26**:490–498
- 6 Watts J. M., et al. (2009) **Architecture and secondary structure of an entire HIV-1 RNA genome** *Nature* **460**:711–716
- 7 Tomezsko P. J., et al. (2020) **Determination of RNA structural diversity and its role in HIV-1 RNA splicing** *Nature* **582**:438–442 <https://doi.org/10.1038/s41586-020-2253-5>
- 8 Ye L., et al. (2022) **Short- and long-range interactions in the HIV-1 5’ UTR regulate genome dimerization and packaging** *Nat Struct Mol Biol* **29**:306–319 <https://doi.org/10.1038/s41594-022-00746-2>
- 9 Ding P., et al. (2021) **5’-Cap sequestration is an essential determinant of HIV-1 genome packaging** *Proc Natl Acad Sci U S A* **118** <https://doi.org/10.1073/pnas.2112475118>
- 10 Chen J., et al. (2020) **Visualizing the translation and packaging of HIV-1 full-length RNA** *Proceedings of the National Academy of Sciences* **117**:6145–6155
- 11 Sherpa C., Rausch J. W., Le Grice S. F., Hammariskjold M. L., Rekosh D (2015) **The HIV-1 Rev response element (RRE) adopts alternative conformations that promote different rates of virus replication** *Nucleic Acids Res* **43**:4676–4686 <https://doi.org/10.1093/nar/gkv313>
- 12 Jackson P. E., Tebit D. M., Rekosh D., Hammariskjold M. L (2016) **Rev-RRE Functional Activity Differs Substantially Among Primary HIV-1 Isolates** *AIDS Res Hum Retroviruses* **32**:923–934 <https://doi.org/10.1089/AID.2016.0047>
- 13 Saldi T., Riemondy K., Erickson B., Bentley D. L (2021) **Alternative RNA structures formed during transcription depend on elongation rate and modify RNA processing** *Molecular cell* **81**:1789–1801
- 14 Le Sage V., et al. (2020) **Mapping of Influenza Virus RNA-RNA Interactions Reveals a Flexible Network** *Cell Rep* **31** <https://doi.org/10.1016/j.celrep.2020.107823>
- 15 Miyamoto S., Noda T (2020) **In vitro vRNA-vRNA interactions in the H1N1 influenza A virus genome** *Microbiol Immunol* **64**:202–209 <https://doi.org/10.1111/1348-0421.12766>

- 16 Li P., et al. 2018) **Integrative Analysis of Zika Virus Genome RNA Structure Reveals Critical Determinants of Viral Infectivity** *Cell Host Microbe* **24**:875–886 <https://doi.org/10.1016/j.chom.2018.10.011>
- 17 Ziv O., et al. 2018) **COMRADES determines in vivo RNA structures and interactions** *Nat Methods* **15**:785–788 <https://doi.org/10.1038/s41592-018-0121-0>
- 18 Huber R. G., et al. 2019) **Structure mapping of dengue and Zika viruses reveals functional long-range interactions** *Nature communications* **10**:1408 <https://doi.org/10.1038/s41467-019-09391-8>
- 19 Ziv O., et al. 2020) **The Short- and Long-Range RNA-RNA Interactome of SARS-CoV-2** *Mol Cell* <https://doi.org/10.1016/j.molcel.2020.11.004>
- 20 Cao C., et al. 2021) **The architecture of the SARS-CoV-2 RNA genome inside virion** *Nature communications* **12** <https://doi.org/10.1038/s41467-021-22785-x>
- 21 Yang S. L., et al. 2021) **Comprehensive mapping of SARS-CoV-2 interactions in vivo reveals functional virus-host interactions** *Nature communications* **12**:5113 <https://doi.org/10.1038/s41467-021-25357-1>
- 22 Zhang Y., et al. 2021) **In vivo structure and dynamics of the SARS-CoV-2 RNA genome** *Nature communications* **12**:5695 <https://doi.org/10.1038/s41467-021-25999-1>
- 23 Gabryelska M. M., et al. 2022) **Global mapping of RNA homodimers in living cells** *Genome Res* <https://doi.org/10.1101/gr.275900.121>
- 24 Zhang M., et al. 2022) **Classification and clustering of RNA crosslink-ligation data reveal complex structures and homodimers** *Genome Res* **32**:968–985 <https://doi.org/10.1101/gr.275979.121>
- 25 Jäger R., et al. 2015) **Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci** *Nature communications* **6**:6178
- 26 Dryden N. H., et al. 2014) **Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C** *Genome research* **24**:1854–1868
- 27 Sahlen P., et al. 2015) **Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution** *Genome Biol* **16** <https://doi.org/10.1186/s13059-015-0727-9>
- 28 Mifsud B., et al. 2015) **Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C** *Nature genetics* **47**:598–606
- 29 Chang S., et al. 2015) **Comparison of susceptibility of HIV-1 variants to antiretroviral drugs by genotypic and recombinant virus phenotypic analyses** *International Journal of Infectious Diseases* **37**:86–92
- 30 Travis A. J., Moody J., Helwak A., Tollervey D., Kudla G 2014) **Hyb: a bioinformatics pipeline for the analysis of CLASH (crosslinking, ligation and sequencing of hybrids) data** *Methods* **65**:263–273 <https://doi.org/10.1016/j.ymeth.2013.10.015>
- 31 Stephenson J. D., et al. 2013) **Three-dimensional RNA structure of the major HIV-1 packaging signal region** *Structure* **21**:951–962

- 32 Mailler E., et al. 2016) **The Life-Cycle of the HIV-1 Gag-RNA Complex** *Viruses* **8** <https://doi.org/10.3390/v8090248>
- 33 Ding P., et al. 2020) **Identification of the initial nucleocapsid recognition element in the HIV-1 RNA packaging signal** *Proc Natl Acad Sci U S A* **117**:17737–17746 <https://doi.org/10.1073/pnas.2008519117>
- 34 Zhang K., et al. 2018) **Structure of the 30 kDa HIV-1 RNA Dimerization Signal by a Hybrid Cryo-EM, NMR, and Molecular Dynamics Approach** *Structure* **26**:490–498 <https://doi.org/10.1016/j.str.2018.01.001>
- 35 Änkö M.-L., et al. 2012) **The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes** *Genome biology* **13**:1–17
- 36 Sertznig H., Hillebrand F., Erkelenz S., Schaal H., Widera M 2018) **Behind the scenes of HIV-1 replication: Alternative splicing as the dependency factor on the quiet** *Virology* **516**:176–188 <https://doi.org/10.1016/j.virol.2018.01.011>
- 37 Nguyen Quang N., et al. 2020) **Dynamic nanopore long-read sequencing analysis of HIV-1 splicing events during the early steps of infection** *Retrovirology* **17** <https://doi.org/10.1186/s12977-020-00533-1>
- 38 Kutluay S. B., et al. 2014) **Global changes in the RNA binding specificity of HIV-1 gag regulate virion genesis** *Cell* **159**:1096–1109 <https://doi.org/10.1016/j.cell.2014.09.057>
- 39 Crane E., et al. 2015) **Condensin-driven remodelling of X chromosome topology during dosage compensation** *Nature* **523**:240–244 <https://doi.org/10.1038/nature14450>
- 40 Martin P., et al. 2015) **Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci** *Nature communications* **6**
- 41 Baxter J. S., et al. 2018) **Capture Hi-C identifies putative target genes at 33 breast cancer risk loci** *Nature communications* **9**:1028
- 42 Orlando G., et al. 2018) **Promoter capture Hi-C-based identification of recurrent noncoding mutations in colorectal cancer** *Nature genetics* **50**:1375
- 43 Chesi A., et al. 2019) **Genome-scale Capture C promoter interactions implicate effector genes at GWAS loci for bone mineral density** *Nature communications* **10**:1260
- 44 Li G., De Clercq E 2016) **HIV genome-wide protein associations: a review of 30 years of research** *Microbiology and Molecular Biology Reviews* **80**:679–731
- 45 Dadonaite B., et al. 2019) **The structure of the influenza A virus genome** *Nature Microbiology* **4**:1781–1789 <https://doi.org/10.1038/s41564-019-0513-7>
- 46 Yao H., et al. 2020) **Molecular Architecture of the SARS-CoV-2 Virus** *Cell* **183**:730–738 <https://doi.org/10.1016/j.cell.2020.09.018>
- 47 Becker J. T., Sherer N. M 2017) **Subcellular Localization of HIV-1 gag-pol mRNAs Regulates Sites of Virion Assembly** *J Virol* **91** <https://doi.org/10.1128/JVI.02315-16>
- 48 Chen T., et al. 2021) **The genome sequence archive family: toward explosive data growth and diverse data types.** *Genomics Proteomics and Bioinformatics* **19**:578–583

- 49 CNCB-NGDC Members and Partners 2022) **Database resources of the national genomics data center, China national center for bioinformation in 2022** *Nucleic Acids Research* **50**:D27–D38

Author information

Yan Zhang[#]

Laboratory of Advanced Biotechnology, Beijing Institute of Biotechnology, Beijing, China
ORCID iD: [0000-0003-4006-663X](https://orcid.org/0000-0003-4006-663X)

[#]The authors wish to note that the first three authors contributed equally

Jingwan Han[#]

Institute of Microbiology and Epidemiology, Beijing, China

[#]The authors wish to note that the first three authors contributed equally

Dejian Xie[#]

Laboratory of Advanced Biotechnology, Beijing Institute of Biotechnology, Beijing, China
ORCID iD: [0009-0000-3025-3568](https://orcid.org/0009-0000-3025-3568)

[#]The authors wish to note that the first three authors contributed equally

Wenlong Shen

Laboratory of Advanced Biotechnology, Beijing Institute of Biotechnology, Beijing, China
ORCID iD: [0000-0002-2293-5590](https://orcid.org/0000-0002-2293-5590)

Ping Li

Laboratory of Advanced Biotechnology, Beijing Institute of Biotechnology, Beijing, China

Jian You Lau

MRC Human Genetics Unit, University of Edinburgh, Edinburgh, United Kingdom

Jingyun Li

Institute of Microbiology and Epidemiology, Beijing, China

For correspondence: lijyjk@163.com

Lin Li

Institute of Microbiology and Epidemiology, Beijing, China
ORCID iD: [0000-0001-6124-5237](https://orcid.org/0000-0001-6124-5237)

For correspondence: dearwood@sina.com

Grzegorz Kudla

MRC Human Genetics Unit, University of Edinburgh, Edinburgh, United Kingdom

ORCID iD: [0000-0002-7924-2744](https://orcid.org/0000-0002-7924-2744)

For correspondence: gkudla@gmail.com

Zhihu Zhao

Laboratory of Advanced Biotechnology, Beijing Institute of Biotechnology, Beijing, China

ORCID iD: [0000-0001-7081-8914](https://orcid.org/0000-0001-7081-8914)

For correspondence: zhaozh@bmi.ac.cn

Editors

Reviewing Editor

Volker Dötsch

Goethe University Frankfurt, Frankfurt am Main, Germany

Senior Editor

Volker Dötsch

Goethe University Frankfurt, Frankfurt am Main, Germany

Reviewer #1 (Public review):

This paper focuses on secondary structure and homodimers in the HIV genome. The authors introduce a new method called HiCapR which reveals secondary structure, homodimer, and long-range interactions in the HIV genome. The experimental design and data analysis are well-documented and statistically sound.

Comments on revisions:

The authors have addressed key questions and highlighted the advantages of HiCapR.

<https://doi.org/10.7554/eLife.102550.2.sa2>

Reviewer #2 (Public review):

Summary:

In the manuscript "Mapping HIV-1 RNA Structure, Homodimers, Long-Range Interactions and 1 persistent domains by HiCapR" Zhang et al report results from an omics-type approach to mapping RNA crosslinks within the HIV RNA genome under different conditions i.e. in infected cells and in virions. Reportedly, they used a previously published method which, in the present case, was improved for application to RNAs of low abundance.

Their claims include the detection of numerous long-range interactions, some of which differ between cellular and virion RNA. Further claims concern the detection and analysis of homodimers.

Strengths:

(1) The method developed here works with extremely little viral RNA input and allows for the comparison of RNA from infected cells versus virions.

(2) The findings, if validated properly, are certainly interesting to the community.

Weaknesses:

(1) On the communication level, the present version of the manuscript suffers from a number of shortcomings. I may be insufficiently familiar with habits in this community, but for RNA aficionados just a little bit outside of the viral-RNA-X-link community, the original method (reference 22) and the presumed improvement here are far too little explained, namely in something like three lines (98-100). This is not at all conducive to further reading.

(2) Experimentally, the manuscript seems to be based on a single biological replicate, so there is strong concern about reproducibility.

(3) The authors perform an extensive computational analysis from a limited number of datasets, which are in thorough need of experimental validation

Comments on revisions:

The authors have made cosmetic changes with regards to the problems I raised. 1 - Reproducibility: the rebuttal letter says there are now 3 replicates, but there is only data for 2 in the supplement. The generation of biological replicates needs to be precisely stated, e.g. taken on different days, from separate cultures, or from neighbouring dishes on the same day etc. I think, the manuscript would greatly benefit from the comparison of at least 3 replicates that were not generated on the same day. Given that the authors report a r^2 of 0.99 between the sets they have, this seems quite plausible.

The validation of the dimerisation sites is marginally better, but the authors should read up on significant digits and how precise K_d values can be determined.

The authors state that they want to make several of the experiments that would address my issues in the future in the context of another study. I find that disappointing, and correspondingly the present datasets insufficient for further endorsement.

<https://doi.org/10.7554/eLife.102550.2.sa1>

Author response:

The following is the authors' response to the original reviews.

Public Reviews:

Reviewer #1 (Public review)

This paper focuses on secondary structure and homodimers in the HIV genome. The authors introduce a new method called HiCapR which reveals secondary structure, homodimer, and long-range interactions in the HIV genome. The experimental design and data analysis are well-documented and statistically sound. However, the manuscript could be further improved in the following aspects.

Major comments:

(1) Please give the full name of an abbreviation the first time it appears in the paper, for example, in L37, "5' UTR" "RRE".

Thank you for your suggestion. We have added the full name of these abbreviations.

(2) The introduction could be strengthened by discussing the limitations of existing methods for studying HIV RNA structures and interactions and highlighting the specific advantages of the HiCapR method.

Thank you for your insightful suggestion. We have modified sentences in the introduction section (line 66 -line 71, line 80-line 81 in the revised manuscript).

| (3) Please reorganize Results Part 1.

Thank you for your advice. We have reorganized results part 1. We hope the revision provides a logical flow and clarity to the results, making it easier for readers to follow the progression of the study and the significance of the findings regarding to the HiCapR method.

| (4) Is there any reason that the authors mention "genome structure of SARS-CoV-2" in L95?

Thank you for your insightful question. We have deleted this sentence in the revised paper.

Initially, the mention of our previous work on SARS-CoV-2 serves two purposes: firstly, to demonstrate our capability to perform proximity ligation assays on viral samples; and secondly, to underscore the necessity of the hybridization step, which is particularly relevant for the study of HIV.

Unlike SARS-CoV-2, which is highly abundant in infected cells and does not require post-library hybridization, HIV-1 presents a unique challenge due to its typically low viral RNA input within cells. The simplified SPLASH protocol, while effective for more abundant viral RNAs, does not provide the necessary coverage for high-resolution analysis when applied directly to HIV samples.

Now, we have deleted this sentence according to your comments, and discuss the technical difference elsewhere.

| (5) L102: Please clarify the purpose of comparing "NL4-3" and "GX2005002." Additionally, could you explain what NL4-3 and GX2005002 are? The connection between NL4-3, GX2005002, and HIV appears to be missing.

Thank you for your question, and we apologize for the misleading. "NL4-3" and "GX2005002" are two distinct HIV-1 strains that exhibit different prevalence patterns in various geographical regions. The NL4-3 strain is a well-characterized laboratory strain that is widely used in HIV research and is representative of the HIV-1 subtype B, which is highly prevalent in Europe and the Americas. On the other hand, GX2005002 is a primary isolate of the CRF01_AE subtype, which is one of the most prevalent strains in Southeast Asia, particularly in China.

The reason for comparing these two strains in our study is twofold. Firstly, it allows us to assess the applicability and versatility of our HiCapR method across different HIV-1 strains that may have distinct genetic and structural features. This is crucial for understanding the potential broad utility of our method in studying various HIV-1 strains globally. Secondly, by comparing these strains, we can begin to elucidate any strain-specific differences in RNA structure, homodimer formation, and long-range interactions, which may have implications for viral pathogenesis, transmission, and response to therapeutic interventions.

The connection between NL4-3, GX2005002, and HIV lies in their representation of different subtypes of the HIV-1 virus, which exhibit genetic diversity and are associated with different geographical distributions. This diversity is epidemiologically and clinically relevant, as it may be associated with different pathogenesis and resistance mechanisms, and might have implications for vaccine development and treatment strategies.

| (6) Figure 1A is not able to clearly present the innovation point of HiCapR.

Thank you for your comment. We have revised this figure to more clearly illustrate the steps and principles of the post-library capture process using HIV pooled probes hybridization and streptavidin pull down to enrich HIV RNA-derived chimeras.

(7) Please compare the contact metrics detected by HiCapR and current techniques like SHAPE on the local interactions to assess the accuracy of HiCapR in capturing local RNA interactions relative to established methods.

Thank you for your request to compare the contact metrics detected by HiCapR and current techniques like SHAPE on local interactions to assess the accuracy of HiCapR in capturing local RNA interactions relative to established methods.

In this study, HiCapR has demonstrated its ability to identify key structural elements within the HIV genome, including TAR, polyA, SL1, SL2, and SL3, as well as the polyA-SL1 in the monomeric conformation. These elements are crucial for understanding the local RNA structures involved in HIV replication and pathogenesis. By visualizing the base pairing probability as a heatmap, we have identified the most stable base pairs in the 5' UTR of HIV, which is consistent across both NL4-3 and GX2005002 strains (Figure 2D). This consistency suggests robustness in the overall structure despite sequence variations and alternative RNA conformations, indicating a high level of agreement between HiCapR and SHAPE methods in detecting local interactions.

Furthermore, HiCapR not only confirms the presence of known structural elements but also reveals alternative conformations of the 5'UTR that support the alternative conformations found in SHAPE analysis. This additional layer of information provides a more comprehensive view of the RNA structures, highlighting HiCapR's ability to capture local RNA interactions with a high degree of accuracy comparable to established methods like SHAPE.

(8) The paper needs further language editing.

We have thoroughly revised the paper. We hope it's improved significantly.

Reviewer #2 (Public review):

Summary:

In the manuscript "Mapping HIV-1 RNA Structure, Homodimers, Long-Range Interactions and 1 persistent domains by HiCapR" Zhang et al report results from an omics-type approach to mapping RNA crosslinks within the HIV RNA genome under different conditions i.e. in infected cells and in virions. Reportedly, they used a previously published method which, in the present case, was improved for application to RNAs of low abundance.

Their claims include the detection of numerous long-range interactions, some of which differ between cellular and virion RNA. Further claims concern the detection and analysis of homodimers.

Strengths:

(1) The method developed here works with extremely little viral RNA input and allows for the comparison of RNA from infected cells versus virions.

(2) The findings, if validated properly, are certainly interesting to the community.

Thank you for your comprehensive review and insightful comments on our manuscript. We appreciate your recognition of the strengths of our HiCapR method and the potential interest

of our findings to the scientific community.

Weaknesses:

(1) On the communication level, the present version of the manuscript suffers from a number of shortcomings. I may be insufficiently familiar with habits in this community, but for RNA aficionados just a little bit outside of the viral-RNA-X-link community, the original method (reference 22) and the presumed improvement here are far too little explained, namely in something like three lines (98-100). This is not at all conducive to further reading.

Thank you for your feedback on the clarity of our manuscript, particularly regarding the explanation of the HiCapR method and its improvements over the original method mentioned in reference 22

In response to your feedback, we expand on the description of the HiCapR method in the revised manuscript to ensure that it is accessible to a broader audience. We will provide a more thorough comparison between HiCapR and the original method, detailing the specific improvements and how they enable the analysis of low-abundance viral RNAs like HIV. This will include:

Post-library Hybridization: Unlike the original method, HiCapR incorporates a post-library hybridization step. This innovation allows for the capture of target RNA involved in interactions after library construction, offering additional flexibility and enhancing the resolution of the analysis.

Enhanced Sensitivity: HiCapR has been optimized to work with extremely low viral RNA input, which is a significant advancement over the original method. This is crucial for studying viruses like HIV, where obtaining high quantities of viral RNA can be challenging. As a matter of fact,

(2) Experimentally, the manuscript seems to be based on a single biological replicate, so there is strong concern about reproducibility.

Thank you for raising the issue of reproducibility in our study. We understand the importance of experimental replication in ensuring the reliability of our findings. In response to your concern, we would like to provide the following clarification and additional details regarding the reproducibility of our HiCapR experiments:

Replicates in HiCapR Experiments: All ligation and control samples in our HiCapR experiments were performed in three biological replicates. This was done to ensure the high reproducibility of our results. The high degree of correlation ($r > 0.99$) between these replicates underscores the reliability of our findings.

Dimer Validation Experiments: To validate the dimer formation of RRE and 5'-UTR, we employed multiple independent methods, including Native agarose gel electrophoresis, Agilent 4200 TapeStation Capillary electrophoresis, and Biomolecular Binding Kinetics Assays. These methods provide complementary perspectives on the dimer formation, enhancing the robustness of our validation process. The data presented in Figure 3C and Supplementary figure S12 are representative results from these experiments, which consistently support our findings on dimer formation.

Agreement Between Cellular and Virion RNA: Our study also demonstrates a significant similarity between virions in the supernatant and infected cells from the same viral strain, as shown in Supplementary Figure S3. This consistency further validates the reproducibility and

reliability of our HiCapR method in capturing RNA structures and interactions under different conditions.

Consistency across two strains: Our study includes a comprehensive analysis of two distinct HIV-1 strains, NL4-3 and GX2005002, which are prevalent in Europe and Southeast Asia, respectively. The consistency in our findings across these strains serves as a strong indicator of the reproducibility and general applicability of our HiCapR method. Specifically, presence of key structural elements such as TAR, polyA, SL1, SL2, and SL3 in both NL4-3 and GX2005002 strains, suggests a robust structural framework that is conserved across different strains, despite sequence variations. Additionally, our study reveals approximately 20 candidate dimer peaks conserved between the NL4-3 and GX2005002 strains along the genome. The conservation of these dimer peaks across strains indicates a reproducible pattern of dimerization.

(3) The authors perform an extensive computational analysis from a limited number of datasets, which are in thorough need of experimental validation

Thank you for your comment.

In response to your concern, we would like to clarify that while our manuscript does present an extensive computational analysis, we have also conducted a series of experiments. Specifically, we have validated dimer formation using multiple independent methods (afore discussed).

Given the time-consuming nature of additional experiments, we have chosen to share the HiCapR data with the community in a timely manner. This approach allows for more immediate communication and evaluation of the data on HIV structure, which we believe is valuable for advancing the field.

We are committed to further investigating the functional implications of our structural findings. We plan to conduct more experiments to explore the functional linking between the structural insights of HIV, which will help to deepen our understanding of the virus's replication and potential antiviral strategies.

Recommendations for the authors:

Reviewer #1 (Recommendations for the authors):

I suggest a major revision of the manuscript.

Minor comments:

(1) The article lacks consistency in its presentation. The expression of the proper noun is wrong in the paper. For example, (a) L89, "RNA:RNA interaction" → RNA-RNA interaction; (b) L431, "SARS-COV-2" → SARS-CoV-2;

We are sorry for the inconsistency. We have corrected the mistakes.

(2) "We identified dimers based on the methodology described in23." is not a complete sentence.

Thank you for your insightful comment. We have revised the sentence to provide a complete and clear description of our methodology. The revised sentence is as follows: "Homodimers were identified in accordance with the methods previously reported in the literature."

Reviewer #2 (Recommendations for the authors):

(1) The authors perform an extensive computational analysis from a limited number of datasets, which are in thorough need of experimental validation. There is a single series on in vitro validation of the interaction of an homodimerization site, described in five lines (278-283) plus the Figure panel 3c with a very brief legend, and an extremely minimalist Figure S12. The panel to Figure 3c contains Kd values which have not been assessed for significant digits.

Thank you for your constructive feedback on our manuscript.

We acknowledge that our computational analysis is based on a limited number of datasets. Due to the initial exploratory nature of our study and the logistical challenges of generating additional datasets, we have focused on in-depth analysis of the available data. We are currently working on further validating our findings and are committed to publishing these results in a follow-up study.

Regarding Experimental Validation:

We agree that the initial description of our in vitro validation of the homodimerization site was concise. To address this, we have expanded the description of our experimental procedures. Specifically, we have detailed the methods used for the in vitro transcription, the preparation of RNA samples, and the use of the Octet R8 platform for biomolecular binding kinetics assays.

For the Kd values presented in Figure 3c. We have now included standard error of the mean and have defined the significant digits in the figure legend. This revision provides a more accurate representation of the binding affinities.

(2) As a further example to be experimentally validated, splice sites are discussed after lines 354, for which unsophisticated validation techniques such as targeted RT-PCR are widely accepted.

In response to your comment, we would like to clarify that the splice sites mentioned in our study are well-established and widely recognized in the literature. They have been previously characterized and are considered canonical within the HIV research community. Given their established nature, we have relied on this foundational knowledge in our analysis.

However, we concur with the importance of validating the regulatory role of homodimers in splicing, which is a significant aspect of HIV biology. While we have provided evidence for the presence of these homodimers and their potential implications for splicing, we acknowledge the need for further functional studies to elucidate their mechanistic role.

Due to the scope and length constraints of the current manuscript, we have chosen to focus on the structural and interaction analyses provided by HiCapR. The functional validation of these homodimers and their impact on splicing will be pursued in subsequent studies, which we plan to initiate promptly. We believe that a dedicated follow-up study will allow for a more in-depth exploration of this complex and important aspect of HIV gene regulation.

We are committed to advancing our understanding of the functional significance of these homodimers in the context of HIV splicing and will ensure that this line of investigation is thoroughly addressed in our future work.

Thank you again for your valuable feedback. We look forward to contributing further to the field with our ongoing research.

<https://doi.org/10.7554/eLife.102550.2.sa0>