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

Yan Zhang[1#], Jingwan Han[2#], Dejian Xie[1#], Wenlong Shen[1], Ping Li[1], Jian You Lau[3], Jingyun Li[2*], Lin Li[2*], Grzegorz Kudla[3*], and Zhihu Zhao[1*]

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

2 Institute of Microbiology and Epidemiology, Beijing, China

3 MRC Human Genetics Unit, University of Edinburgh, Edinburgh EH4 2XU, United Kingdom

* To whom correspondence should be addressed: Zhihu Zhao, Email: zhaozh@bmi.ac.cn.

Correspondence may also be addressed to Grzegorz Kudla Email: gkudla@gmail.com, Jingyun Li: lijyjk@163.com or Lin Li: dearwood@sina.com;

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

Keywords: HIV-1, RNA structure, Viral Genome, HiCapR, Genomic Interactions

Competing interests: The authors declare no competing interests.

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 encluding homodimer sites are persistent in virions and are involved in dimerization.

Abstract

Human Immunodeficiency Virus (HIV) persists as a leading global health issue. A significant knowledge gap exists in our understanding of long-range interactions of 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'-UTR and 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

47 within virions that are instrumental in the dimerization process. These domains are
48 preserved throughout the packaging process.

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

52 ## Main

53 Human Immunodeficiency Virus (HIV) remains a significant global health concern
54 despite advancements in antiretroviral therapy even with the newest report of total
55 7 AIDS patients been completely "cued" [1] [2,3]. Research on HIV-1 RNA structure and
56 function has intensified, focusing on assembly, release, and maturation processes,
57 using NMR[4], cryo-EM[5] and chemical probing[6-8]. However, our knowledge of the
58 global architecture and the long-range interactions within the HIV-1 RNA genome
59 remains limited, primarily due to the complexity and dynamic nature of these
60 structures.

61 The HIV genome, comprising two copies of positive-sense RNA, has been extensively
62 studied to understand its regulatory elements, such as the 5'-UTR, crucial for
63 dimerization and selective packaging into viral particles[9]. Additionally, interactions
64 involving the nucleocapsid protein (NC) and the dimerization signal (Ψ) play pivotal
65 roles in the assembly and maturation of HIV virions[10]. Furthermore, the Rev
66 Response Element (RRE) in the HIV genome, responsible for nucleocytoplasmic
67 transport of viral RNA, exhibits structural variability impacting virus replication
68 rates[11,12]. While current techniques like SHAPE provide insights into local RNA
69 structures[7,8,13], studying long-range RNA interactions within HIV under
70 physiological conditions remains a challenge.

71 Proximity ligation-based methods have been shown to be particularly useful for
72 studying long-range RNA interactions, as they can overcome the RNA length
73 limitations of traditional techniques. These methods have been applied to the study
74 of various viruses, including influenza[14,15], Zika[16-18], and SARS-CoV-2[19-22], and have
75 led to the identification of numerous RNA structures that are closely related to the
76 virus lifecycle. We and others have recently shown that RNA proximity ligation data

77 also contains information about RNA homodimerization events[23,24]. Nevertheless, as
78 of now, there are no proximity ligation studies of long-range interactions or dimeric
79 interactions of the HIV RNA genome.

80 In this study, we modified our previous protocol[22] by integrating a hybridization-
81 based capture of HIV-1 sequence after library construction, resulting in a method
82 we call High throughput Capture of RNA interactions (HiCapR). This advancement
83 enabled the comprehensive capture and analysis of the complete HIV RNA genome.
84 By applying HiCapR to both infected cells and virions, we uncover a distinct genomic
85 compression pattern in virions, highlighting the critical role of global genome
86 folding in HIV packaging and assembly. Our findings not only reveal the presence of
87 persistent genomic domains within virions that facilitate whole genome
88 dimerization, even the transgenerational inheritance but also a significant reduction
89 of long-range RNA:RNA interactions in virions compared to infected cells. These
90 insights provide a new perspective on the functional implications of HIV RNA
91 structure in splicing and packaging processes and may lead to the identification of
92 new targets for antiviral intervention.

93 ## results

94 ### 1. Capturing in vivo HIV genome RNA structure in infected cells and in virion

95 In a previous study, we characterized the genome structure of SARS-CoV-2 [22], which
96 was sufficiently abundant in cells, obviating the need for enrichment. To effectively
97 capture and analyze full-length low abundant HIV RNA-RNA interactions in infected
98 cells, we developed an improved version of the protocol, which we call **Hi**gh
99 throughput **Cap**ture of **R**NA interactions (**HiCapR**). The principle of hybridization-
100 based capture incorporated NGS library construction is well established and widely
101 employed in capture Hi-C methods[25-28] and in RNA proximity ligation [17,19]. Applying
102 this strategy, we compared the NL4-3 and GX2005002 strains [29], which represent
103 the prevalent strains in Europe and Southeast Asia, respectively. The overall design
104 of this study is presented in Figure 1A, and the detailed methods are provided in the
105 methods section. Supplementary Figure S1 illustrates the bioinformatic pipeline

106 used to detect RNA-RNA interactions as well as homodimers, which are discussed
107 later in this paper.

108 Approximately 16 million raw reads were obtained per sequencing library. Chimeric
109 reads were identified using the hyb technique as previously described[22], with a
110 database comprising human spliced mRNAs, noncoding RNAs[30] and the HIV genome
111 sequence as reference (see Supplementary Table 1). The alignment rate exceeded
112 90%, and over 95% of the reads contained HIV sequences, indicating an efficient
113 capture of HIV RNA transcripts. On average, we obtained reads with depths of
114 coverage of approximately 11,500x across the genome (Supplementary table 1). The
115 uniform distribution of reads across the HIV genome post-capture (Supplementary
116 Figures S2) indicating HiCapR's effectiveness in capturing the entire HIV RNA
117 genome with minimal bias.

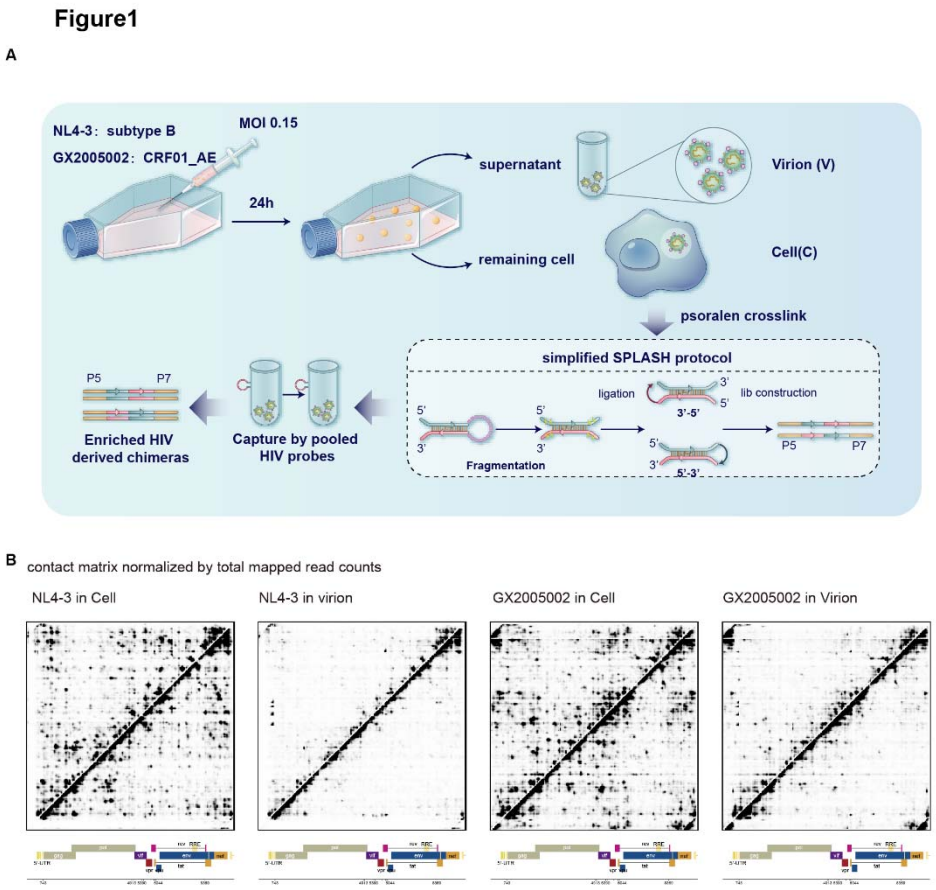
118 The overall duplication rate remained below 50%, with the lower duplication rate in
119 the ligation group compared to the non-ligation group (supplementary table 1),
120 possibly due to the increased library complexity resulting from RNA proximity
121 ligation.

122 In proximity ligation assays, the chimeric rate serves as a crucial indicator of ligation
123 efficiency. We observed an approximate 1% chimeric rate in the non-ligation
124 controls, and around 9% of chimeric reads in the ligation group, aligning with the
125 widely reported range in the literature. The proportion of 5'-3' and 3'-5' chimeras
126 was nearly equal in long-range (>1kb) interactions, whereas proximal (<100-nt)
127 interactions were enriched for 3'-5' chimeras, resulting from the similarity of
128 proximal 5'-3' chimeras to nonchimeric sequencing reads, in agreement with
129 previous studies[19,22] (supplementary table 1).

130 We performed a Pearson correlation analysis and clustering of contact read counts
131 from various samples, including ligated (+L) and non-ligated control samples (-L).
132 Our results revealed a high degree of correlation (r>0.99) between biological
133 replicates, indicating the high reproducibility. Additionally, we observed a
134 significant similarity between virions in the supernatant and infected cells from the
135 same viral strain (Supplementary Figure S3). Using a similar approach as described

136 in our previous paper, we generated a contact matrix for the global HIV-1 genome,
137 revealing clear local and long-range interactions (Figure 1B and Supplementary
138 Figure S4).

139 In summary, HiCapR demonstrated high reproducibility, low bias, reliability and
140 efficiency in capturing HIV transcripts. Therefore, it can be applied as a robust tool
141 for analyzing the RNA structure of the HIV genome, and perhaps the other similar
142 RNA virus.



143

144 **Figure 1 Experimental design of HiCapR and overall HIV-1 contact matrix**

145 A) Experimental design of HiCapR for profiling RNA-RNA interactions and
146 dynamics of the HIV genome. Based on the simplified SPLASH protocol, HiCapR
147 incorporates a probe-based hybridization method to enrich HIV RNA chimeras
148 from the SPLASH library. The HiCapR method has been applied to two strains of

149 HIV, NL4-3 and GX2005002, in both infected cells and their corresponding virus
150 particles.

151 B) Comprehensive contact matrix derived from infected cells and virions of both
152 NL4-3 and GX2005002 HIV-1 strains in the HiCapR experimental groups. The
153 heatmap displays the average count of chimeras per 1 million mapped reads,
154 combining data from two biological replicates for each sample.

155

156 ## 2. Local RNA structure heterogeneity, dynamics and robustness in HIV 5'-UTR

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

160 Chimeras analysis support the presence of key structural elements, including TAR,
161 polyA, SL1, SL2, and SL3, as well as polyA-SL1 in the monomeric conformation in the
162 HIV genome (Figure 2A and Supplementary Figure S5). Despite the pivotal role of
163 the 5'-UTR in replication, packaging, transcription and translation, we noted that the
164 5'-UTR sequence of HIV is not highly conserved. A comparison between the NL4-3
165 and GX2005002 strains revealed notable insertions and deletions in the U5 region,
166 along with point mutations in the loop of the SL1 core region, which is critical for
167 HIV dimerization(Figure 2B). These findings raise questions about the conservation
168 of the structural integrity of this region. We therefore applied the comradesFold
169 algorithm [17] to our HiCapR data to generate 1000 structure predictions of the 5'-
170 UTR (extending 100 nt downstream of the AUG start codon) for each sample, and
171 then performed MDS analysis after aligning the viral genome coordinates of both
172 strains (Figure 2C). This analysis revealed that the reported dimer and monomer
173 structures clustered together in NL4-3 but form distinct clusters in
174 GX2005002(Figure 2C). Based on these folded structures, we calculated the base
175 pairing probability for each base pair, which involves determining the number of
176 folded structures supporting a specific base pair divided by the total number of
177 structures. Visualizing this base pairing probability as a heatmap identifies the most
178 stable base pairs in the 5' UTR of HIV. We observed a consistent presence of key

179 structural elements such as polyA, TAR, SL1, SL2, and SL3 in both NL4-3 and
180 GX2005002 strains (Figure 2D), suggesting robustness in the overall structure
181 despite sequence variations and alternative RNA conformations.

182

Figure2

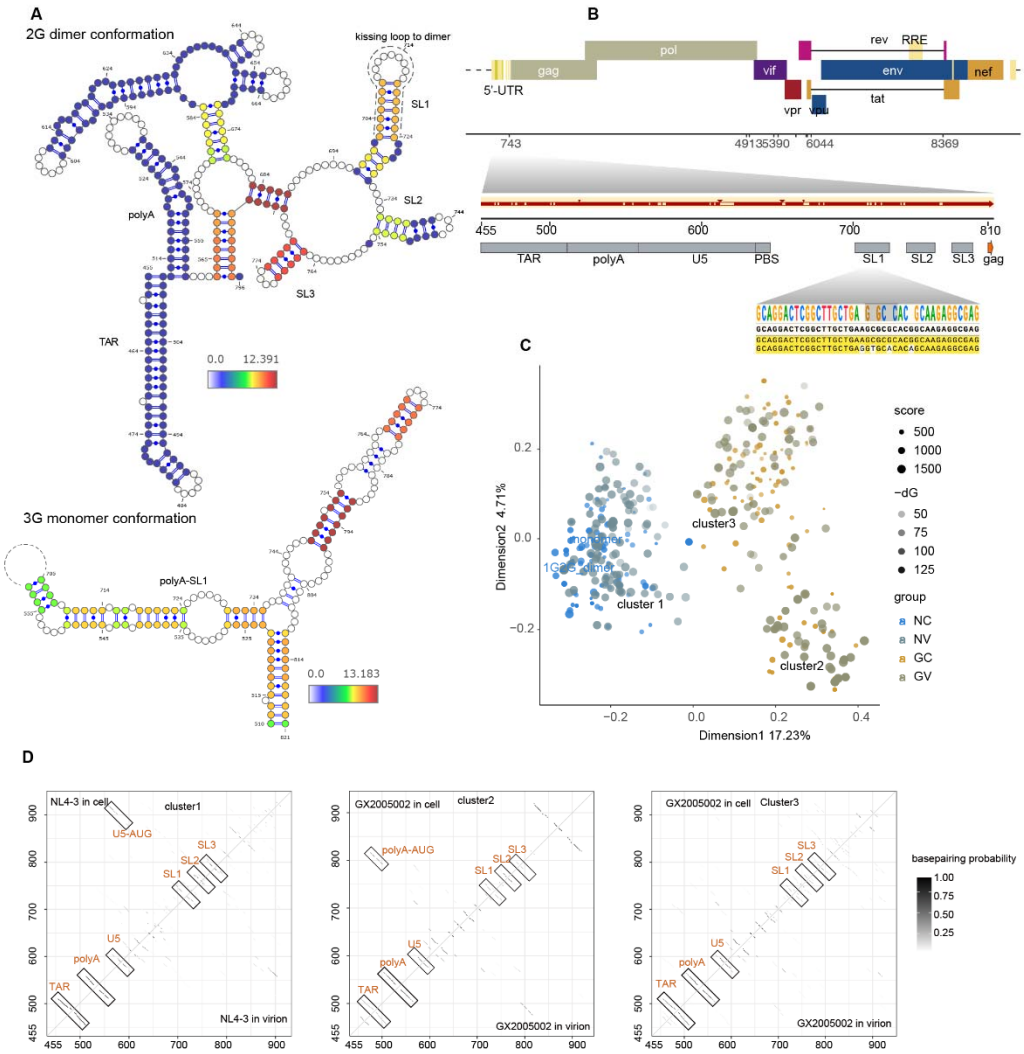


Figure 2 conformations of HIV-1 5'-UTR

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

192 GX2005002 compared to NL4-3. Small triangle arrows indicate insertions in the
193 GX2005002 5'-UTR. At the top, a seqlogo displays the consensus nucleotides in
194 the SL1 region.

195 C) Multidimensional scaling (MDS) plot clustering each of the 1,000
196 computationally predicted structures of the 5'-UTR in two strains under two
197 conditions.

198 D) Basepairing probability matrices were calculated from 1000 computationally
199 predicted structures, with color indicating the percentage of structures
200 supporting the specific basepair.

201

202

203 ## 3. Novel homodimerization sites and their implications in RNA splicing and viral
204 assembly

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

216 The strongest homodimerization signal was found in the 5'-UTR region, which
217 aligns with previous studies. The 5'-UTR region is well-known for its role in
218 triggering HIV dimerization. Previous literature has highlighted the importance of
219 SL1 (DIS) and SL3 (Ψ) in 5'-UTR[8,31,32]. Our data shows that the SL1, SL2, and SL3
220 regions all have supportive dimeric chimeras (Supplementary Figure S7). We

221 derived a Dimerization Score by calculating the reads that support base pairing of
222 homodimeric chimeras (Supplementary Figure S8), analogous to the COMRADES
223 Score described previously [17,19,22]. The base pairing of homodimers in the SL1 region
224 of the HIV genome is consistent with previous data from NMR and cyro-EM studies
225 [33,34]. Additionally, we observed that variations in nucleotides within the SL1 region
226 do not alter the base pairing pattern of dimers between the NL4-3 and GX2005002
227 strains (Figure 3B) which is egret with previous structure robustness revealed by
228 permutation and structure prediction results

229 Strikingly, in addition to the known homodimer within the SL1 and SL3 regions, we
230 observed homodimerization distributed along the entire length of the HIV genome
231 (Supplementary Figure 9). Homodimers were present in both infected cells and
232 virions of both strains, with approximately 20 peaks of dimerization that are
233 conserved between the NL4-3 and GX2005002 strains along the genome
234 (Suplementary table2).

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

245 The identification of multiple dimers in HIV RNA beyond the 5'-UTR region raises
246 intriguing questions regarding their potential roles serving within the HIV genome.
247 To this end, we analyzed the sequence within dimer peaks, identifying a prevalent
248 AG-rich motif rich, present in nearly every peak. This motif highly? resembles the
249 RNA binding motif of Serine/Arginine-Rich Splicing Factor (SRSF) (Supplementary
250 Figure S9B,S9C and S10), which are essential RNA-binding proteins involved in pre-
251 mRNA splicing and alternative splicing regulation[35], serving as the main type of HIV-

252 1 splicing factors[36]. This result suggests a potential role of dimerization in RNA
253 splicing processes within the HIV genome.

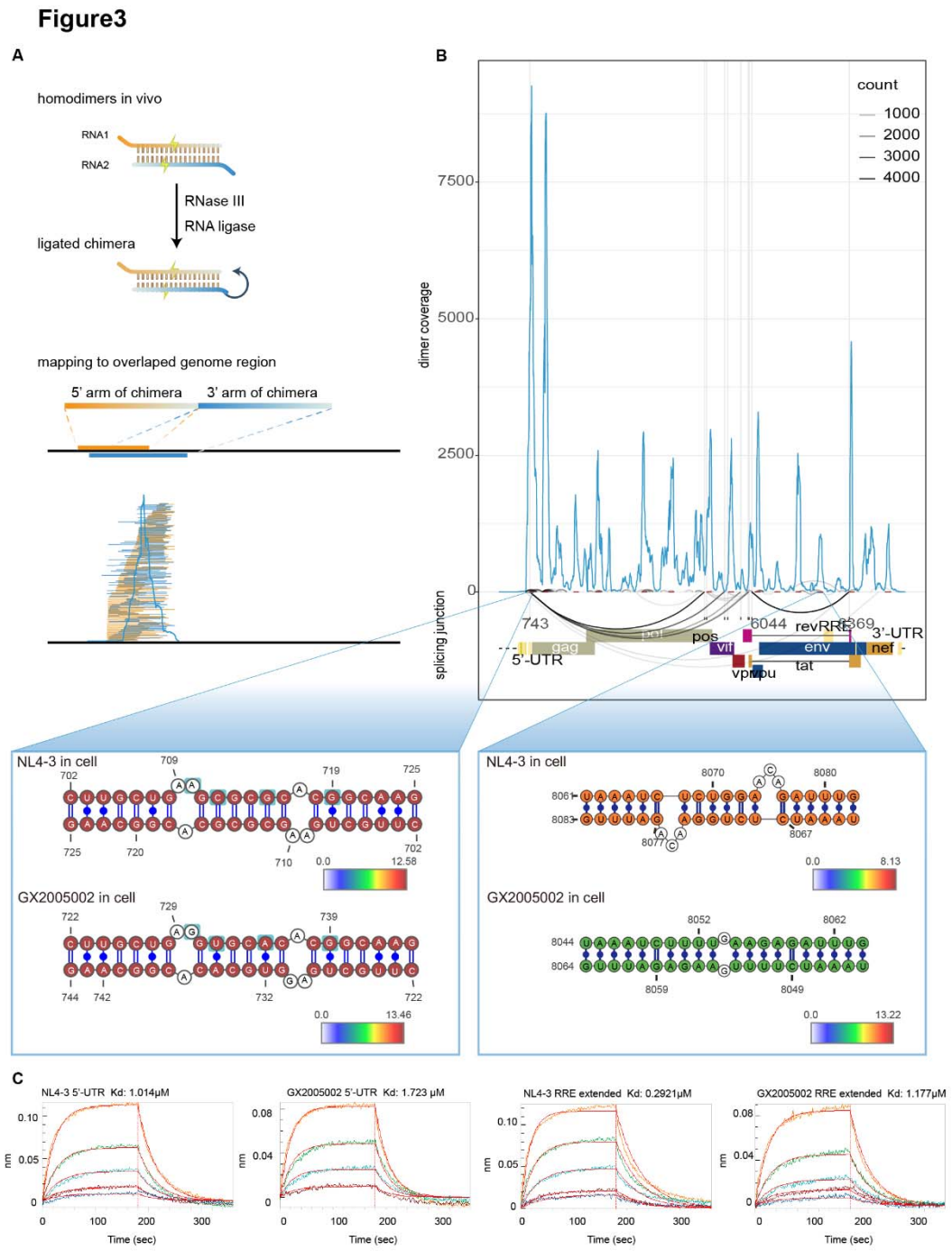
254 5'-3' discontinuous reads were identified in non-ligation control data, enabling the
255 inference of splicing junction sites. These inferred sites showed strong concordance
256 with canonical splicing sites and recent nanopore sequencing data[37] (Figure 3B).
257 Notably, almost every junction site, including splicing donor and acceptor sites,
258 exhibited dimer peaks around them. Additionally, the majority of dimeric chimeras
259 covered these junction sites and enriched downstream of acceptor sites
260 (Supplementary Figure S10), suggesting that dimerization process/events takes
261 place on unspliced genomic RNA. Interestingly, dimers surrounding splicing
262 regulatory elements (as summarized in[36]) form stable base pairing supported by
263 numerous chimeras (Supplementary Figure S10). This observation underscores the
264 close proximity and potential functional relationship between dimerization events
265 and RNA splicing processes at these critical sites within the HIV genome, suggesting
266 a potential role of dimerization in splicing processes or assisting in the transport of
267 unspliced RNA out of the nucleus.

268 According to a previous report that utilized CLIP-seq to analyze the RNA binding
269 specificity of the Gag protein in both cells and virions/From a previous CLIP-seq
270 analyzing of Gag protein's RNA binding specificity, it was observed that Gag exhibits
271 a preference for binding to multiple elements, such as psi and RRE, in HIV infected
272 cells[38]. Interestingly, significant dimer peaks were also identified in these regions
273 (Supplementary Figure S9A). This suggests a potential connection between Gag
274 binding and dimerization of the HIV genome.By plotting contact matrices derived
275 from non-overlapping and dimeric chimeras around the RRE region, we were able to
276 identify two distinct dimer peaks flanking the RRE region (Supplementary Figure
277 S11A).

278 To validate these novel homodimers, we synthesized extended RRE RNA fragments
279 with two dimer sites using in vitro transcription and confirmed their ability to form
280 a dimeric conformation through annealing and non-denaturing gel electrophoresis
281 (Supplementary Figure S12A), as well as Agilent Tapestation 4200 capillary

282 electrophoresis (Supplementary Figure S12B) and Bio-Layer Interferometry (BLI)
283 technology (Figure 3C).

284



285

286

287 **Figure 3 Identification and validation of homodimers in HIV-1 genome**

288 A) Visual depiction of the RNA homodimer formation process: Inter-ligated
289 fragments (arms) originating from homodimeric RNA molecules generate
290 chimeras, where each arm aligns to overlapping coordinates on the HIV-1
291 genome. This process enables the plotting of coverage and specific details of
292 dimers by utilizing the positions and counts of these overlapping chimeras.

293 B) distribution of homodimers throughout the HIV-1 genome. The blue line plot
294 showcases the homodimer coverage derived from ligated samples of NL4-3
295 infected cells along the HIV-1 genome. Arc plots exhibit discontinuous reads in
296 non-ligated samples of NL4-3 infected cells, with dark red segments indicating
297 peaks of homodimers. The lower panels depict the base pairing of homodimers
298 in the SL1 region of the 5'-UTR and downstream of the RRE region. Color
299 mapping indicates the log2 transformed dimer score.

300 C) Evaluation of the self-binding of dimers using Bio-Layer Interferometry (BLI).
301 The data presented is the result of analysis from three independent experiments,
302 providing insights into the interaction and binding behavior of the dimers.

303

304 **4. Interplay between dimeric peaks and 3D genome organization**

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

315 suggesting the involvement of stable local structures in the formation of dimers
316 between two copies of the HIV genomes.

317 We observed the meta matrices around dimer sites are similar to genomic domains,
318 which have been previously described in SARS-COV-2[22]. This observation prompted
319 us to investigate the potential existence of genomic domains in the HIV RNA
320 genome. Using similar methods[39], we calculated insulation scores for two strains of
321 the virus in both cells and virions, as depicted in Figure 4A.

322 In this way, we identified 31, 33 HIV genomic domains in infected cells and 36, 32
323 domains in virions for NL4-3 and GX2005002 strains, respectively. The high
324 correlation of insulation scores between cells and virions and consistent domain
325 boundaries are observed (Figure 4B), and more importantly, the boundary
326 strengths did not show any significant differences between cells and virions (Figure
327 4C), suggesting conserved, stable and persistent genomic domains during the
328 analyzed stages.

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

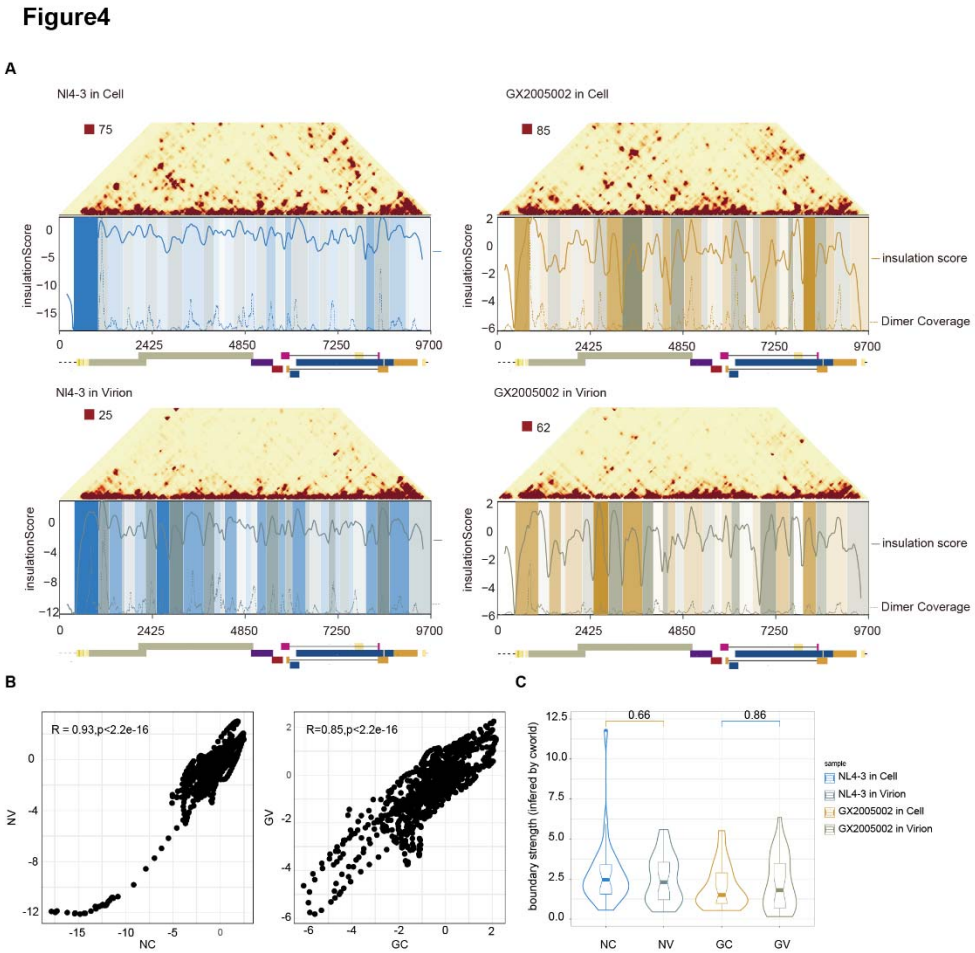


Figure 4 genome domains along HIV-1 genome

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.

346

347 ## 5. Dynamic changes in global and long-range interactions throughout HIV-1 life cycle.

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

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

359 Specifically, we employed viewpoint analysis to examine interactions involving the
360 5'-UTR. The results, depicted in Figure 5B, reveal multiple peaks of 5'-UTR
361 interactions across the HIV-1 genome. Interestingly, for the majority of these peaks,
362 the signals are lower in virions compared to cells. However, cyclization interactions
363 (5'-UTR-3'-UTR) at the ends are still maintained in virions (Figure 5C). Intriguingly,
364 we also identified multiple interaction peaks located near crucial elements, such as
365 the frameshift element (Supplementary Figure S16A), rev start codon
366 (Supplementary Figure S16B), RRE, and nuclear exporting signal (NES)
367 (Supplementary Figure S16C), indicating complex roles of 5'-UTR in HIV life cycle.
368 Among these interactions, some are strain specific, such as the interaction between
369 5'-UTR and FSE is specific to the NL4-3 strain, while interaction between 5'-UTR and
370 8.5K region (near NES) are highly enriched in GX2005002 strain (Supplementary
371 Figure S16C). As mentioned earlier, the majority of interactions decrease in the
372 virion, with the notable exception of interactions in the 5'-UTR and 6K region (near
373 D4 splicing donor site), which significantly increase in the virion. (Supplementary
374 Figure S16B). The enhanced interactions in virions are consistent across both
375 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.

Figure5

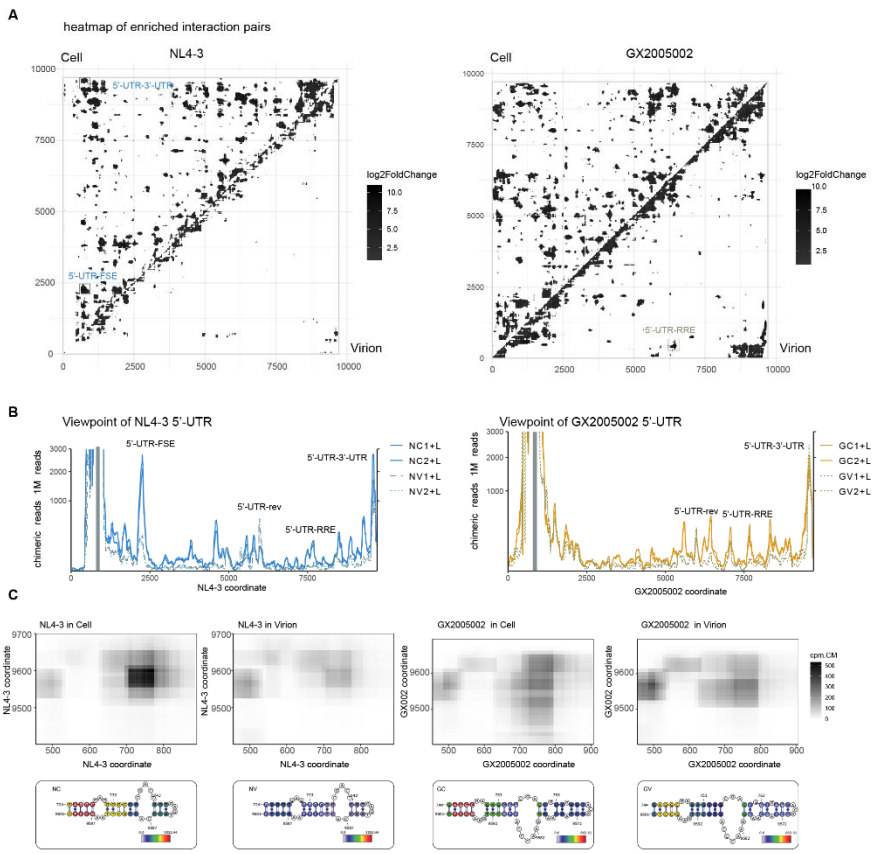


Figure 5 long range interaction in HIV-1 genome

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.

discussion

1. 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-C[40-43], 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.

417 ## 2. New insight of HIV RNA local and long-range interactions

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

423 In addition to its local folding, the 5'-UTR engages in extensive and dynamic long-
424 range interactions, notably with the 3'-UTR, suggesting genome cyclization akin to
425 other viruses (Figure 4). This phenomenon, also observed in Zika, influenza, and
426 SARS-COV-2[15,17,19,22], hints at broader biological implications warranting further
427 exploration. Additionally, the 5'-UTR interacts with key import elements in the HIV
428 genome, such as NES, RRE, and the rev start codon. Given the Rev protein's roles in
429 RNA transport and nuclear export[44], the 5'-UTR may also contribute to these
430 processes.

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

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

444 The contrasting loss of long-range interactions in the HIV-1 genome compared to
445 SARS-CoV-2 [22] suggests distinct folding processes between these viruses. This
446 disparity underscores the unique characteristics of each virus and their genomic

447 structures. Speculations indicate that the HIV-1 genome may adopt a rod-like
448 structure within viral particles, unlike the spherical compression seen in SARS-COV-
449 2 [46]. Further investigation is needed to unravel the mechanisms and biological
450 significance of this compression in the HIV-1 genome.

451 ### 3. Emerging Insights into HIV-1 Dimerization through newly identified dimer sites

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

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

475 Interestingly, the observation that candidate dimer sites are located within genomic
476 domains suggests that dimerization may be influenced by the local RNA folding

477 environment (Supplementary Figure S12). It is possible that the dense local
478 interaction around these sites may facilitate dimerization by bringing the two
479 regions of the genome into close proximity. Alternatively, dimerization may play a
480 role in shaping the local chromatin environment by promoting even initiating the
481 formation of genomic domains. Further research is needed to fully understand the
482 relationship between RNA genomic domains and dimer sites in the HIV-1 genome.
483 However, the identification of these candidate dimer sites within genomic domains
484 provides a starting point for investigating the role of local genome RNA interactions
485 in HIV-1 replication and dimerization. These findings may have implications for the
486 development of novel antiviral therapies that target the dimerization process and
487 may provide new avenues for future research in this field.

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

494 methods

495 Cell culture and virus infection

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

506 ## HicapR method

507 First, extracted crosslinked RNA was treated as in simplified SPLASH protocol [22].

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

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

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

526

527 ## Data preprocessing and chimeric reads identification

528 The data preprocessing and identification of chimeric reads were conducted as
529 previously described. The reference genome used for NL4-3 strain was a
530 combination of NL4-3 genome sequence
531 (https://www.ncbi.nlm.nih.gov/nuccore/AF003887) and human spliced mRNAs
532 and noncoding RNAs described in [30], while for GX2005002 strain, the reference
533 genome used was a combination of GX2005002 genome sequence

534 (https://www.ncbi.nlm.nih.gov/nuccore/KP178420) and human spliced mRNAs
535 and noncoding RNAs as above.

536 First, we used pear to merge the overlapped reads:

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

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

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

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

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

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

544 **Chimeras for RNA-RNA interaction and dimer identification.**

545 We identified dimers based on the methodology described in[23]. Specifically, we
546 utilized the hub pipeline and filtered chimeras in .hyb files. To calculate the degree
547 of overlap between the two arms of chimeric reads, we used the formula
548 L=1+min(e1,e2)-max(s1,s2), where e1 and e2 represent the ends of each arm of
549 chimeric reads, while s1 and s2 represent their starts. To ensure that possible
550 circular RNAs, which may be produced by single RNA end-to-end ligation, are
551 filtered out, we applied the following condition: (e1 < e2) OR (s1 < s2). A schematic
552 diagram of dimer chimeras is provided in Supplementary Figure S1.

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

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

559 ### Dimer score and dimer base pairing

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

565 ### Local structure folding and MDS plot

566 Initial local structure folding was performed similarly to a previous report[17]. Base
567 pairing scores were first calculated using comradesMakeConstraints function in
568 COMRADES package(https://github.com/gkudla/comrades):

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

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

574 ### Calculation of base pairing probability for local structure

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

579 ### Dimer validation

580 **RNA preparation.** Purified NL4-3 and GX2005002 5'-UTR and extended RRE PCR
581 products (200 ng) were utilized as templates for RNA in vitro transcription with T7
582 RNA polymerase (Vazyme). The reaction was then incubated at 37 °C for 16 hours,
583 followed by DNase I treatment for 30 minutes at 37 °C. The RNA was subsequently
584 purified using MagicPure RNA Beads (TransGen) through gel purification.

585 **Native agarose gel electrophoresis.** RNA (600 ng) was heated to 95°C and then
586 slowly cooled to room temperature in either high salt buffer (50mM Tris-HCl pH 7.5,

587 140mM KCl, 10mM NaCl) or low salt buffer (high salt buffer diluted 1:10 with
588 water). Samples were loaded with native loading dye (0.17% Bromophenol Blue and
589 40% (vol/vol) sucrose) on 2% agarose gel prepared with 1× tris-borate magnesium
590 (TBM) buffer (89 mM Tris base, 89 mM boric acid and 2 mM MgCl2) and
591 fractionated at 100 V for 85 min at room temperature.

592 **Agilent 4200 TapeStation Capillary electrophoresis**. We prepared the RNA
593 samples as described above and loaded them onto the TapeStation RNA screentape
594 without the heating and denaturing step. Subsequently, we initiated the device for
595 electrophoresis analysis, which ran automatically without the need for manual
596 intervention. After the electrophoresis analysis was completed, the system
597 generated electropherograms and relevant data regarding the RNA samples,
598 including their size distribution and concentration. This method allowed us to
599 quickly and accurately evaluate the conformational features of the RNA.

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

607 Data availability

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

612 Code availability

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

Acknowledgements

This work was funded by a National Key Technologies R&D Program (2018YFA0900801) awarded to Y.Z

The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Supplementary Figure legends

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

642 suggests that the HiCapR method is able to capture the entire HIV genome with low
643 bias.

644 ## Supplementary Figure S3

645 The heatmap displays the correlation between samples, which was measured using
646 Pearson's coefficients.

647 ## Supplementary Figure S4

648 Related to Figure 1B, but with the inclusion of information from individual
649 replicates and data from non-ligation controls, offering a more detailed view of the
650 interactions and distributions presented in Figure 1B.

651 ## Supplementary Figure S5

652 The established conformation of the HIV-1 5'-UTR, supported by chimeras in NL4-3
653 and GX2005002, both in cells and in virions. Colors of the nucleotides indicate the
654 log2 transformed base pairing scores.

655 ## Supplementary Figure S6

656 The stacked bar chart displays the number of 3'-5', 5'-3', and dimer chimeras in
657 various ranges across the sample groups. The panel on the left pertains to NL4-3,
658 while the one on the right pertains to GX2005002.

659 ## Supplementary Figure S7

660 (A) Heatmaps simultaneously plot contact matrices calculated from non-overlapped
661 chimeras, which is informative for interactions and dimer chimeras which is
662 derived for homodimers. The coordination in heatplota is RNA-based, with the
663 numbers in brackets indicating the corresponding coordinates in the DNA
664 genome.

665 (B) Homodimer chimeras details around splicing sites show the 5'-arm represented
666 by blue lines and the 3'-arm represented in yellow, providing a visual
667 representation of the interactions in these regions.

668 ## Supplementary Figure S8

669 Dimer score matrices around 5'-UTR of NL4-3 and GX2005002 strain in different
670 stages.

671 ## Supplementary Figure S9

672 (A) Homodimer coverage along HIV-1 genome in NL 4-3 and GX2005002 both in
673 cells and in virion samples. Gag-CLIP data from Kutluay, S. B., et al. (2014) are
674 also shown. Bottom panel depict delta G of each100-nt window in NL4-3 genome.

675 (B) Distribution of enriched motif in homodimer peaks of NL4-3, region similar to
676 SRSF motif was underlined in the seqlog.

677 (C) Distribution of enriched motif in homodimer peaks of GX2005002, region similar
678 to SRSF motif was underlined in the seqlog.

679 ## Supplementary Figure S10

680 Top: Homodimer chimeras details around splicing sites show the 5'-arm
681 represented by blue lines and the 3'-arm represented in yellow, providing a visual
682 representation of the interactions in these regions. Known splicing regulatory
683 elements are also indicated.

684 Bottom: basepairing of homodimers around splicing sites as indicated. Color code
685 indicates log2 transformed base pairing score supporting each base-pair.

686 ## Supplementary Figure S11

687 (A) Contact matrix and homodimeric chimeras surrounding the RRE region in the
688 HIV genome. The left panel displays the contact matrix derived from non-
689 overlapping chimeras and the contact matrix derived from dimeric chimeras.
690 The right panel presents specific details of the dimeric chimeras around the RRE
691 elements.

692 (B) Base pairing of homodimers in the extended RRE region in specific samples.
693 Color mapping indicates the log2 transformed dimer score.

694 ## Supplementary Figure S12

695 (A) Native agarose gel confirming dimeric conformation of RRE (extended to 620nt)

696 (B) Agilent tapestation 4200 capillary electrophoresis confirming dimeric
697 conformation of 5'-UTR and RRE (extended to 620nt)

698 ## Supplementary Figure S13

699 Predicted structure of extended RRE. Color code indicates log2 transformed base
700 pairing score supporting each base-pair.

701 ## Supplementary Figure S14

702 Average contact matrices around peak centers. Heatmaps showing the normalized
703 average interaction frequencies for all peak centers as well as their flanking regions
704 (±0.5 peak width). The heatmaps were binned at 10 nt resolution.

705 ## Supplementary Figure S15

706 A) histograms show distributions of different types of interactions across the HIV
707 genome. The x-axis represents the span of the interactions, while the y-axis
708 represents the frequency of each interaction type.

709 B) Contact probability decay curves (Ps curves) showing progressive
710 reconfiguration HIV genomes. Both two strains of the virus exhibit faster decay
711 rates within virions than cells.

712 ## Supplementary Figure S16

713 Contact matrices and base pairing details between 5'-UTR and 2.2K (A), 6K (B) and
714 7.6~8.5K (C).

715 ## Supplementary data

716 **Supplementary table1**

717 Mapping statistics

718 **Supplementary table2**

719 Homodimer peaks information

720 **Supplementary data**

721 Designed probes for enriching HIV chimeras in fasta format

722

723 ## Refferences

724

725

726 1 Mallapaty, S. Seventh patient 'cured'of HIV: why scientists are excited. *Nature*
727 **632**, 235-236 (2024).
728 2 Freed, E. O. HIV-1 assembly, release and maturation. *Nat Rev Microbiol* **13**,
729 484-496 (2015). https://doi.org/10.1038/nrmicro3490
730 3 Emery, A. & Swanstrom, R. HIV-1: To Splice or Not to Splice, That Is the
731 Question. *Viruses* **13**, 181 (2021).
732 4 Keane, S. C. *et al.* RNA structure. Structure of the HIV-1 RNA packaging signal.
733 *Science* **348**, 917-921 (2015). https://doi.org/10.1126/science.aaa9266
734 5 Zhang, K. *et al.* Structure of the 30 kDa HIV-1 RNA dimerization signal by a
735 hybrid Cryo-EM, NMR, and molecular dynamics approach. *Structure* **26**, 490-
736 498. e493 (2018).
737 6 Watts, J. M. *et al.* Architecture and secondary structure of an entire HIV-1
738 RNA genome. *Nature* **460**, 711-716 (2009).
739 7 Tomezsko, P. J. *et al.* Determination of RNA structural diversity and its role in
740 HIV-1 RNA splicing. *Nature* **582**, 438-442 (2020).
741 https://doi.org/10.1038/s41586-020-2253-5
742 8 Ye, L. *et al.* Short- and long-range interactions in the HIV-1 5' UTR regulate
743 genome dimerization and packaging. *Nat Struct Mol Biol* **29**, 306-319 (2022).
744 https://doi.org/10.1038/s41594-022-00746-2
745 9 Ding, P. *et al.* 5'-Cap sequestration is an essential determinant of HIV-1
746 genome packaging. *Proc Natl Acad Sci U S A* **118** (2021).
747 https://doi.org/10.1073/pnas.2112475118
748 10 Chen, J. *et al.* Visualizing the translation and packaging of HIV-1 full-length
749 RNA. *Proceedings of the National Academy of Sciences* **117**, 6145-6155
750 (2020).
751 11 Sherpa, C., Rausch, J. W., Le Grice, S. F., Hammarskjold, M. L. & Rekosh, D. The
752 HIV-1 Rev response element (RRE) adopts alternative conformations that
753 promote different rates of virus replication. *Nucleic Acids Res* **43**, 4676-4686
754 (2015). https://doi.org/10.1093/nar/gkv313
755 12 Jackson, P. E., Tebit, D. M., Rekosh, D. & Hammarskjold, M. L. Rev-RRE
756 Functional Activity Differs Substantially Among Primary HIV-1 Isolates. *AIDS*

757 *Res Hum Retroviruses* **32**, 923-934 (2016).
758 https://doi.org/10.1089/AID.2016.0047
759 13 Saldi, T., Riemondy, K., Erickson, B. & Bentley, D. L. Alternative RNA
760 structures formed during transcription depend on elongation rate and
761 modify RNA processing. *Molecular cell* **81**, 1789-1801. e1785 (2021).
762 14 Le Sage, V. *et al.* Mapping of Influenza Virus RNA-RNA Interactions Reveals a
763 Flexible Network. *Cell Rep* **31**, 107823 (2020).
764 https://doi.org/10.1016/j.celrep.2020.107823
765 15 Miyamoto, S. & Noda, T. In vitro vRNA-vRNA interactions in the H1N1
766 influenza A virus genome. *Microbiol Immunol* **64**, 202-209 (2020).
767 https://doi.org/10.1111/1348-0421.12766
768 16 Li, P. *et al.* Integrative Analysis of Zika Virus Genome RNA Structure Reveals
769 Critical Determinants of Viral Infectivity. *Cell Host Microbe* **24**, 875-886 e875
770 (2018). https://doi.org/10.1016/j.chom.2018.10.011
771 17 Ziv, O. *et al.* COMRADES determines in vivo RNA structures and interactions.
772 *Nat Methods* **15**, 785-788 (2018). https://doi.org/10.1038/s41592-018-
773 0121-0
774 18 Huber, R. G. *et al.* Structure mapping of dengue and Zika viruses reveals
775 functional long-range interactions. *Nature communications* **10**, 1408 (2019).
776 https://doi.org/10.1038/s41467-019-09391-8
777 19 Ziv, O. *et al.* The Short- and Long-Range RNA-RNA Interactome of SARS-CoV-
778 2. *Mol Cell* (2020). https://doi.org/10.1016/j.molcel.2020.11.004
779 20 Cao, C. *et al.* The architecture of the SARS-CoV-2 RNA genome inside virion.
780 *Nature communications* **12** (2021). https://doi.org/10.1038/s41467-021-
781 22785-x
782 21 Yang, S. L. *et al.* Comprehensive mapping of SARS-CoV-2 interactions in vivo
783 reveals functional virus-host interactions. *Nature communications* **12**, 5113
784 (2021). https://doi.org/10.1038/s41467-021-25357-1
785 22 Zhang, Y. *et al.* In vivo structure and dynamics of the SARS-CoV-2 RNA
786 genome. *Nature communications* **12**, 5695 (2021).
787 https://doi.org/10.1038/s41467-021-25999-1
788 23 Gabryelska, M. M. *et al.* Global mapping of RNA homodimers in living cells.
789 *Genome Res* (2022). https://doi.org/10.1101/gr.275900.121
790 24 Zhang, M. *et al.* Classification and clustering of RNA crosslink-ligation data
791 reveal complex structures and homodimers. *Genome Res* **32**, 968-985 (2022).
792 https://doi.org/10.1101/gr.275979.121
793 25 Jäger, R. *et al.* Capture Hi-C identifies the chromatin interactome of colorectal
794 cancer risk loci. *Nature communications* **6**, 6178 (2015).
795 26 Dryden, N. H. *et al.* Unbiased analysis of potential targets of breast cancer
796 susceptibility loci by Capture Hi-C. *Genome research* **24**, 1854-1868 (2014).
797 27 Sahlen, P. *et al.* Genome-wide mapping of promoter-anchored interactions
798 with close to single-enhancer resolution. *Genome Biol* **16**, 156 (2015).
799 https://doi.org/10.1186/s13059-015-0727-9
800 28 Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with
801 high-resolution capture Hi-C. *Nature genetics* **47**, 598-606 (2015).

29 Chang, S. *et al.* 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 (2015).

30 Travis, A. J., Moody, J., Helwak, A., Tollervey, D. & Kudla, G. Hyb: a bioinformatics pipeline for the analysis of CLASH (crosslinking, ligation and sequencing of hybrids) data. *Methods* **65**, 263-273 (2014). https://doi.org/10.1016/j.ymeth.2013.10.015

31 Stephenson, J. D. *et al.* Three-dimensional RNA structure of the major HIV-1 packaging signal region. *Structure* **21**, 951-962 (2013).

32 Mailler, E. *et al.* The Life-Cycle of the HIV-1 Gag-RNA Complex. *Viruses* **8** (2016). https://doi.org/10.3390/v8090248

33 Ding, P. *et al.* Identification of the initial nucleocapsid recognition element in the HIV-1 RNA packaging signal. *Proc Natl Acad Sci U S A* **117**, 17737-17746 (2020). https://doi.org/10.1073/pnas.2008519117

34 Zhang, K. *et al.* Structure of the 30 kDa HIV-1 RNA Dimerization Signal by a Hybrid Cryo-EM, NMR, and Molecular Dynamics Approach. *Structure* **26**, 490-498 e493 (2018). https://doi.org/10.1016/j.str.2018.01.001

35 Änkö, M.-L. *et al.* The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes. *Genome biology* **13**, 1-17 (2012).

36 Sertznig, H., Hillebrand, F., Erkelenz, S., Schaal, H. & Widera, M. Behind the scenes of HIV-1 replication: Alternative splicing as the dependency factor on the quiet. *Virology* **516**, 176-188 (2018). https://doi.org/10.1016/j.virol.2018.01.011

37 Nguyen Quang, N. *et al.* Dynamic nanopore long-read sequencing analysis of HIV-1 splicing events during the early steps of infection. *Retrovirology* **17**, 25 (2020). https://doi.org/10.1186/s12977-020-00533-1

38 Kutluay, S. B. *et al.* Global changes in the RNA binding specificity of HIV-1 gag regulate virion genesis. *Cell* **159**, 1096-1109 (2014). https://doi.org/10.1016/j.cell.2014.09.057

39 Crane, E. *et al.* Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* **523**, 240-244 (2015). https://doi.org/10.1038/nature14450

40 Martin, P. *et al.* Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. *Nature communications* **6**, 10069 (2015).

41 Baxter, J. S. *et al.* Capture Hi-C identifies putative target genes at 33 breast cancer risk loci. *Nature communications* **9**, 1028 (2018).

42 Orlando, G. *et al.* Promoter capture Hi-C-based identification of recurrent noncoding mutations in colorectal cancer. *Nature genetics* **50**, 1375 (2018).

43 Chesi, A. *et al.* Genome-scale Capture C promoter interactions implicate effector genes at GWAS loci for bone mineral density. *Nature communications* **10**, 1260 (2019).

44 Li, G. & De Clercq, E. HIV genome-wide protein associations: a review of 30 years of research. *Microbiology and Molecular Biology Reviews* **80**, 679-731 (2016).

848 45 Dadonaite, B. *et al.* The structure of the influenza A virus genome. *Nature*
849 *Microbiology* **4**, 1781-1789 (2019). https://doi.org/10.1038/s41564-019-
850 0513-7
851 46 Yao, H. *et al.* Molecular Architecture of the SARS-CoV-2 Virus. *Cell* **183**, 730-
852 738 e713 (2020). https://doi.org/10.1016/j.cell.2020.09.018
853 47 Becker, J. T. & Sherer, N. M. Subcellular Localization of HIV-1 gag-pol mRNAs
854 Regulates Sites of Virion Assembly. *J Virol* **91** (2017).
855 https://doi.org/10.1128/JVI.02315-16
856 48 Chen, T. *et al.* The genome sequence archive family: toward explosive data
857 growth and diverse data types. *Genomics, Proteomics and Bioinformatics* **19**,
858 578-583 (2021).
859 49 Database resources of the national genomics data center, China national
860 center for bioinformation in 2022. *Nucleic Acids Research* **50**, D27-D38
861 (2022).