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 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

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 RNA-
51 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 seven AIDS patients been completely "cued" [1 2,3]. Research on HIV-1 RNA structure
56 and function has intensified, focusing on assembly, release, and maturation
57 processes, using NMR[4], cryo-EM[5] and chemical probing[6-8]. However, our knowledge
58 of the global architecture and the long-range interactions within the HIV-1 RNA
59 genome remains limited, primarily due to the complexity and dynamic nature of
60 these 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 RRE in the
66 HIV genome, responsible for nucleocytoplasmic transport of viral RNA, exhibits
67 structural variability impacting virus replication rates[11,12]. While current techniques
68 like SHAPE began to provide insights into local RNA structures[7,8,13], studying the
69 intricate long-range RNA interactions and homodimers within the HIV genome
70 under physiological conditions remains a significant challenge, primarily due to the
71 limitations of these methods in capturing the full spectrum of RNA interactions and
72 the dynamic nature of the viral genome.

73 Proximity ligation-based methods have been shown to be particularly useful for
74 studying long-range RNA interactions, as they can overcome the RNA length
75 limitations of traditional techniques. These methods have been applied to the study
76 of various viruses, including influenza[14,15], Zika[16-18], and SARS-CoV-2[19-22], and have

77 led to the identification of numerous RNA structures that are closely related to the
78 virus lifecycle. We and others have recently shown that RNA proximity ligation data
79 also contains information about RNA homodimerization events[23,24]. Nevertheless, as
80 of now, there are no proximity ligation studies of long-range interactions or dimeric
81 interactions of the HIV RNA genome. Probably due to high requirements of starting
82 material, making it very difficult for low abundant HIV sample treatment.

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

96 results

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

98 To effectively capture and analyze full-length low abundant HIV RNA-RNA
99 interactions in infected cells, we developed an improved version of the protocol,
100 which we call **Hi**gh throughput **Cap**ture of **R**NA interactions (**HiCapR**). This method
101 incorporates hybridization-based capture with next-generation sequencing (NGS)
102 library construction, a principle well established in capture Hi-C methods [25-28] and
103 in RNA proximity ligation [17,19]. Applying this strategy, we compared the NL4-3 and
104 GX2005002 strains [29],which are two distinct HIV-1 strains that exhibit different
105 prevalence patterns in various geographical regions. The NL4-3 strain is a well-
106 characterized laboratory strain that is widely used in HIV research and is

107 representative of the HIV-1 subtype B, which is highly prevalent in Europe and the
108 Americas. On the other hand, GX2005002 is a primary isolate of the CRF01_AE
109 subtype, which is one of the most prevalent strains in Southeast Asia, particularly in
110 China. The overall design of this study is presented in Figure 1A, with detailed
111 methods provided in the methods section. Supplementary Figure S1 illustrates the
112 bioinformatic pipeline used to detect RNA-RNA interactions and
113 homodimers.Approximately 16 million raw reads were obtained per sequencing
114 library, with over 95% of the reads containing HIV sequences, indicating efficient
115 capture of HIV RNA transcripts. The alignment rate exceeded 90%, and we achieved
116 an average depth of coverage of approximately 11,500X across the genome
117 (Supplementary Table 1). The uniform distribution of reads post-capture
118 (Supplementary Figures S2) demonstrates HiCapR's effectiveness in capturing the
119 entire HIV RNA genome with minimal bias. The overall duplication rate remained
120 below 50%, with a lower rate in the ligation group, suggesting increased library
121 complexity due to RNA proximity ligation.

122 Chimeric reads were identified using the hyb technique as previously described[22],
123 with a database comprising human spliced mRNAs, noncoding RNAs[30] and the HIV
124 genome sequence as reference. The chimeric rate, a crucial indicator of ligation
125 efficiency, was approximately 1% in non-ligation controls and around 9% in the
126 ligation group, aligning with the widely reported range. The proportion of 5'-3' and
127 3'-5' chimeras was nearly equal in long-range interactions, while proximal
128 interactions were enriched for 3'-5' chimeras. Pearson correlation analysis revealed
129 a high degree of correlation ($r>0.99$) between biological replicates, indicating high
130 reproducibility. Additionally, a significant similarity was observed between virions
131 in the supernatant and infected cells from the same viral strain (Supplementary
132 Figure S3).

133 Using a similar approach as described in our previous paper [22], we generated a
134 contact matrix for the global HIV-1 genome, revealing clear local and long-range
135 interactions (Figure 1B and Supplementary Figure S4). In summary, HiCapR
136 demonstrated high reproducibility, low bias, reliability, and efficiency in capturing

137 HIV transcripts, making it a robust tool for analyzing the RNA structure of the HIV
138 genome and potentially other similar RNA viruses.

139

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

141 A) Experimental design of HiCapR for profiling RNA-RNA interactions and
142 dynamics of the HIV genome. Based on the simplified SPLASH protocol, HiCapR
143 incorporates a post-library probe-based hybridization and streptavidin
144 pulldown method to enrich HIV RNA chimeras from the SPLASH library. The
145 HiCapR method has been applied to two strains of HIV, NL4-3 and GX2005002,
146 in both infected cells and their corresponding virus particles.

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

151

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

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

156 Chimeras analysis support the presence of key structural elements, including TAR,
157 polyA, SL1, SL2, and SL3, as well as polyA-SL1 in the monomeric conformation in the
158 HIV genome (Figure 2A and Supplementary Figure S5). Despite the pivotal role of
159 the 5'-UTR in replication, packaging, transcription and translation, we noted that the
160 5'-UTR sequence of HIV is not highly conserved across different HIV strains. A
161 comparison between the NL4-3 and GX2005002 strains revealed notable insertions
162 and deletions in the U5 region, along with point mutations in the loop of the SL1
163 core region, which is critical for HIV dimerization (Figure 2B). These findings raise
164 questions about the conservation of the structural integrity of this region. We
165 therefore applied the comradesFold algorithm [17] to our HiCapR data to generate

166 1000 structure predictions of the 5'-UTR (extending 100 nt downstream of the AUG
167 start codon) for each sample, and then performed MDS analysis after aligning the
168 viral genome coordinates of both strains (Figure 2C). This analysis revealed that the
169 reported dimer and monomer structures clustered together in NL4-3 but form
170 distinct clusters in GX2005002(Figure 2C). Based on these folded structures, we
171 calculated the base pairing probability for each base pair, which involves
172 determining the number of folded structures supporting a specific base pair divided
173 by the total number of structures. Visualizing this base pairing probability as a
174 heatmap identifies the most stable base pairs in the 5' UTR of HIV. We observed a
175 consistent presence of key structural elements such as polyA, TAR, SL1, SL2, and SL3
176 in both NL4-3 and GX2005002 strains (Figure 2D), suggesting robustness in the
177 overall structure despite sequence variations and alternative RNA conformations.

178

Figure2

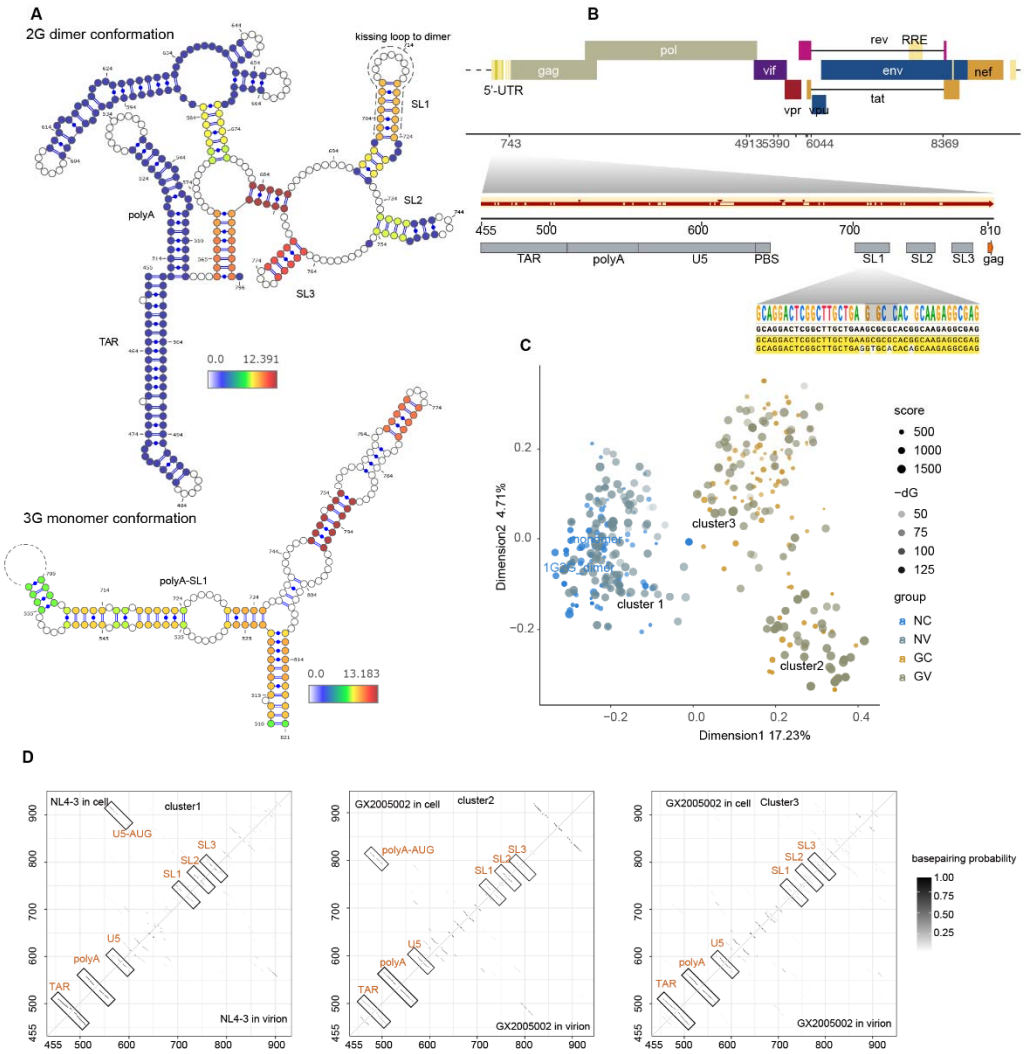


Figure 2 conformations of HIV-1 5'-UTR

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

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

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

194 D) Basepairing probability matrices were calculated from 1000 computationally
195 predicted structures, with color indicating the percentage of structures
196 supporting the specific basepair.

197

198

199 **3. Novel homodimerization sites and their implications in RNA splicing and viral**
200 **assembly**

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

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

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

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

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

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

248 1 splicing factors[36]. This result suggests a potential role of dimerization in RNA
249 splicing processes within the HIV genome.

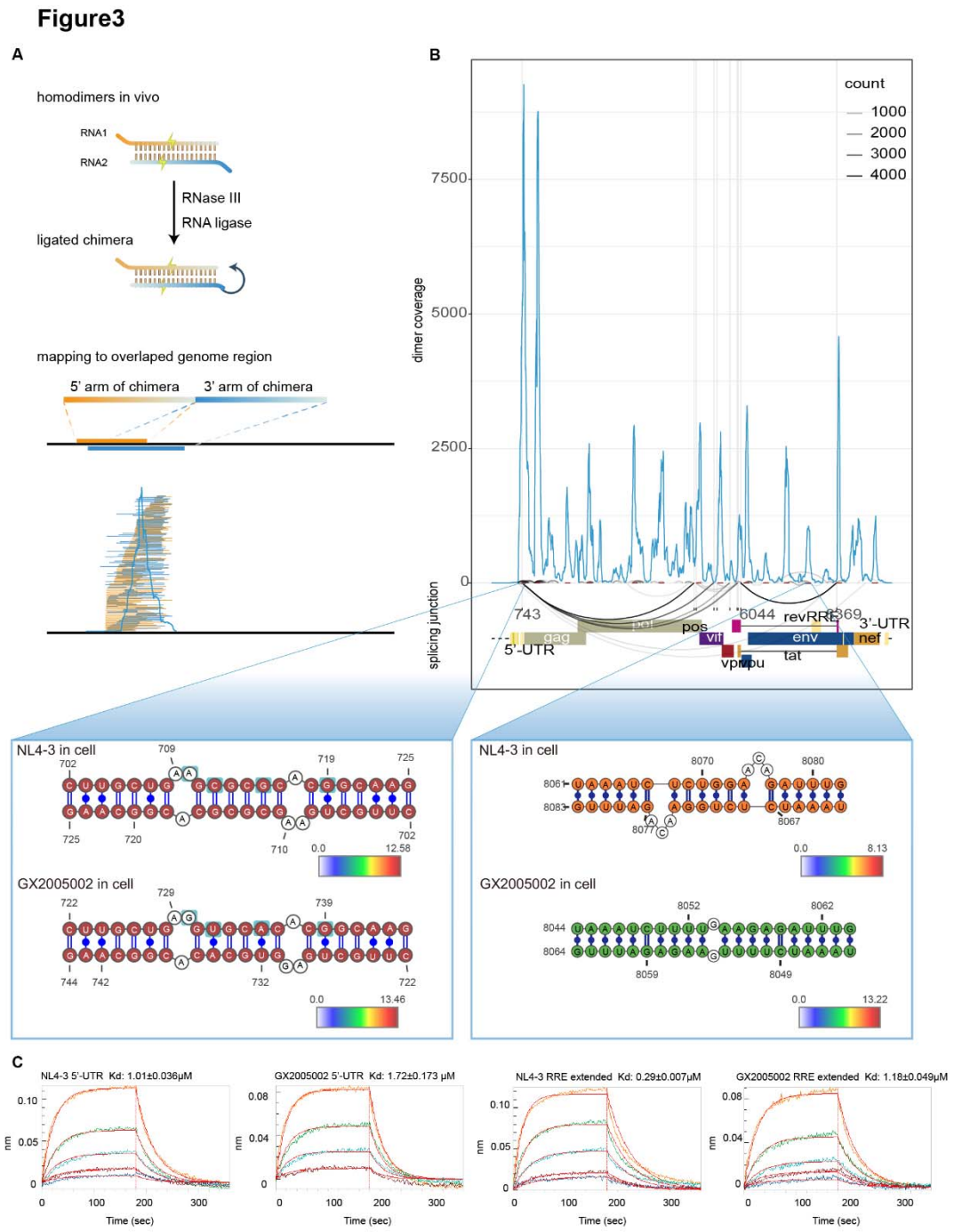
250 5'-3' discontinuous reads were identified in non-ligation control data, enabling the
251 inference of splicing junction sites. These inferred sites showed strong concordance
252 with canonical splicing sites and recent nanopore sequencing data[37] (Figure 3B).
253 Notably, almost every junction site, including splicing donor and acceptor sites,
254 exhibited dimer peaks around them. Additionally, the majority of dimeric chimeras
255 covered these junction sites and enriched downstream of acceptor sites
256 (Supplementary Figure S10), suggesting that dimerization process/events takes
257 place on unspliced genomic RNA. Interestingly, dimers surrounding splicing
258 regulatory elements (as summarized in[36]) form stable base pairing supported by
259 numerous chimeras (Supplementary Figure S10). This observation underscores the
260 close proximity and potential functional relationship between dimerization events
261 and RNA splicing processes at these critical sites within the HIV genome, suggesting
262 a potential role of dimerization in splicing processes or assisting in the transport of
263 unspliced RNA out of the nucleus.

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

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

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

280



281

282

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

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

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

296 C) Assessment of dimer self-binding using Bio-Layer Interferometry (BLI). The data
297 presented here are from three separate experiments, offering insights into how
298 dimers interact and bind to themselves. The dissociation constant (Kd),
299 indicated by the mean ± standard deviation, was determined from these
300 experiments.

301

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

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

311 the flanking regions located 1/2 upstream and 1/2 downstream (Supplementary
312 Figure S12). This analysis unveiled significant local interactions around dimer sites,
313 suggesting the involvement of stable local structures in the formation of dimers
314 between two copies of the HIV genomes.

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

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

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

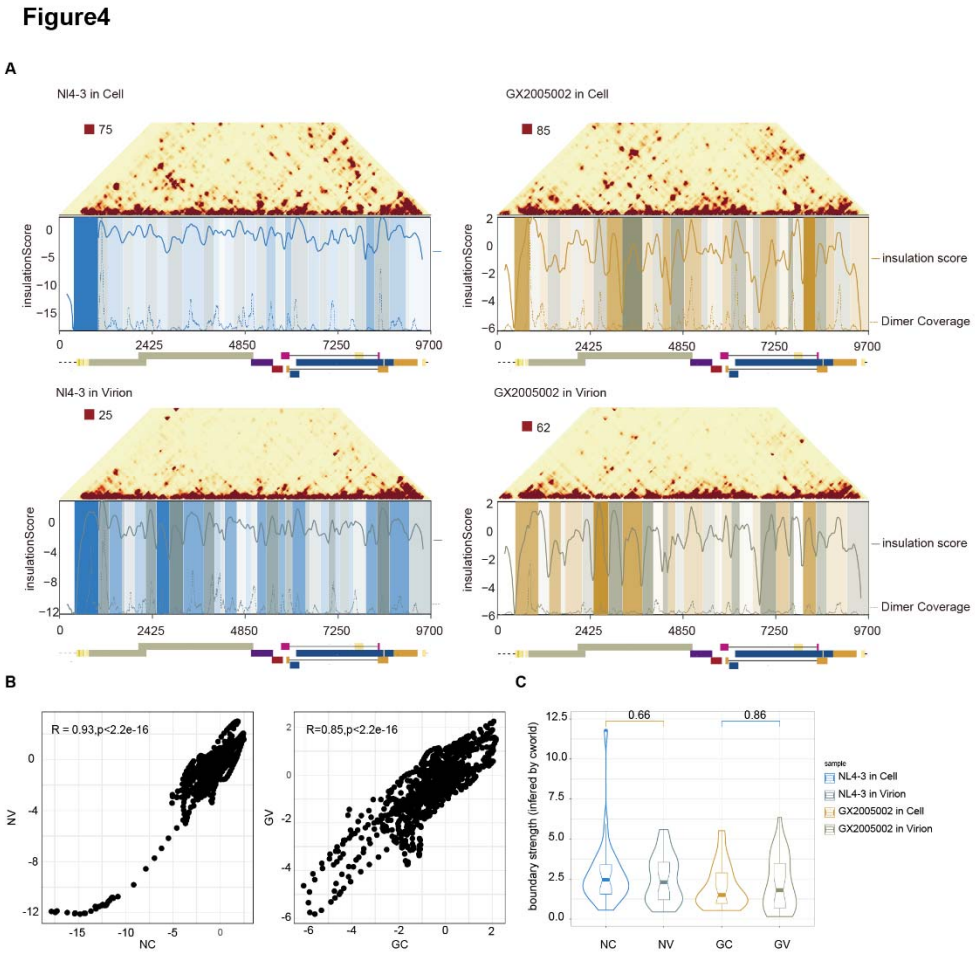


Figure 4 genome domains along HIV-1 genome

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.

344

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

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

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

357 Specifically, we employed viewpoint analysis to examine interactions involving the
358 5'-UTR. The results, depicted in Figure 5B, reveal multiple peaks of 5'-UTR
359 interactions across the HIV-1 genome. Interestingly, for the majority of these peaks,
360 the signals are lower in virions compared to cells. However, cyclization interactions
361 (5'-UTR-3'-UTR) at the ends are still maintained in virions (Figure 5C). Intriguingly,
362 we also identified multiple interaction peaks located near crucial elements, such as
363 the frameshift element (Supplementary Figure S16A), rev start codon
364 (Supplementary Figure S16B), RRE, and nuclear exporting signal (NES)
365 (Supplementary Figure S16C), indicating complex roles of 5'-UTR in HIV life cycle.
366 Among these interactions, some are strain specific, such as the interaction between
367 5'-UTR and FSE is specific to the NL4-3 strain, while interaction between 5'-UTR and
368 8.5K region (near NES) are highly enriched in GX2005002 strain (Supplementary
369 Figure S16C). As mentioned earlier, the majority of interactions decrease in the
370 virion, with the notable exception of interactions in the 5'-UTR and 6K region (near
371 D4 splicing donor site), which significantly increase in the virion. (Supplementary
372 Figure S16B). The enhanced interactions in virions are consistent across both
373 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

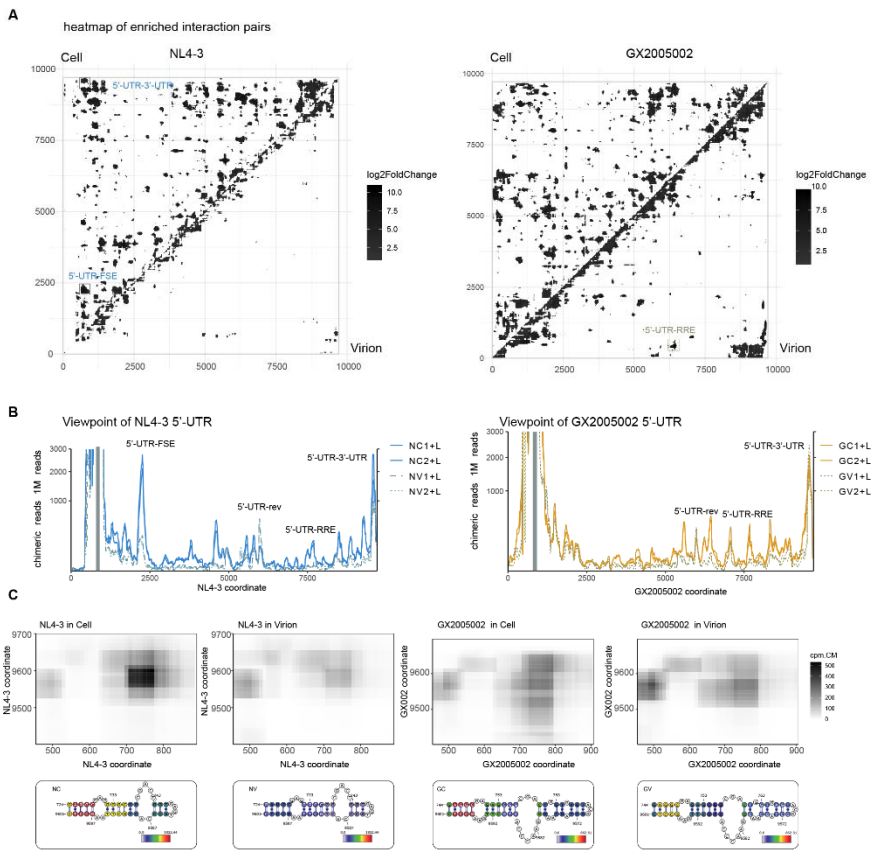


Figure 5 long range interaction in HIV-1 genome

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

386 B) Viewpoint lines depicts the binding positions of the 5'-UTR along the HIV-1
387 genome. The gray rectangles indicate the viewpoint regions. The colors of the
388 lines represent specific samples, with samples from virions shown as dashed
389 lines.

390 C) Contact matrices and base pairing details between 5'-UTR and 3'-UTR. The top
391 panels display heatmaps indicating contact probability, with the color bar
392 indicating chimeric reads per 1M reads in each specific sample. The bottom
393 panels show base pairing colored by base pairing scores.

394

395 **discussion**

396 **1. HicapR：a reliable and efficient method for exploring structures of viral RNA**
397 **genomes**

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

404 HiCapR uses a similar hybridization-based capture strategy as COMRADES[17,19]
405 except that in HiCapR, this capture step occurs after library construction, whereas in
406 COMRADES, it takes place prior to proximity ligation. The post-library-capture
407 principle employed in HiCapR is a well-established approach utilized in other
408 methods such as capture Hi-C or capture-C[40-43], and it offers additional flexibility
409 relative to previous protocols. Our results demonstrate remarkable sensitivity and
410 reliability of the HiCapR technique in capturing low-abundance HIV RNA structure,
411 providing an unprecedented resolution and comprehensive approach to elucidate
412 the intricate RNA structures within the HIV-1 genome. This novel proximity ligation
413 method transcends the limitations of traditional techniques, offering a robust
414 framework for dissecting the complex architecture of viral RNA.

415 **2. New insight of HIV RNA local and long-range interactions**

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

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

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

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

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

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

449 **3. Emerging Insights into HIV-1 Dimerization through newly identified dimer sites**

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

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

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

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

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

492 **methods**

493 **Cell culture and virus infection**

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

504 **HicapR method**

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

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

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

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

524

525 **Data preprocessing and chimeric reads identification**

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

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

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

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

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

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

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

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

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

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

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

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

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

558 **Dimer score and dimer base pairing**

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

564 **Local structure folding and MDS plot**

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

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

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

573 **Calculation of base pairing probability for local structure**

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

578 **Dimer validation**

579 **RNA preparation.** 5'-UTR and RRE DNA fragments were amplified by RT-PCR using
580 RNA from infected cells, with primers listed in Supplementary table3. Purified NL4-
581 3 and GX2005002 5'-UTR and extended RRE PCR products (200 ng) were utilized as
582 templates for RNA in vitro transcription with T7 RNA polymerase (Vazyme),
583 following manufacturer's protocol. The reaction was then incubated at 37 °C for 16
584 hours, followed by DNase I treatment for 30 minutes at 37 °C. The RNA was
585 subsequently purified using MagicPure RNA Beads (TransGen) through gel
586 purification.

587 **Native agarose gel electrophoresis.** RNA (600 ng) was heated to 95°C and then
588 slowly cooled to room temperature in either high salt buffer (50mM Tris-HCl pH 7.5,
589 140mM KCl, 10mM NaCl) or low salt buffer (high salt buffer diluted 1:10 with
590 water). Samples were loaded with native loading dye (0.17% Bromophenol Blue and
591 40% (vol/vol) sucrose) on 2% agarose gel prepared with 1× tris-borate magnesium
592 (TBM) buffer (89 mM Tris base, 89 mM boric acid and 2 mM MgCl2) and
593 fractionated at 100 V for 85 min at room temperature. RNA in the PAGE gel was
594 visualized using GelRed nucleic acid gel stain(Thomas scientific).

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

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

610 Data availability

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

615 **Code availability**

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

618 **Supplementary Figure legends**

619 **Supplementary Figure S1**

620 This figure illustrates the workflow for identifying chimeric reads supporting RNA-
621 RNA interactions as well as homodimers. The rounded rectangle in the upper right
622 corner of the figure displays the start and end positions of the 5'- and 3'-arms of
623 chimeric reads. These arms can be ligated in multiple ways to form chimeras, some
624 of which are thought to be produced exclusively by homodimers.

625 First, sequencing reads are mapped to the HIV RNA genome, and chimeric reads are
626 identified using hyb pipeline. Dimeric chimeras are then filtered to select only those
627 that can be formed exclusively by dimerization. To reduce background noise,
628 especially contaminated from homodimers, we selected non-overlapping chimeras
629 for our analysis of RNA interactions.

630 This approach allowed us to identify both RNA-RNA interactions and candidate
631 dimer sites and predict secondary structures with high coverage chimeras.

632 **Supplementary Figure S2**

633 HIV Genome coverage of HiCapR data.

634 (A) The organization of the HIV NL4-3 genome follows NCBI annotation and is
635 specified.

636 (B) and (C) The sequencing data from each group was mapped to the HIV genome
637 and then genome coverage was calculated. The results showed that both NL4-3 (B)
638 and GX2005002 (C) strains of the virus, whether in infected cells or virions, and
639 whether RNAs are ligated or not, had uniformly complete genome coverage. This
640 suggests that the HiCapR method is able to capture the entire HIV genome with low
641 bias.

642 **Supplementary Figure S3**

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

645 **Supplementary Figure S4**

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

649 **Supplementary Figure S5**

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

653 **Supplementary Figure S6**

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

657 **Supplementary Figure S7**

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

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

666 **Supplementary Figure S8**

667 Dimer score matrices around 5’-UTR of NL4-3 and GX2005002 strain in different
668 stages.

669 **Supplementary Figure S9**

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

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

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

677 **Supplementary Figure S10**

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

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

684 **Supplementary Figure S11**

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

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

692 **Supplementary Figure S12**

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

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

696 **Supplementary Figure S13**

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

699 **Supplementary Figure S14**

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

703 **Supplementary Figure S15**

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

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

710 **Supplementary Figure S16**

711 Contact matrices and base pairing details between 5'-UTR and 2.2K (A), 6K (B) and
712 7.6~8.5K (C).

713 **Supplementary data**

714 **Supplementary table1**

715 Mapping statistics

716 **Supplementary table2**

717 Homodimer peaks information

718 Supplementary table3

719 Primers used for amplifying 5'-UTR and extended RRE

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).
802 29 Chang, S. *et al.* Comparison of susceptibility of HIV-1 variants to
803 antiretroviral drugs by genotypic and recombinant virus phenotypic
804 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).

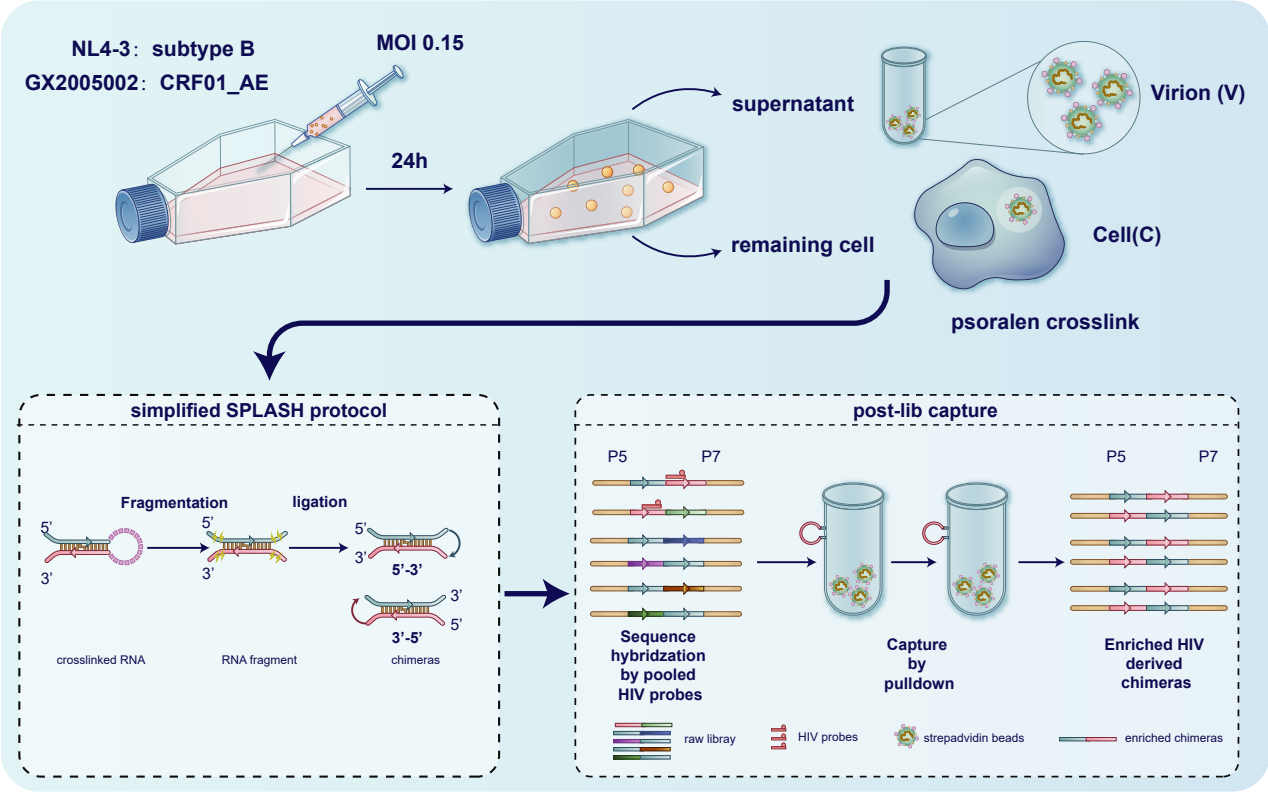
45 Dadonaite, B. *et al.* The structure of the influenza A virus genome. *Nature Microbiology* **4**, 1781-1789 (2019). https://doi.org/10.1038/s41564-019-0513-7

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).

862

Figure1

A



B contact matrix normalized by total mapped read counts

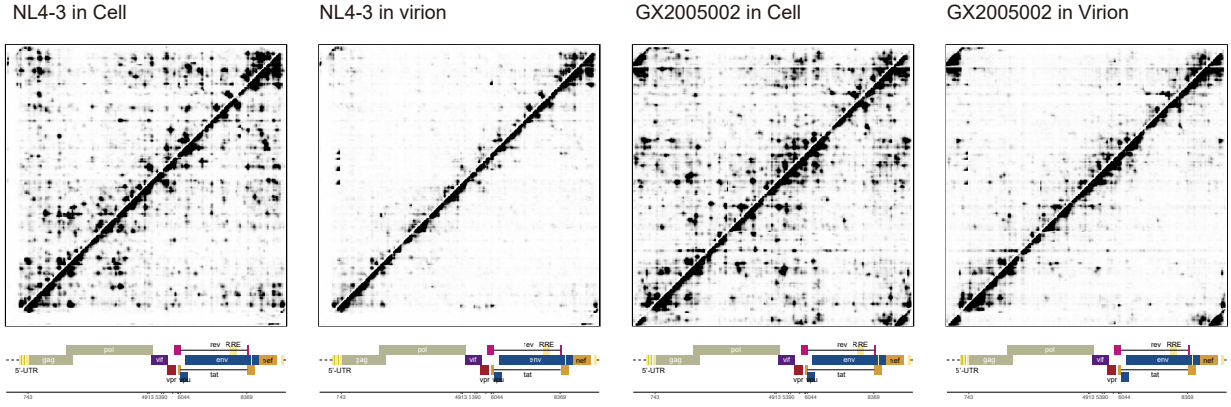
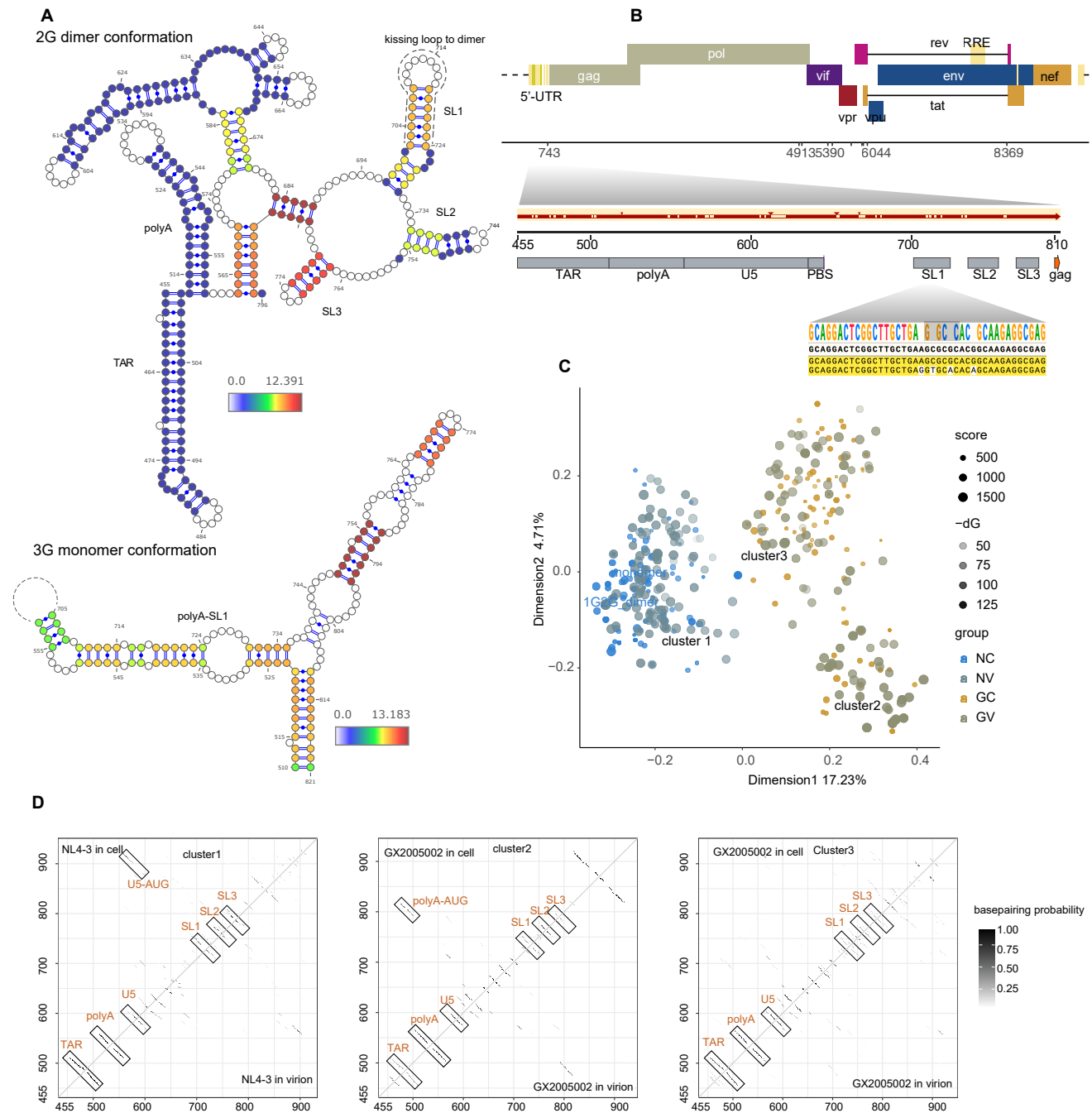


Figure2



A
2G dimer conformation
kissing loop to dimer
SL1
SL2
SL3
polyA
TAR
0.0 12.391
3G monomer conformation
polyA-SL1
0.0 13.183
B
pol
gag
vif
env
nef
rev
RRE
tat
vpr
vpu
5'-UTR
743
49135390
6044
8369
455
500
600
700
810
TAR
polyA
U5
PBS
SL1
SL2
SL3
gag
C
Dimension2 4.71%
Dimension1 17.23%
cluster 1
cluster2
cluster3
score
500
1000
1500
−dG
50
75
100
125
group
NC
NV
GC
GV
D
NL4-3 in cell
cluster1
U5-AUG
SL3
SL2
SL1
U5
polyA
TAR
NL4-3 in virion
GX2005002 in cell
cluster2
polyA-AUG
GX2005002 in virion
GX2005002 in cell
Cluster3
basepairing probability
1.00
0.75
0.50
0.25

Figure3

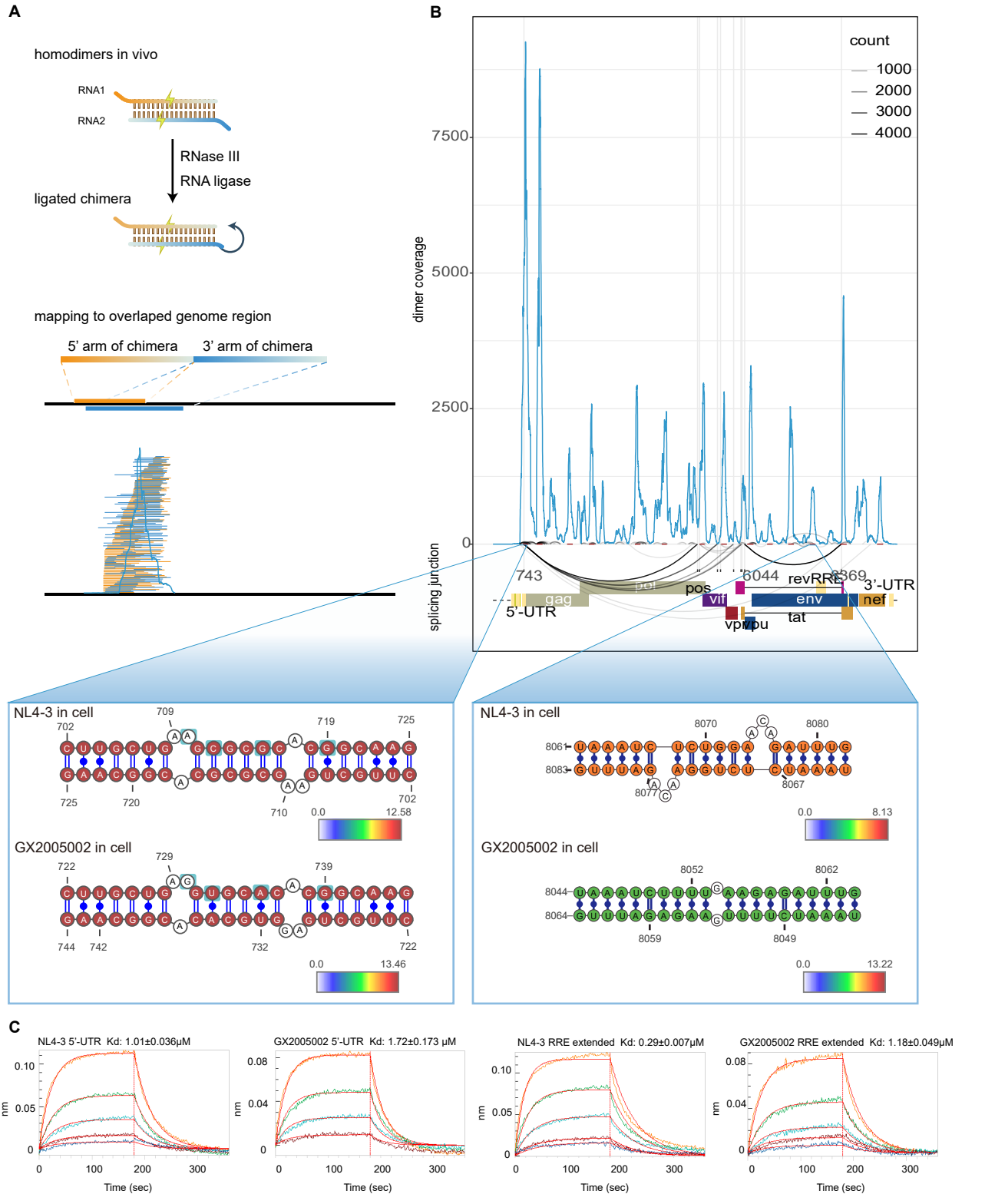

A
homodimers in vivo
RNA1
RNA2
RNase III
RNA ligase
ligated chimera
mapping to overlaped genome region
5' arm of chimera
3' arm of chimera
B
count
1000
2000
3000
4000
dimer coverage
7500
5000
2500
0
splicing junction
743
6044
revRRE
6369
3'-UTR
pol
pos
gag
vif
env
nef
5'-UTR
tat
vpr
vpu
NL4-3 in cell
GX2005002 in cell
C
NL4-3 5'-UTR Kd: 1.01±0.036μM
GX2005002 5'-UTR Kd: 1.72±0.173 μM
NL4-3 RRE extended Kd: 0.29±0.007μM
GX2005002 RRE extended Kd: 1.18±0.049μM
nm
Time (sec)

Figure4

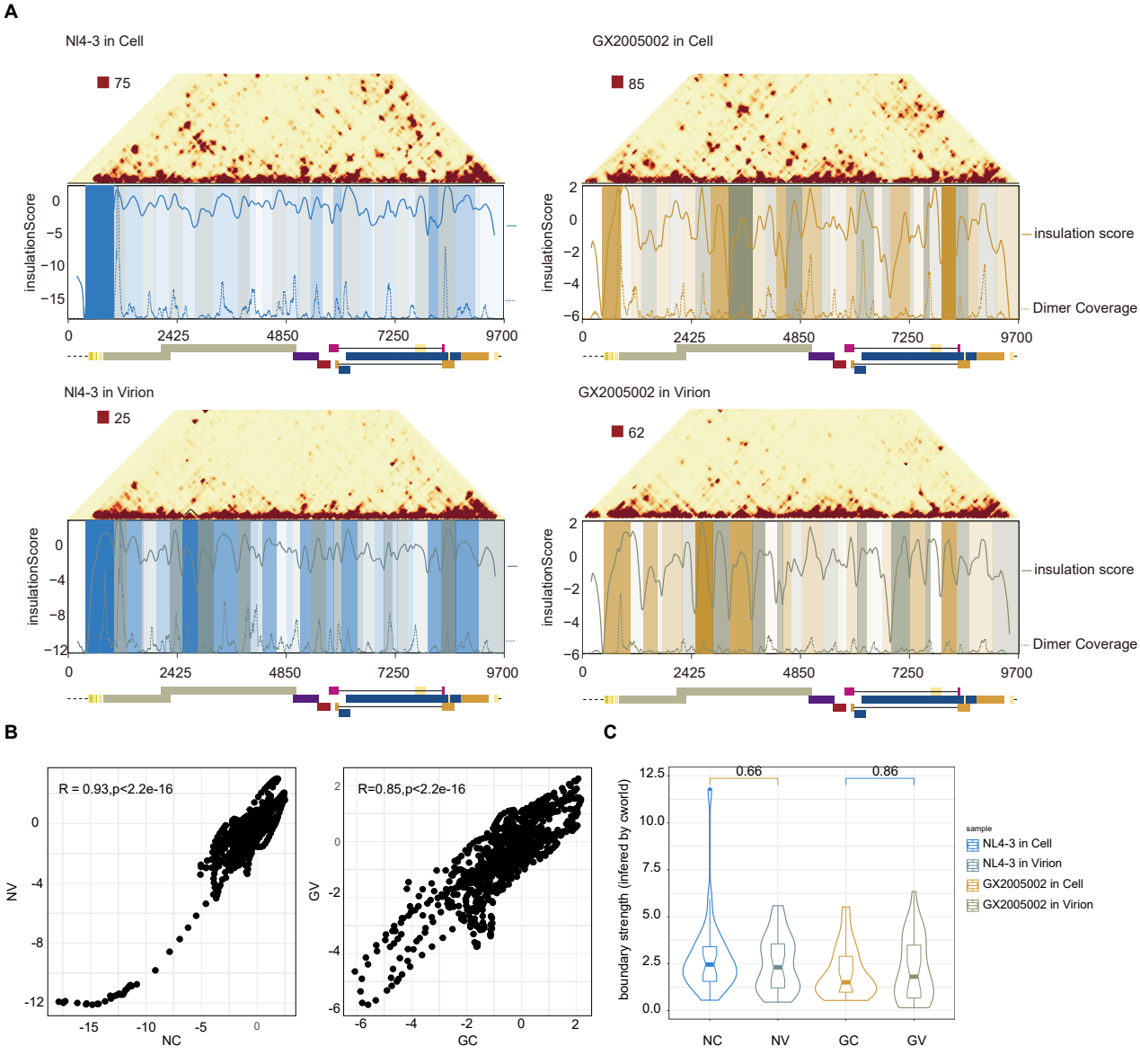
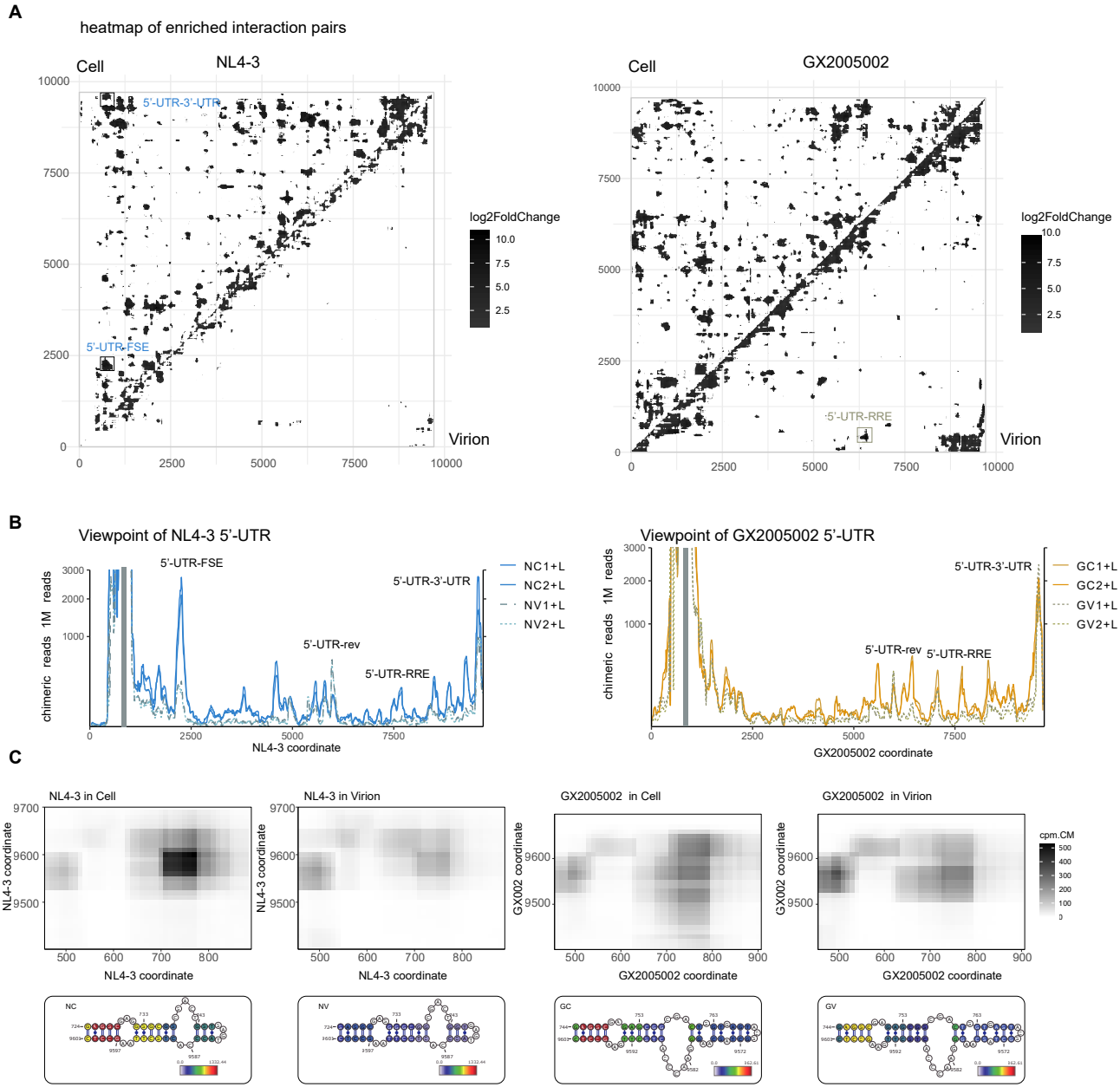

A
NI4-3 in Cell
75
insulationScore
0
−5
−10
−15
0
2425
4850
7250
9700
GX2005002 in Cell
85
insulationScore
2
0
−2
−4
−6
insulation score
Dimer Coverage
NI4-3 in Virion
25
insulationScore
0
−4
−8
−12
GX2005002 in Virion
62
B
R = 0.93,p<2.2e-16
NV
NC
R=0.85,p<2.2e-16
GV
GC
C
0.66
0.86
boundary strength (infered by cworld)
12.5
10.0
7.5
5.0
2.5
0.0
NC
NV
GC
GV
sample
NL4-3 in Cell
NL4-3 in Virion
GX2005002 in Cell
GX2005002 in Virion

Figure5


A
heatmap of enriched interaction pairs
Cell
NL4-3
5'-UTR-3'-UTR
5'-UTR-FSE
Virion
log2FoldChange
10.0
7.5
5.0
2.5
GX2005002
5'-UTR-RRE
B
Viewpoint of NL4-3 5'-UTR
5'-UTR-FSE
5'-UTR-3'-UTR
5'-UTR-rev
5'-UTR-RRE
NC1+L
NC2+L
NV1+L
NV2+L
chimeric reads 1M reads
NL4-3 coordinate
Viewpoint of GX2005002 5'-UTR
GC1+L
GC2+L
GV1+L
GV2+L
GX2005002 coordinate
C
NL4-3 in Cell
NL4-3 in Virion
GX2005002 in Cell
GX2005002 in Virion
NL4-3 coordinate
GX002 coordinate
GX2005002 coordinate
cpm.CM
NC
NV
GC
GV