Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Structural analysis of the 5' terminal of human immunodeficiency virus type 1 genome RNAs transcribed from proviruses with mutations in the transcription start site

**Yuka Kashiyama**
Chiba Institute of Technology
**Gota Kawai**
Chiba Institute of Technology
**Takeshi Yoshida**
Japan Institute for Health Security
**Youichi Suzuki**
youichi.suzuki@ompu.ac.jp

Osaka Medical and Pharmaceutical University

**Article**

**Keywords:** HIV-1, genomic RNA, functional role of transcript, mutation at the transcription start site, NMR

**Posted Date:** October 27th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11237882/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Structural analysis of the 5' terminal of human immunodeficiency virus type 1 genome RNAs transcribed from proviruses with mutations in the transcription start site

Yuka Kashiyama[1], Gota Kawai[1,#], Takeshi Yoshida[2] and Youichi Suzuki[3,*]

[1] Department of Life Science, Graduate School of Advanced Engineering, Chiba Institute of Technology, Chiba, Japan

[2] AIDS Research Center, National Institute of Infectious Diseases, Japan Institute for Health Security, Tokyo, Japan

[3] Laboratory of Biosafety Research, Faculty of Medicine, Osaka Medical and Pharmaceutical University, Takatsuki, Japan

* To whom correspondence should be addressed. Email: youichi.suzuki@ompu.ac.jp

[#] Present address: Resorna Inc., Chiba, Japan.

**ABSTRACT**

Recent studies have shown that the functional roles of RNAs transcribed from the transcription start site containing three G residues in human immunodeficiency virus type 1 (HIV-1) are regulated by their structural stability, which is determined by the number of G residues at the 5′ terminus. In contrast, most proviruses in newly infected cells generated by HIV-1 have been shown to retain mutations introduced at the transcription start site. When C or U mutations are introduced at the third G residue, transcription from that position is unlikely to occur, suggesting that transcription instead occurs from the first or second G residue. In the absence of mutations, these transcripts are thought to be used for translation rather than incorporated into genomic RNA. Therefore, the mechanisms by which the functions of these RNAs are determined remain unclear. In this study, we designed model RNAs containing mutations and analyzed their structures and stabilities. RNAs transcribed from the second G residue with GC or GU at the 5'-end formed stable structures similar to those of RNAs transcribed from the third G residue in the native virus, which functions as ic RNA. These results suggest that, in the presence of these mutations, RNA transcribed from the second G functions as genomic RNA.

Keywords: HIV-1; genomic RNA; functional role of transcript; mutation at the transcription start site; NMR

## INTRODUCTION

The human immunodeficiency virus type 1 (HIV-1) genomic RNA is known to be transcribed from one of three continuous G residues in the provirus [1-4]. Masuda et al. demonstrated that, among the RNAs synthesized from the three transcription start sites, RNA transcribed from the third G residue is incorporated into virions as genomic RNA, whereas RNA transcribed from the first G residue is used for translation as mRNA in the cytoplasm (Fig. 1A) [1]. This difference has been shown to result from differences in the stability of the 5′ untranslated region (UTR) structure caused by the number of G residues at the 5′ terminus [2-4]. Hereafter, the RNA transcribed from the third G is referred to as G1G, the RNA transcribed from the second G residue as G2GG, and the RNA transcribed from the first G residue as G3GGG (Fig. 1A).

Yoshida et al. introduced mutations into the three G residues at the transcription start site of the HIV-1 provirus and examined their effects (Fig. 1B) [5]. G1G RNAs were primarily selected as templates for provirus DNA synthesis rather than for other RNAs. However, when the GGG sequence at the transcription start site was replaced with GGC or GGU, the mutant proviruses were capable of producing infectious viruses. Proviruses containing GGG and GGC were generated from GGC at frequencies of 29% and 71%, respectively, whereas those containing GGG and GGU were generated from GGU at frequencies of 17% and 83%, respectively (Table 1) [5]. These findings indicated that transcripts capable of functioning as genomic RNA were produced. Although the GGG sequence is also present in the 3′ UTR, the mutations were retained, indicating that HIV-1 reverse transcriptase must synthesize DNA even in the presence of mismatches. According to a study by Yoshida et al., the abundance of G1C or G1U in viral particles produced from cells containing proviruses with mismatches was low [5]. Therefore, RNAs transcribed from proviruses containing mismatches are expected to include G2GC or G2GU, which are transcribed from the second G residue, and G3GGC or G3GGU, which

are transcribed from the first G residue (Fig. 1C). However, the structural differences among these RNAs remain unclear.

In the present study, model RNAs representing the four transcripts were designed and analyzed to determine the structural and functional differences among these mismatch-containing transcripts. The results suggest that G2GC and G2GU preferentially function as genomic RNAs because they exhibit properties similar to those of G1G. In contrast, G3GGC and G3GGU showed properties similar to those of G3GGG, suggesting that they preferentially function as mRNAs.

## RESULTS

### Design of model RNAs

Fig. 2 shows the secondary structure of the 5′ terminal regions of the genomic RNA and the model RNAs used in this study. In this study, we focused on the region at the boundary between the TAR and PolyA stems (Fig. 2A).

Based on the study by Obayashi et al. [4], model RNAs were designed by introducing mutations into the RNAs used in that study. Because the TAR-side stem was found to be too short and the loop structure was disrupted, RNAs extended by five base pairs were used in the present study (Fig. 2B and C; TP-G1G-46, TP-G2GG-47, and TP-G3GGG-48). For comparison, stem-loop RNAs corresponding to the TAR side were also designed (Fig. 2B and C; TAR-G1G-29, TAR-G2GG-32, and TAR-G3GGG-33).

Next, model RNAs corresponding to RNAs transcribed from proviruses containing mutations were generated. For the C mutant, TP-G2GC-47 and TP-G3GGC-48 were prepared. For the U mutant, TP-G2GU-47 and TP-G3GGU-48 were prepared. In addition, TAR-G2GC-30, TAR-G3GGC-33, TAR-G2GU-30, and TAR-G3GGU-33, which correspond to the TAR-side stems of the respective RNA, were designed.

As described below, the designed RNAs formed the intended RNA structures (Supplemental Fig. S1).

**Structural comparison of the model RNAs without mutations**

Fig. 3 shows the results obtained using the model RNAs without mutations. To confirm the conformational characteristics of the model RNAs without mutations, polyacrylamide gel electrophoresis (PAGE) experiments were performed under two conditions: with 7 M urea (Fig. 3A, left panel) and without urea (urea-free; Fig. 3A, right panel). TP-G1G-46 (Fig. 3A, lane 1) migrated more slowly during PAGE with 7 M urea and more rapidly during PAGE without urea than TP-G2GG-47 or TP-G3GGG-48 (Fig. 3A, lanes 2 and 3, respectively). These results indicate that, under urea conditions, the structure of TP-G1G-46 was disrupted, resulting in slower migration, whereas TP-G2GG-47 and TP-G3GGG-48 retained some structural features and therefore adopted more compact conformations that migrated more rapidly. In contrast, under urea-free conditions, TP-G1G-46 migrated more rapidly as it formed a compact structure in which the two stems were connected, whereas TP-G2GG-47 and TP-G3GGG-48 migrated more slowly because the two stems remained independent (Fig. 3A).

Melting temperature analysis showed that TP-G1G-46 exhibited a single Tm value of 62.7 °C (Fig. 3B, red line), whereas TP-G2GG-47 and TP-G3GGG-48 exhibited two Tm values, one at approximately 45 °C and the other at 68 °C (Fig. 3B, green dash-dotted and green dashed lines, respectively). These results are consistent with the electrophoretic mobility data. The melting curves of the fragments are shown in Supplemental Fig. S2, and the Tm values are summarized in Table 1. For all RNAs, the TAR-side fragments showed Tm values comparable to those of the corresponding full-length RNAs. The Tm values observed at approximately 45 °C for TP-G2GG-47 and TP-

G3GGG-48 are likely attributable to the Poly(A)-side stem. The absence of this Tm transition in TP-G1G-46 indicates that the two stems melt cooperatively.

Furthermore, comparison of the NMR spectra of the three RNAs and their corresponding fragments confirmed that the TAR-side stem was formed in all three RNAs (Supplemental Fig. S2). In the NMR spectrum of TP-G1G-46, signals corresponding to U104, G105, G2, and G3 were clearly observed, whereas these signals were not detected in the spectra of the other RNAs (Fig. 3C, Supplemental Fig. S3). Because these signals originate from base pairs at the boundary between the TAR and PolyA stems, these results indicate that the two stems are connected in TP-G1G-46. This finding is consistent with the electrophoretic mobility and Tm analyses.

Taken together, these results are consistent with previous findings [4] and demonstrate that this model system is suitable for investigating the functions of these RNAs *in vivo*.

**Structural comparison of the model RNAs with mutations**

Fig. 4 shows the results obtained using the model RNAs with mutations. TP-G2GC-47 and TP-G2GU-47 migrated more slowly during PAGE with 7 M urea (Fig. 4A, left panel) and more rapidly during PAGE without urea (Fig. 4A, right panel). These migration patterns were similar to those of TP-G1G-46. In contrast, TP-G3GGU-48 migrated more rapidly during PAGE with urea and more slowly during PAGE without urea, showing migration patterns similar to those of TP-G2GG-47 and TP-G3GGG-48 (Fig. 4A). However, TP-G3GGC-48 showed a different migration pattern from that of TP-G3GGU-48 and differed from those of TP-G2G-47 and TP-G3G-48.

Melting temperature analysis showed that TP-G2GC-47 and TP-G2GU-47 each exhibited a single Tm value of approximately 60 °C, whereas TP-G3GGU-48 exhibited

two Tm values, one at approximately 40 °C and the other at 68.8°C (Fig. 4B). In contrast, TP-G3GGC-48 showed a Tm value of 65.8 °C, which was similar to that of TP-G1G-46.

These results are consistent with the electrophoretic mobility data. The Tm values of the fragments are listed in Table 1. For all RNAs, the TAR-side fragments showed Tm values comparable to those of the corresponding full-length RNAs.

Furthermore, comparison of the NMR spectra of the four RNAs with those of their corresponding fragments confirmed that the TAR-side stem was formed in all four RNAs (Fig. 4C; Supplemental Fig. S4, S5). In the NMR spectra of TP-G2GC-47 and TP-G2GU-47, signals corresponding to U104, G105, G1, and G2 were clearly observed, whereas these signals were not detected in the spectra of the other RNAs. Because these signals originate from base pairs at the boundary between the TAR and PolyA stems, these results indicate that the two stems are connected in TP-G2GC-47 and TP-G2GU-47, as observed for TP-G1G-46.

**Structure determination**

The solution structures of TP-G1G-46, TP-G2GC-47, and TP-G2GU-47 were determined. Base-pairing restraints were applied to eight base pairs on the UUCG loop side of the TAR stem and to six base pairs on the polyA stem. Additional restraints were applied to three base pairs, including the 5′ terminus in TP-G1G-46; to G1–C58, G2–C57, and U4–A56 in TP-G2GC-47; and to the three base pairs, including the 5′ terminus, in TP-G2GU-47. The formation of these base pairs was confirmed by NMR analysis. RNA A-form backbone restraints were also applied to the base-paired regions. In addition, RNA A-form restraints were also applied to the backbone between C58 and C59. Two hydrogen-bond restraints were applied to the UUCG and GAAA loops. Structural calculations were performed using the distance restraints obtained from the NOESY spectra. For RNAs with mismatches, additional distance restraints were applied to promote stacked

structures around the bulged regions (Supplemental Table S1). The restraints and structural statistics are summarized in Supplemental Table S1. Fig. 5 shows the three lowest-energy structures used in the MD simulations. In each structure, the TAR and PolyA stems formed continuous stems.

**MD simulations**

MD simulations were performed to examine the dynamics of each RNA. For TP-G1G-46, TP-G2GC-47, and TP-G2GU-47, 200-ns MD simulations were carried out for each of the three lowest-energy structures. The base pairs in the PolyA and TAR stems were maintained throughout the 200-ns simulations, indicating that the PolyA and TAR stem structures were retained during this period (Supplemental Figs. S6–S8). The distance between the G residue at the 5′ terminus and G105 at the 3′ terminus did not increase substantially (Fig. 6). Conformational stability was supported by small conformational fluctuations observed in all regions, except the loops and bulged residues (Fig. 7). These results were consistent with the $T_m$ data. Thus, the structure in which the TAR and PolyA stems are connected was stably maintained.

## DISCUSSION

The present study demonstrated that, among HIV-1 genomic RNAs with mutations at the 5′-terminus, G2GC and G2GU form stable structures, whereas G3GGC and G3GGU form relatively flexible structures (Fig. 8). Previous studies investigating the structural and functional roles of these transcripts revealed that G1G forms a stable structure and functions as genomic RNA, whereas G3GGG forms a flexible structure and functions as mRNA [1–4]. Based on these findings, G2GC and G2GU, which form stable structures, preferentially function as genomic RNA and increase the production of viral genomes, whereas G3GGC and G3GGU, which form rather flexible structures, preferentially

function as mRNA. However, the fractions of G2GC and G2GU in viral particles were slightly lower than those of G3GGC and G3GGU [5]. This finding suggests that G2GC and G2GU may not function efficiently as genomic RNA, and that G3GGC and G3GGU may also be incorporated into viral particles as genomic RNA.

Yoshida et al. [5] showed, using infection experiments with mutants in which the GGG sequence at the transcription start site was replaced with GGC or GGU, that the mutant sequences were transmitted during reinfection and that new proviruses were generated. However, it remained unclear what structures were formed by RNAs transcribed from proviruses retaining these mutations and what roles these RNAs play in subsequent viral replication. In the present study, G2GC and G2GU formed stable structures in which the TAR and polyA stems were continuous, similar to G1G. These results suggest that these RNAs function as genomic RNAs. In contrast, G3GGU formed a structure similar to that of G3GGG, whereas G3GGC showed partially distinct behavior and exhibited structural features different from those of G2GC. Taken together, these findings suggest that, as indicated by Yoshida et al., mutation-containing RNAs may retain their mutations during subsequent infection cycles and function similarly to wild-type RNAs.

In this study, structural analyses of mutation-containing RNAs clarified how their structural characteristics are determined by the transcription start site. We are currently further analyzing reverse transcription reactions to determine how HIV-1 reverse transcriptase tolerates mismatches and proceeds with DNA synthesis during reinfection with these RNAs.

## METHODS

### Preparation of RNA samples for NMR measurements

Secondary structures were predicted using the program vs_subopt [6] with a Kuhn length

of 7. The nucleotide sequences of the fragments used for NMR measurements are summarized in Supplemental Table S1. RNA samples were purchased from Hokkaido System Sciences. Each sample was dissolved in water, after which NMR buffer and $D_2O$ were added to obtain a final solution containing 20 mM sodium phosphate buffer (pH 6.5), 50 mM NaCl, and 5% $D_2O$.

**PAGE experiment**

The conformation of each RNA sample was analyzed by 10% PAGE in the presence and absence of 7 M urea. Conventional Tris-borate-ethylenediaminetetraacetic acid buffer was used for electrophoresis. The pre-stained Marker for Small RNA Plus (BioDynamics Laboratory Inc.) was used as a size marker for urea-containing PAGE. The gels were then stained with Toluidine Blue O.

**UV melting experiment**

UV melting curves were measured for the RNA fragments, and the melting temperatures (Tm) were obtained. Each RNA sample was prepared in 20 mM sodium phosphate buffer (pH 6.5) containing 50 mM NaCl. Each sample was prepared to a final volume of 115 µL, with an absorbance at 260 nm of 0.2–0.3 at 25 °C. Melting curves at 260 nm were measured with a V-730BIO UV–Vis spectrophotometer (JASCO Corporation) from 25 to 95 °C at a rate of 1 degree/min. Measurements were repeated three times for each RNA fragment. The Tm values were calculated using the program Spectra Manager (JASCO Corporation). The Tm values were further confirmed by curve fitting with the second-order function for each peak in the first derivatives of the UV melting curves using a custom-developed program.

**Measurements and analyses of NMR spectra**

All NMR spectra were measured using an AvanceNeo 600 NMR spectrometer (Bruker Biospin). The sample concentration was 0.15 mM, and the measurement temperature was 288 K. Water signals were suppressed using a jump-and-return pulse [7] with a delay of 50 μs for the one-dimensional imino proton spectra. For the other spectra, a 3-9-19 pulse [8] was used for water suppression, with delays of 70 μs for measurements covering all regions and 139 μs for measurements focusing on the CH region. NMR spectra were measured and analyzed using a conventional method [9]. Homonuclear Hartmann–Hahn spectroscopy (HOHAHA) spectra were measured with a mixing time of 50 ms, and Nuclear Overhauser Effect Spectroscopy (NOESY) spectra were measured with a mixing time of 200 ms. For measurements of the HOHAHA spectra, the clean-HOHAHA method was used [10]. The observed data were processed by TopSpin (Bruker Biospin) and analyzed by Sparky [11]. A half sine-bell or Gaussian window function was used.

**Structure calculations**

For the determination of the structure of TP-G1G-46, 121 nuclear Overhauser effect (NOE) distance restraints for non-exchangeable protons were obtained using the NOESY spectrum (a mixing time of 200 ms) measured at 288 K in $H_2O$ (5% $D_2O$). Twenty-three distance constraints to maintain hydrogen bonds were used for base pairs for which formation was confirmed by imino proton analysis. Constraints for maintaining the planarity of the base pairs were also used. Sugar pucker was judged by the signal between H1′ and H2′ in the HOHAHA spectra measured with a mixing time of 50 ms. Dihedral constraints for sugar pucker and the RNA-A stem structure, including the *anti*-conformation, were used for the stem regions. Structure calculations were performed 100 times by CNS_SOLVE [12] with the protocol described previously [13] to obtain 15 accepted structures. The 10 structures with the lowest energies were selected, and a

minimized average structure was calculated. Molecular images were produced with UCSF Chimera [14]. Atomic coordinates and NMR information for the reported structure have been deposited with the Protein Data Bank under accession number 46KZ and BMRB ID 36879 for TP-G1G-47.

The solution structures of TP-G2GC-47 and TP-G2GU-47 were also calculated using the same method as described above. The NMR signals corresponding to the bulged region were broadened, and inter-residue NOEs could not be observed. To prepare the initial models for the MD simulations, the formation of stacked base pairs in the mismatch region for TP-G2GC-47 and TP-G2GU-47, as indicated in Fig. 2, was assumed. Restraints for hydrogen bonding and base-pair planarity were introduced for these regions, and three additional restraints between imino protons in these regions were introduced. Thus, these were treated as the model structures to be used for MD analysis and were not deposited as the experimental structures.

**Molecular dynamics simulation**

Molecular dynamics (MD) simulations were performed using AMBER22 [15] according to a protocol similar to that described previously [13]. For each of TP-G1G-47, TP-G2GC, and TP-G2GU, the three lowest-energy structures were used as the initial structures. The RNA.OL3 force field was used. Productive simulations under constant-volume conditions without positional restraints were performed for 200 ns (200,000,000 steps). The trajectories from the productive simulations were processed by the program ptraj in the AMBER suite. Structural fluctuations were evaluated by superimposing each structure in the trajectory onto the initial structure of the corresponding MD run. For each residue, values corresponding to the crystallographic B-factor were calculated using the relationship $B = 8\pi^2 RMSF^2/3$.

## REFERENCES


1. Masuda, T. et al. Fate of HIV-1 cDNA intermediates during reverse transcription is dictated by transcription initiation site of virus genomic RNA. *Sci. Rep.* **5**, 17680 (2015).
2. Kharytonchyk, S. et al. Transcriptional start site heterogeneity modulates the structure and function of the HIV-1 genome. *PNAS* **113**, 13378–13383 (2016).
3. Brown, J. D. et al. Structural basis for transcription start site control of HIV-1 RNA fate. *Science* **368**, 413–417 (2020).
4. Obayashi, C. M. et al., Influence of the 5′-terminal sequences on the 5′-UTR structure of HIV-1 genomic RNA *Sci. Rep.* **11**, 10920 (2021).
5. Yoshida, T. et al. HIV-1 RNAs whose transcription initiates from the third deoxyguanosine of GGG tract in the 5′ long terminal repeat serve as a dominant genome for efficient provirus DNA formation, *J. Virol.* **98**, e01825-23 (2024).
6. Dawson, W. et al. A new entropy model for RNA: part III. Is the folding free energy landscape of RNA funnel shaped? *J Nucleic Acids Investigation* **5**: 2652 (2014).
7. Plateau, P. & Guéron, M. Exchangeable proton NMR without base-line distortion, using new strong-pulse sequences. *J. Am. Chem. Soc.* **104**, 7310-7311 (1982).
8. Piotto, M. et al. Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions. *J. Biomol. NMR* **2**, 661–665 (1992).
9. Sakamoto, et al. *NMR Studies on RNA in Experimental Approaches of NMR Spectroscopy* (Springer, 2018).
10. Griesinger, C. et al. Clean TOCSY for $^{1}$H Spin System Identification in Macromolecules, *J Am. Chem. Soc.* **110**, 7870–7872 (1988).
11. Goddard, T. D. & Kneller, D. G. SPARKY 3. University of California, San Francisco (2008).
12. Brünger, A. T. et al. Crystallography and NMR System (CNS): A New Software

System for Macromolecular Structure Determination. *Acta Cryst* **D54**, 905-921 (1998).

13. Otsu, M. et al. Solution structure of a reverse transcriptase recognition site of a LINE RNA from zebrafish. *J. Biochem.* **162**, 279–285 (2017).
14. Pettersen, E. F. et al. UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
15. Case, D. A. et al. Amber 2023, University of California, San Francisco (2023).

**Acknowledgments**

This research was supported by AMED under Grant Number JP24fk0410064.

**Author contributions**

YK: Investigation, Formal analysis, Visualization, and Writing – original draft.

GK: Conceptualization, Investigation, Formal analysis, and Writing – original draft and review & editing.

TY: Conceptualization, Formal analysis, and Writing –review & editing.

YS: Conceptualization, Writing –review & editing, Supervision, and Project administration.

**Competing interests**

The authors declare no competing interests.

**Supplementary Data**

Supplemental material is available for this article.

**Data availability**

Atomic coordinates and NMR information for the reported structure have been deposited with the Protein Data Bank under accession number 46KZ and BMRB ID 36879 for TP-G1G-47. The structures of TP-G2GC-48 and TP-G2GU-48 were determined with additional distance restraints and, thus, these were treated as the model structures to be used for MD analysis and were not deposited as the experimentally determined structures.

**Figure Legends**

**Fig. 1.** Mutations introduced into the HIV-1 provirus. (**A**) Transcription start sites of the HIV-1 provirus and resulting transcripts. Three types of transcripts, G1G, G2GG, and G3GGG, are generated from the GGG sequence in the provirus. (**B**) Replication of HIV-1 with mutations at the transcription start site, as analyzed by Yoshida et al. [5]. Mutations introduced into the starting provirus were detected in newly generated provirus with fractions of 71% and 83% for GGC and GGU, respectively. Because the fractions of the G1 forms, G1C and G1U, in the viral particles produced were relatively low, the structures of G2GC, G3GGC, G2GU, and G3GGU were analyzed in this study.

**Fig. 2.** Model RNAs used in this study. (**A**) Schematic drawing of the secondary structure of the HIV-1 5′-UTR. The box indicates the region used to design the model RNAs. (**B**) Secondary structure of the model RNAs: TP-G1G-46 and TAR-G1G-29. TP and TAR indicate models consisting of the TAR-PolyA stem and TAR stem, respectively. Five base pairs, U10-U14/A47-A51, were added to the model RNAs used in the previous study [4] to increase the stability of the TAR stem region even with the mutations. (**C**) Sequences of all model RNAs used in this study. For each model, the first G residue (green) corresponds to the $m^7G$ moiety of the Cap structure. G residues corresponding to the GGG regions in the provirus are indicated in blue. C or U residues transcribed from the

mutated residues are indicated in red.

**Fig. 3.** Analysis of RNAs without mutations. (**A**) PAGE with 7 M urea (left) and without urea (right). Lane 1-3: TP-G1G-46, TP-G2GG-47, and TP-G3GGG-48, respectively. M indicates the length marker of 100, 75, 50, 40, 30, and 20 bases. (**B**) First derivatives of the UV melting curves. Red solid line: TP-G1G-46; green dash-dotted line: TP-G2GG-47; green dashed line: TP-G3GGG-48. (**C**) NMR spectra of the model RNAs. Signals corresponding to base pairs at the junction region between the TAR and PolyA stems are shown in bold, and signals corresponding to the two A-U base pairs, U13 and U14, are also indicated.

**Fig. 4.** Analysis of RNAs with mutations. (**A**) PAGE with 7 M urea (left) and without urea (right). Lane 1-4: TP-G2GC-47, TP-G2GU-47, TP-G3GGC-48, and TP-G3GGU-48, respectively. M indicates the length marker. (**B**) First derivatives of the UV melting curves. Red solid line: TP-G2GC-47, red dashed line: TP-G2GU-47, green dash-dotted line: TP-G3GGC-48, green double-dashed line: TP-G3GGU-48. C: NMR spectra of the model RNAs. Signals corresponding to base pairs at the junction region between the TAR and PolyA stems are shown in bold, and signals corresponding to the two A-U base pairs, U13 and U14, are also indicated. Parentheses indicate tentative assignments.

**Fig. 5.** Solution structures of model RNAs predicted to function as genomic RNAs. The three lowest-energy structures used for the MD simulations were superimposed. Green: 5′ G residue, blue: G residue(s) corresponding to the GGG region, pink: the mutated residue, orange: C5, magenta: G105, and cyan: U106. For TP-G2GC-47 and TP-G2GU-47, the formation of base pairs in the mismatch regions, as indicated in Fig. 2, was assumed, and additional restraints between imino protons in these regions were

introduced because the signals in these regions were broadened and inter-residue NOEs could not be observed. Only the structures of TP-G1G-46 have been deposited in the Protein Data Bank as solution structures.

**Fig. 6.** Distances between the H1 of the 5′-G residues and H1 of G105 during the MD simulations. TP-G1G-46 (**A**), TP-G2GC-47 (**B**), and TP-G2GU-47 (**C**). Blue, lowest-energy structure; orange, second-lowest-energy structure; green, the third-lowest-energy structure.

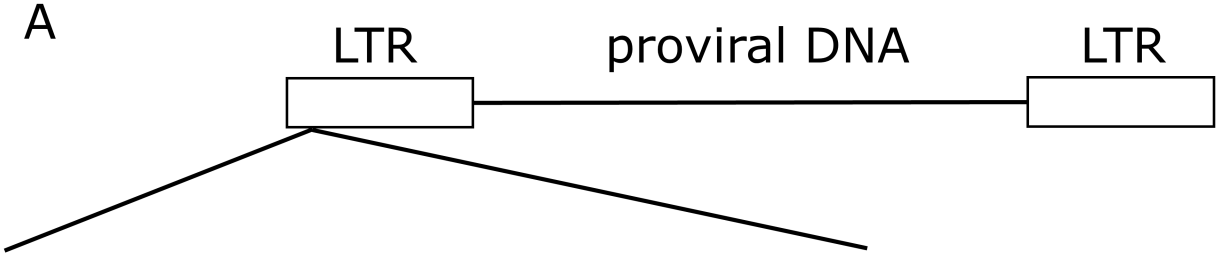
**Fig. 7.** Structural fluctuations of residues during the MD simulations. TP-G1G-46 (**A**), TP-G2GC-47 (**B**), and TP-G2GU-47 (**C**). Blue, the lowest-energy structure; orange, the second-lowest-energy structure; grey, the third-lowest-energy structure. Each histogram starts with the bulged C5 residue and ends with U4. Corresponding regions and sequences are indicated at the top of each histogram. For TP-G2GC-47, the value for the fourth A residue in the GAAA loop in the third-lowest-energy structure was 2428.

**Fig. 8.** Schematic drawing of the structures of the TAR-PolyA junction regions. Left: RNAs forming a continuous TAR-PolyA stem, which are likely to function as genomic RNAs. Right: RNAs forming independent TAR and PolyA stems, which are likely to function as mRNA. Green: 5′ G residue, blue: G residue(s) corresponding to the GGG region, red: mutated residue, orange: C5, magenta: G105, and cyan: U106.

**Table 1. $T_m$ values for model RNAs**

| | <50 | 50-65 | >65 |
|---|---|---|---|
| TP-G1G-46 | | 62.7 | |
| TP-G2G-47 | 44.2 | | 69.6 |
| TP-G3G-48 | 41.4 | | 68.6 |
| TP-G2GC-47 | | 59.8 | |
| TP-G2GU-47 | | 62.8 | |
| TP-G3GGC-48 | | | 65.8 |
| TP-G3GGU-48 | 40.0 | | 68.8 |
| TAR-G1G-29 | | 61.2 | |
| TAR-G2G-32 | | | 67.6 |
| TAR-G3G-33 | | | 67.2 |
| TAR-G2GC-30 | | 60.2 | |
| TAR-G2GU-30 | | 60.3 | |
| TAR-G3GGC-33 | | 62.5 | |
| TAR-G3GGU-33 | | | 66.9 |

A

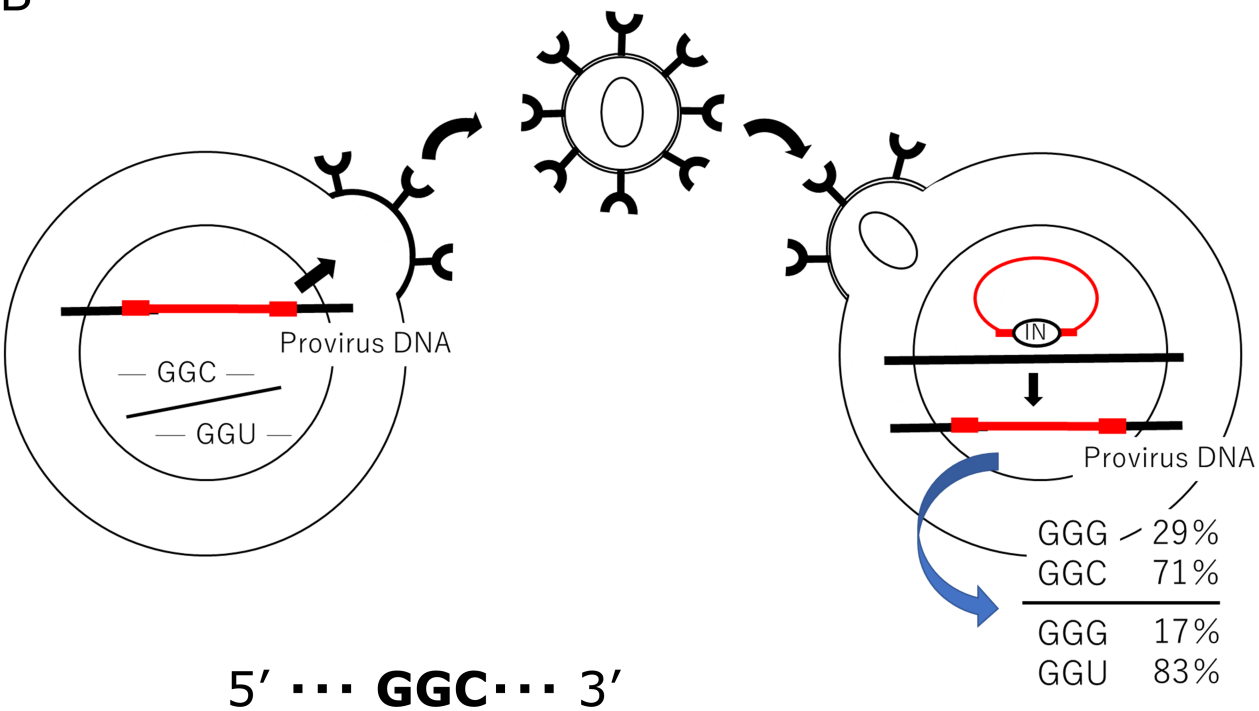



three transcription start sites

5′ ··· **GGG**··· 3′

- G1G — GpppG···
- G2GG — GpppGG···
- G3GGG — GpppGGG···

B



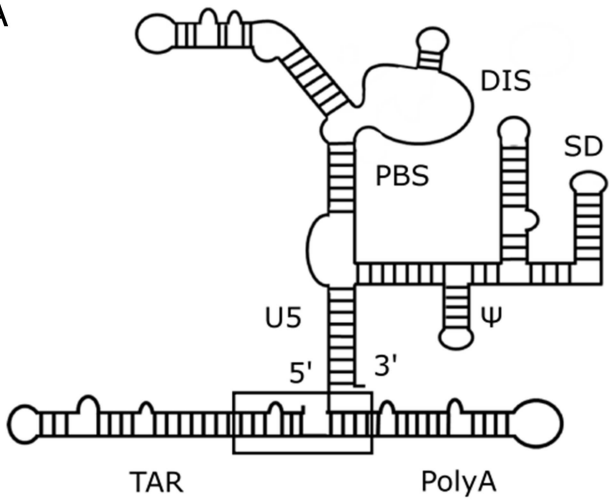

5′ ··· **GGC**··· 3′

- G2GC — GpppGC···
- G3GGC — GpppGGC···

5′ ··· **GGU**··· 3′

- G2GU — GpppGU···
- G3GGU — GpppGGU···

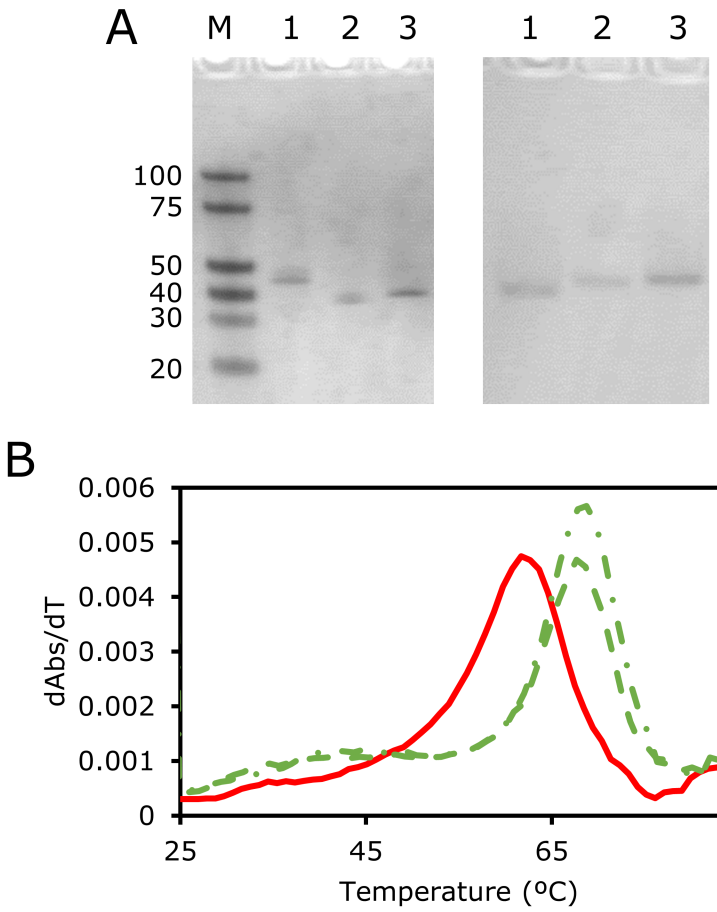
A

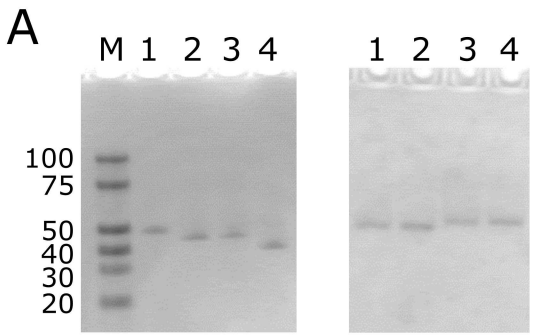


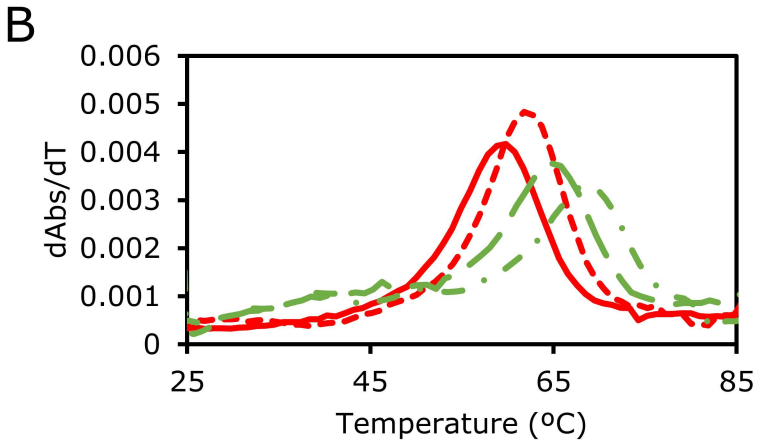

B

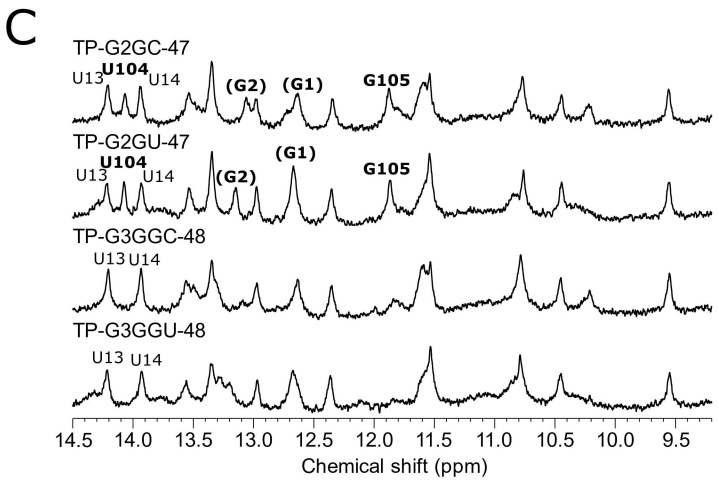
C

```
                2                   47                  100
TP-G1G-46        GGUCUCUCUGGUUuucgAACUAGGGAACCCACUGCgaaaGUAGUGU
TAR-G1G-29       GGUCUCUCUGGUUuucgAACUAGGGAACC
TP-G2GG-47     GGGUCUCUCUGGUUuucgAACUAGGGAACCCACUGCgaaaGUAGUGU
TAR-G2GG-32    GGGUCUCUCUGGUUuucgAACUAGGGAACCCA
TP-G2GC-47     GGCUCUCUCUGGUUuucgAACUAGGGAACCCACUGCgaaaGUAGUGU
TAR-G2GC-30    GGCUCUCUCUGGUUuucgAACUAGGGAACC
TP-G2GU-46     GGUUCUCUCUGGUUuucgAACUAGGGAACCCACUGCgaaaGUAGUGU
TAR-G2GU-30    GGUUCUCUCUGGUUuucgAACUAGGGAACC
TP-G3GGG-48  GGGGUCUCUCUGGUUuucgAACUAGGGAACCCACUGCgaaaGUAGUGU
TAR-G3GGG-33 GGGGUCUCUCUGGUUuucgAACUAGGGAACCCA
TP-G3GGC-48  GGGCUCUCUCUGGUUuucgAACUAGGGAACCCACUGCgaaaGUAGUGU
TAR-G3GGC-33 GGGCUCUCUCUGGUUuucgAACUAGGGAACCCA
TP-G3GGU-48  GGGUUCUCUCUGGUUuucgAACUAGGGAACCCACUGCgaaaGUAGUGU
TAR-G3GGU-33 GGGUUCUCUCUGGUUuucgAACUAGGGAACCCA
```

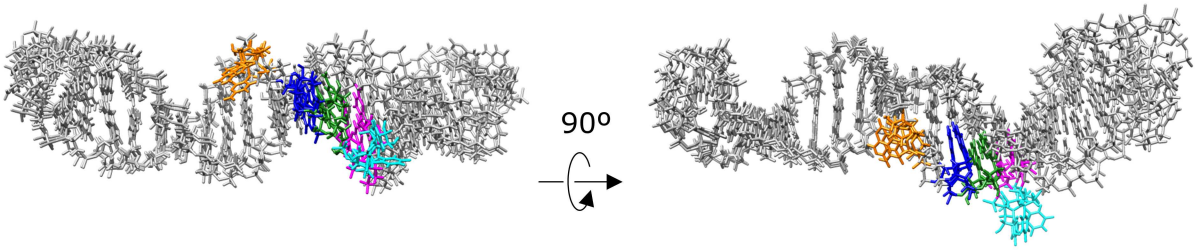

A
M
1
2
3
1
2
3
100
75
50
40
30
20
B
0.006
0.005
0.004
0.003
0.002
0.001
0
dAbs/dT
25
45
65
85
Temperature (ºC)
C
TP-G1G-46
U13
U104
U14
G3
G2
G105
TP-G2G-47
U13 U14
TP-G3G-48
U13 U14
14.5
14.0
13.5
13.0
12.5
12.0
11.5
11.0
10.5
10.0
9.5
Chemical shift (ppm)


A
M 1 2 3 4
1 2 3 4
100
75
50
40
30
20

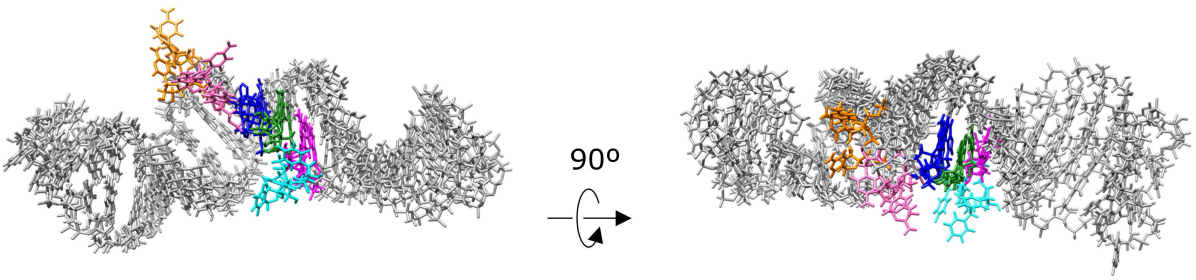

B
0.006
0.005
0.004
0.003
0.002
0.001
0
dAbs/dT
25
45
65
85
Temperature (ºC)


C
TP-G2GC-47
U13 U104 U14
(G2)
(G1)
G105
TP-G2GU-47
U13 U104 U14
(G2)
(G1)
G105
TP-G3GGC-48
U13 U14
TP-G3GGU-48
U13 U14
14.5 14.0 13.5 13.0 12.5 12.0 11.5 11.0 10.5 10.0 9.5
Chemical shift (ppm)

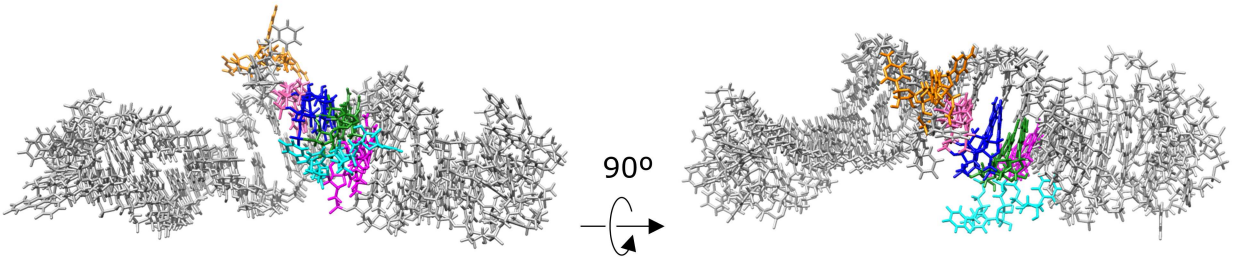
TP-G1G-46

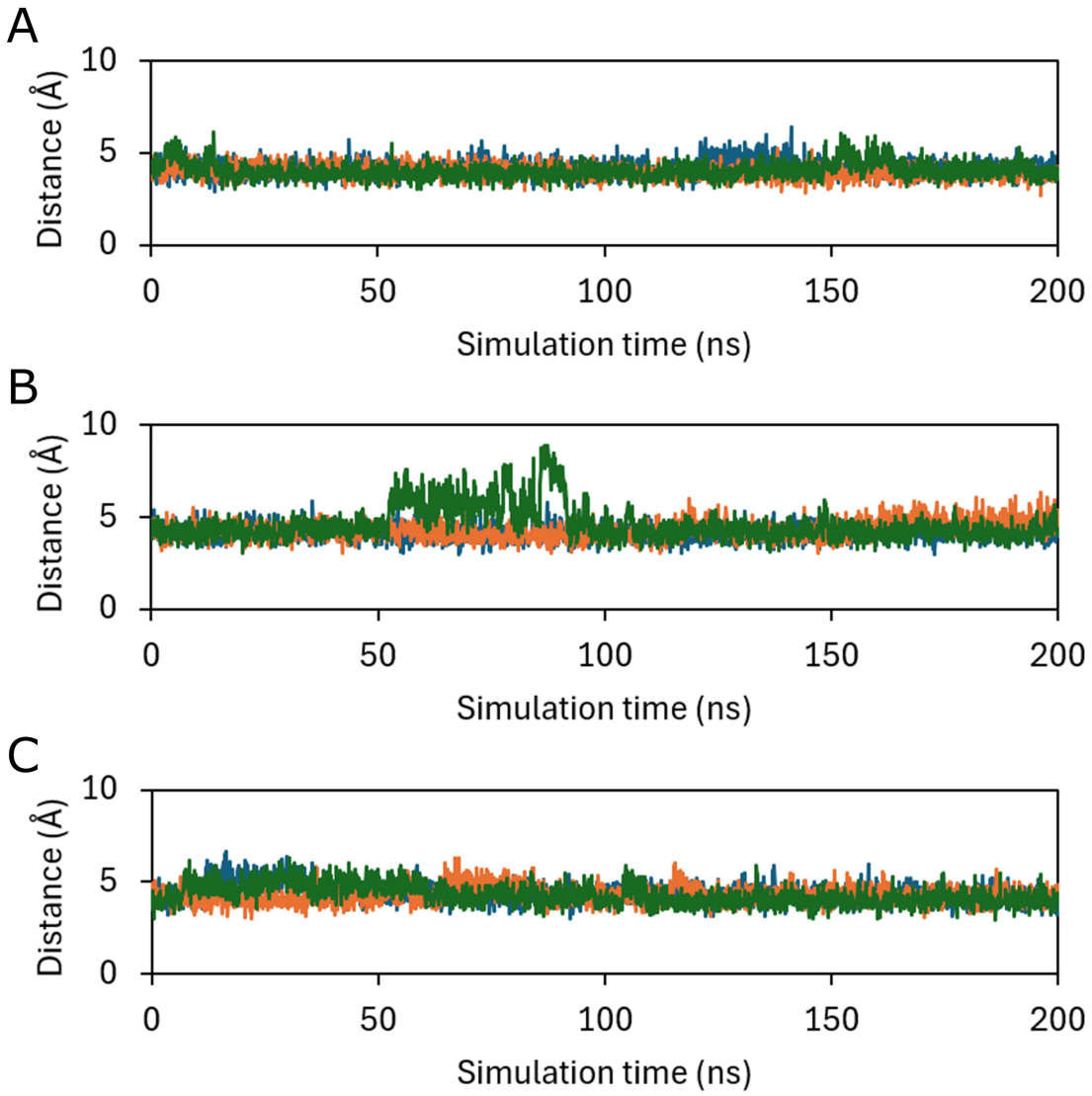


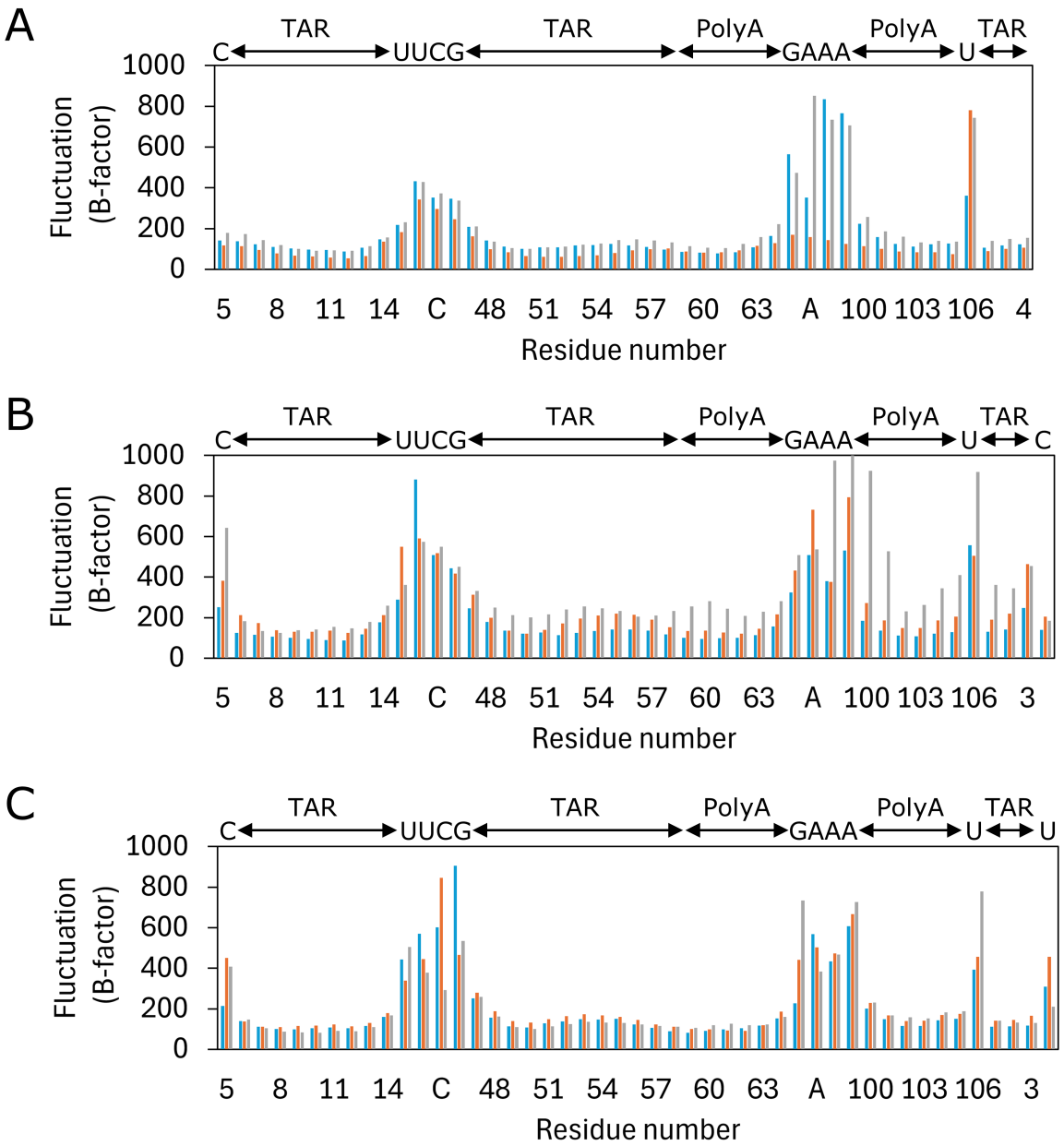

TP-G2GC-47



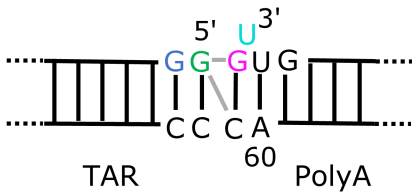

TP-G2GU-47



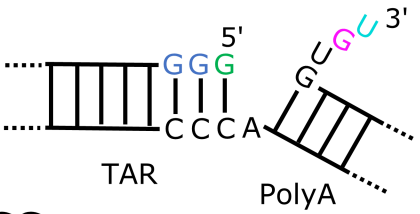

A
Distance (Å)
10
5
0
0
50
100
150
200
Simulation time (ns)
B
Distance (Å)
10
5
0
0
50
100
150
200
Simulation time (ns)
C
Distance (Å)
10
5
0
0
50
100
150
200
Simulation time (ns)


A
TAR
TAR
PolyA
PolyA
TAR
C
UUCG
GAAA
U
Fluctuation (B-factor)
1000
800
600
400
200
0
5 8 11 14 C 48 51 54 57 60 63 A 100 103 106 4
Residue number
B
TAR
TAR
PolyA
PolyA
TAR
C
UUCG
GAAA
U
C
Fluctuation (B-factor)
1000
800
600
400
200
0
5 8 11 14 C 48 51 54 57 60 63 A 100 103 106 3
Residue number
C
TAR
TAR
PolyA
PolyA
TAR
C
UUCG
GAAA
U
U
Fluctuation (B-factor)
1000
800
600
400
200
0
5 8 11 14 C 48 51 54 57 60 63 A 100 103 106 3
Residue number

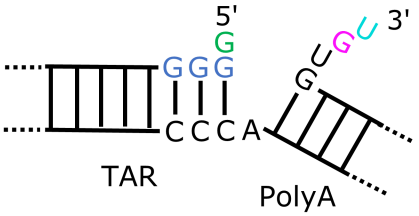
G1G



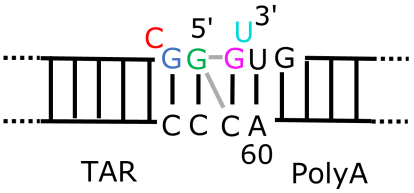

G2GG



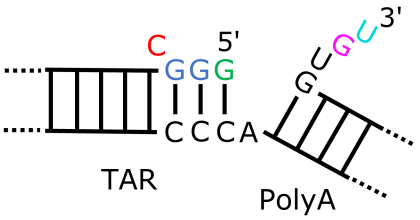

G3GGG



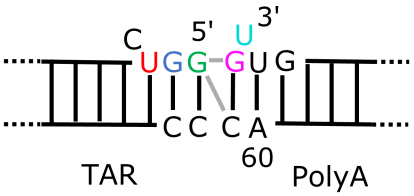

G2GC



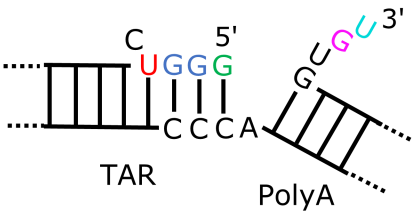

G3GGC




G2GU




G3GGU



## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterials.pdf