The effect of hairpin loop on the structure and gene expression activity of the long-loop G-quadruplex

Subramaniyam Ravichandran, Maria Razzaq, Nazia Parveen, Ambarnil Ghosh and Kyeong Kyu Kim

Department of Precision Medicine, Graduate School of Basic Medical Science (GSBMS), Institute for Antimicrobial Resistance Research and Therapeutics, Sungkyunkwan University School of Medicine, Suwon 16419, Republic of Korea

Received August 05, 2020; Revised August 10, 2021; Editorial Decision August 11, 2021; Accepted August 24, 2021

ABSTRACT

G-quadruplex (G4), a four-stranded DNA or RNA structure containing stacks of guanine tetrads, plays regulatory roles in many cellular functions. So far, conventional G4s containing loops of 1–7 nucleotides have been widely studied. Increasing experimental evidence suggests that unconventional G4s, such as G4s containing long loops (long-loop G4s), play a regulatory role in the genome by forming a stable structure. Other secondary structures such as hairpins in the loop might thus contribute to the stability of long-loop G4s. Therefore, investigation of the effect of the hairpin-loops on the structure and function of G4s is required. In this study, we performed a systematic biochemical investigation of model G4s containing long loops with various sizes and structures. We found that the long-loop G4s are less stable than conventional G4s, but their stability increased when the loop forms a hairpin (hairpin-G4). We also verified the biological significance of hairpin-G4s by showing that hairpin-G4s present in the genome also form stable G4s and regulate gene expression as confirmed by in cellulo reporter assays. This study contributes to expanding the scope and diversity of G4s, thus facilitating future studies on the role of G4s in the human genome.

INTRODUCTION

G-quadruplexes (G4s) are non-B DNA structures characterised by the presence of stacked G-quartets, with each G-quartet containing four guanines bonded by Hoogsteen base pairs in a planar structure. G4s are formed from a single DNA strand in which consecutive G-repeats (G-runs) are separated by short oligonucleotides (∼1 nucleotide). Upon folding into G4s, the G-runs form G-quartets and the intervening short oligonucleotides form loops that connect the G-quartets. Based on these loops, G4s can adopt a variety of folding topologies. In the case of a parallel intramolecular G4, the top and bottom G-quartets are connected by a propeller or chain-reversal loop. Previous reports showed that the stability and conformational polymorphism of G4 DNA are governed by the length and nucleotide composition of the loop. When G4s have a loop 1–10 nucleotides (nt) in length, G4s adopt a parallel conformation. Based on these rules, several prediction tools have been developed to identify potential G4-forming sequences. Many recent studies have indicated that loops that are longer than 10 nt can also form parallel G4s. High-throughput sequencing of G4s (G4-Seq) was performed based on the principle that G4s can act as physical barriers for polymerase progression. By performing the sequencing under G4-stabilizing conditions, the polymerase pause sites were found to overlap with ~700,000 G4-forming sequences, with over 70% consisting of long loops (>10 nt) and bulges. The study of G4s with longer loops is further complicated by the abundance of nested secondary structures such as hairpins. The ability of hairpins to form in short lengths of single stranded DNA and the prevalence of inverted repeats in the human genome make hairpins the most likely type of secondary structure that are easily formed within loop regions of G4; these can also be called hairpin-forming loops. Considering the evidence that G4s containing a hairpin in the loop (hairpin-G4s) are present in the genome, intensive and systematic investigation on hairpin-G4s are necessary for the comprehensive understanding of G4s function in the genome.

A previous report proposed that hairpin-G4s are more stable than G4s containing unstructured loops based on a thermal stability study. However, this study only considered the 6 nt-stem region for the long loop of the hairpin-G4 and investigated the effect of the hairpin-forming loop...
present in the central position (loop 2). Subsequent studies revealed that genome-wide extraction of the putative hairpin-G4s from the human genome was carried out using the schema G_{(3–6)}N_{(1–20)}G_{(3–6)}N_{(1–20)}G_{(3–6)}N_{(1–20)}G_{(3–6)} because they observed that the 20 nt loop length represented 45% of all hairpin-G4s in chromosome 1 (15). In this genome-wide search, 48,508 hairpin-G4 sequences in the promoter regions of 12,315 genes were found. However, considering the correlation between the stability and the number of base pairs in the hairpin, it is plausible to speculate that a loop longer than 20 nt might form more stable hairpins, which contribute to the structure and function of the long-loop G4s. Therefore, it is necessary to more systematically investigate G4s with long loops, including G4s with loops longer than 20 nt, for the comprehensive understanding of the role of non-canonical DNAs in the genome.

In the current study, we extend the existing knowledge on hairpin-G4s (Figure 1) by evaluating long-loop G4s with various loop sizes (13, 23 and 33 nt) and examining the effect of hairpin formation on G4 stability and function. For this purpose, we designed model hairpin-G4s containing one hairpin with various sizes and structures. We used circular dichroism (CD) and nuclear magnetic resonance (NMR) studies to analyse the effect of hairpin-forming propensity, size of the stem-flanking region, and location of the hairpin on hairpin-G4 formation and thermal stability. For identification of hairpin formation in the G4 structure, we developed a SYBR Green binding assay that differentiates the hairpin-loop from the unstructured-loop. Finally, we searched the human genome for hairpin-G4–forming sequences in promoter regions and evaluated their functional effects on gene regulation through in cellulo reporter assays and in vitro polymerase stop assays.

MATERIALS AND METHODS

Oligonucleotides used for this study

All oligonucleotides were purchased from Cosmo Genetech and Macrogen. The oligonucleotides used in this study are listed in Table 1.

Prediction of hairpin structure and free energy ($\Delta G_h$)

The predicted structure and thermodynamic parameters ($\Delta G_h$) for the loops with hairpin-forming regions were obtained using the UNAFold server (http://unafold.rna.albany.edu/?q=DINAMelt/Quickfold) under 100 mM Na$^+$ and 1 mM Mg$^{2+}$ at 37°C.

Circular dichroism (CD) spectroscopy

The CD experiments were performed on a Jasco J-810 CD spectropolarimeter fitted with a Jasco CDS-426F Peltier temperature controller and a 1 mm quartz cuvette (Hellma) with a 220 μl reaction volume. Oligonucleotides (15 μM) were suspended in a buffer containing 10 mM HEPES (pH 7.5) with 100 mM KCl and varying concentrations of MgCl$_2$ (1 and 10 mM), followed by denaturation at 95°C for 5 min. After cooling to room temperature over 2 h, all spectra were measured between 220–320 nm with a speed of 100 nm/min, 1 nm data pitch, and a bandwidth of 1 nm at a constant temperature of 298 K. The readings of three accumulations were averaged and recorded. For CD thermal melting analysis, the samples were heated from 15 to 90°C at a rate of 0.2°C/min and the ellipticity was recorded at the wavelength showing maximum ellipticity every 0.2°C. After subtracting the spectrum of the buffer, the data was converted to the fraction folded and the melting temperature ($T_m$) was obtained at the point where 50% of the fraction was in the unfolded state. Data plotting and curve fitting were performed on Sigma Plot 12.5.

SYBR Green analysis of hairpin-G4s

For fluorescence spectrum measurement, 5 μl of 100 μM pre-formed G4 DNA in buffer containing 10 mM HEPES (pH 7.5) with 100 mM KCl and 1 mM MgCl$_2$ was mixed with 15 μl of 1× SYBR Green I dye for a final DNA concentration of 25 μM. Fluorescence spectra were recorded in 96-well black bottom plates between 505–600 nm after excitation at 475 nm on a Biotek Synergy Neo multiplate reader.
Table 1. List of oligonucleotides used in this study

| Sample name | Sequence | Length |
|-------------|----------|--------|
| SH-G4       | TGGGTGCTGTTGTCACATGATGGGT | 29     |
| SI-1-G4     | TGGGTGCTGTTGTCACATGATGGGT | 29     |
| SI-2-G4     | TGGGTGCTGTTGTCACATGATGGGT | 29     |
| MH-G4       | TGGGTGCTGTTGTCACATGATGGGT | 29     |
| MI-1-G4     | TGGGTGCTGTTGTCACATGATGGGT | 29     |
| MI-2-G4     | TGGGTGCTGTTGTCACATGATGGGT | 29     |
| LH-G4       | TGGGTGCTGTTGTCACATGATGGGT | 29     |
| LI-1-G4     | TGGGTGCTGTTGTCACATGATGGGT | 29     |
| LI-2-G4     | TGGGTGCTGTTGTCACATGATGGGT | 29     |
| SU-G4 LOOP ONLY | TGGGTGCTGTTGTCACATGATGGGT | 13     |
| MU-G4 LOOP ONLY | TGGGTGCTGTTGTCACATGATGGGT | 23     |
| LU-G4 LOOP ONLY | TGGGTGCTGTTGTCACATGATGGGT | 33     |
| SH-G4 LOOP ONLY | TGGGTGCTGTTGTCACATGATGGGT | 13     |
| MH-G4 LOOP ONLY | TGGGTGCTGTTGTCACATGATGGGT | 23     |
| LH-G4 LOOP ONLY | TGGGTGCTGTTGTCACATGATGGGT | 33     |

Electrophoretic mobility analysis

Electrophoretic mobility of hairpin-G4s and unstructured-loop controls was analyzed on 15% polyacrylamide gels run under native conditions. The G4s were pre-formed for native PAGE at a concentration of 1 μM in buffer containing 10 mM HEPES pH 7.5, 100 mM KCl and 1 mM MgCl2 by heating at 95°C for 5 min and allowing to cool to room temperature over 2 h. 5 μl of preformed G4 was loaded with 1 μl of 6× OrangeRuler Loading Dye on a 15% Native PAGE gel supplemented with 100 mM KCl and run at 80 V for 1 hour at 4°C in 1× TBE. GeneRuler Ultra-Low Range 10 bp ladder (Thermo Scientific) was used for checking the migration. After running the gels were stained using SYBR Gold and imaged on a Bio-Rad XRF Gel Doc system.

1H nuclear magnetic resonance (NMR) analysis

1H NMR experiments were performed at 25°C on a 400-MHz Bruker Avance 400 spectrometer fitted with a general-purpose probe in buffer containing in 100 mM KCl and 1 mM MgCl2 in 10 mM HEPES (pH 7.5) and 10% D2O. The oligonucleotide concentration was set to 0.5 mM and 50 μM DSS was added for referencing the spectra. For temperature-dependent NMR experiments, spectra were recorded at 10°C increments from 20 to 80°C. All spectra were processed using Topspin 3.5 (Bruker).

Prediction of hairpin-G4 formation in the human genome

Step 1. We searched the long-loop G4s with one long loop (10nt < loop length < 40nt) and two short loops staining was carried out using SYBR Gold for 15 min and imaged on a Bio-Rad XRF Gel Doc system.
with one nt each (A/T/C/G) using quadparser (17) in the promoter regions, defined as the regions −500 to +100 bp from the transcription start sites (TSS) present in the hg19 human genome. To observe the distribution of the G4s depending on the loop size, we subdivided the long-loop G4s based on their loop size by 10 nt steps. G4 with loop length 1–10 was also included for comparison. Accordingly, we applied the following schema for searching the G4s in the promoter regions of human genome: G(3)N(1−10/10/30−1)G(3)N(1−10/10)G(3)N(1−10/10/30−1)G(3)N(1−10/10/30−1)G(3). 

**Step 2.** Shortlisting of long-loop G4s in step 1 by overlapping with experimentally identified G4s (14).

**Step 3.** Since we tested long-loop G4s with ‘thymine’ in the short loop position and loop length between 13 and 33 in our model study, we further restricted H-G4s with ‘T’ in the short loop site and with loop length less than or equal to 33 nt.

**Step 4.** Identifying the hairpin-G4s (H-G4s) from the long-loop G4s in Step 3 based on the hairpin folding prediction. Hairpin formation was predicted based on the secondary structure and ΔG<sub>h</sub> value calculated from the UNAFold server (http://unafold.rna.albany.edu/?q=DINAMelt/Quickfold) (18) under the condition of 100 mM Na<sup>+</sup> and 1 mM Mg<sup>2+</sup> by considering that the loop forms a stable hairpin when ΔG<sub>h</sub> value is lower.

**Step 5.** Five H-G4s (MCM4, MSI1, TMCC3, CHST1 and NRBPI) were chosen based on the biological relevance and secondary structure among the H-G4s chosen in Step 3.

**Construction of reporter vectors for in cellulo expression analysis**

Candidate sequences containing hairpin-forming loops and their respective non-hairpin-forming mutants were cloned in the pGL3-Promoter vector. Two types of vectors were constructed: one to analyze the cis-acting effect of the sequences and the other to analyse the in cellulo polymerase-stalling ability. For the cis-acting effect, the hairpin-G4s and control sequences were cloned in the XhoI/NheI sites upstream of the SV40 promoter. To analyse the in cellulo stability-dependent effect, the sequences were cloned into the antisense strand between the HindIII/NcoI sites between the SV40 promoter and luciferase gene based on the rationale that G4 formation can only affect transcription by polymerase stalling (19). A single stranded oligonucleotide (ODN) corresponding to the hairpin-G4 forming sequence and its complementary strand were synthesized and annealed to form double-stranded DNA, which was then cloned into the pLG3-Promoter vector to construct the reporter plasmids. The pRL-TK vector expressing Renilla luciferase was used as internal control. We transfectected 900 ng of reporter plasmid along with 100 ng of internal control plasmid using Turbofect (Thermo Scientific) into 3 x 10<sup>5</sup> HEK293T cells seeded 24 h earlier in six-well plates. After 24 h of transfection, the cells were harvested, and luciferase activity was measured using the Dual-Luciferase Reporter Assay Kit (Promega) according to the manufacturer’s instructions. Firefly luciferase reporter expression (F<sub>Luc</sub>) was first normalised to Renilla expression (R<sub>Luc</sub>) and the F<sub>Luc</sub>/R<sub>Luc</sub> ratio was further normalised to the activity from the mutated (G4 without hairpin) vector.

**In vitro T7 RNA polymerase stop assay**

The T7 RNA polymerase stop assay was performed using the protocol based on the study by Tateishi-Karimata et al. (20). Oligonucleotides (Supplementary Table S2) containing the T7 polymerase binding site and a 35 nt spacer followed by the G4 forming sequence were synthesized and annealed with the complementary strand to form a dsDNA T7 polymerase binding site while leaving the spacer and G4-forming site single-stranded. The list of oligonucleotides used for this study is listed in Supplementary Table S2. The reaction was performed by incubating T7 polymerase and the annealed oligonucleotides for 10 mins in 1 x T7 Polymerase Reaction Buffer (NEB), followed by addition of rNTPs and further incubation at 37°C for 90 mins. The reaction was quenched using Stop Buffer (80% formamide, 10mM MgCl<sub>2</sub> and 0.01% Dextran Blue). The samples were heated for 5 min at 95°C before being loaded on a 10% denaturing PAGE gel with 7M urea and electrophoresed for 45 min at 60°C and stained with SYBR Gold stain (Thermo Scientific). The gels were imaged on the Bio-Rad XRF Gel Doc system.

**Statistical analysis**

Significance was calculated using one-sample t-tests by setting the theoretical mean as 1 for the unstructured loop mutants on GraphPad Prism 6. P-values were indicated based on the program “P ≤ 0.05; **P ≤ 0.01; ***P ≤ 0.001.”

**RESULTS**

**Design of model hairpin-G4s: S(U/I/H)-G4s, M(U/I/H)-G4s, and L(U/I/H)-G4s**

In our quest to establish whether the formation of parallel hairpin-G4s with long loops was feasible, we designed model hairpin-G4s with the sequence ‘TG<sub>1</sub>TG<sub>1</sub>TG<sub>1</sub>X<sub>13/23/33</sub>G<sub>3</sub>T’ in which the third loop (X) forms a hairpin. Three loop lengths were chosen for this purpose and the hairpin-G4s were categorised accordingly as short (S, 13 nt), medium (M, 23 nt), and long (L, 33 nt) loops (Supplementary Figure S1). To evaluate the effect of the hairpin stability on hairpin-G4s, we additionally varied the propensity of hairpin formation in three categories based on the predicted free energy of hairpin formation (ΔG<sub>h</sub>) (Table 2) from the UNAFold server (http://unafold.rna.albany.edu/?q=mfold/DNA-Folding-Form) (18,21). The UNAFold server has been shown to predict the secondary structure of nucleic acids and their free energy very accurately (18,21,22) and is especially useful considering our aim to screen H-G4s from the human genome on a large scale. Accordingly, hairpin-G4s were further characterised as low or unstructured (U-G4s),
are characteristic of G4 formation. We propose that the CD spectra of the current model G4s with high or intermediate propensity hairpin-forming tendencies show higher ellipticity than I-2. In total, 12 model hairpin-G4s with loops of different lengths and hairpin-forming propensities were designed: 13 nt short-loop G4s (SU-G4, SI-1-G4, SI-2-G4 and SH-G4), 23 nt medium-loop G4s (MU-G4, MI-1-G4, MI-2-G4, and MH-G4), and 33 nt long-loop G4s (LU-G4, LI-1-G4, LI-2-G4, and LH-G4) (Table 1).

### Correlation between the stabilities of the hairpin in the loop and the hairpin-G4

CD spectroscopy is a technique widely used to characterise G4 structures. G4s exhibit characteristic peaks between the 220 and 320 nm regions in the CD spectrum (23). Parallel G4s exhibit a peak at ~260 nm and a trough at ~240 nm, as evidenced by the CD spectra of the parallel G4 pu22myc from the C-MYC promoter, which was used as a control (23,24). The CD spectra of model G4s revealed that model G4s with high or intermediate propensity hairpin-forming loops (H-G4s or I-G4s) form parallel-stranded G4s with the 260 and 240 nm peaks (Figure 2A). However, in the case of U-G4s, 1–4 nm spectral shifts were observed, suggesting that the G4 structure is perturbed, possibly due to the increased flexibility of the loop. Consistently, the spectral shifts were more pronounced in MU-G4s and LU-G4s with longer unstructured loops (23 and 33 nt) than SU-G4 (13 nt). We also investigated whether there is any contribution of the long loops to the CD spectra of the long-loop G4 by observing the CD spectra of the loop region only (Supplementary Figure S2). While the spectral signature was found near 250 and 280 nm, there was no CD peak at 262 nm and even ellipticity was almost zero, representing the presence of the unstructured or hairpin DNA (23,25). Therefore, we can propose that the CD spectra of the current model G4s are characteristic of G4 formation.

Next, we performed CD thermal melting analysis, which is widely used for evaluating the stability of the G4 structure (26). We measured the $T_m$ by monitoring the ellipticity change at the 262 nm between 15 and 90°C (Supplementary Figure S3). Regardless of the length of the loop, the $T_m$ of H-G4s was higher than that of U-G4s and I-G4s (Figure 2B and Table 3). The difference in $T_m$ was more pronounced in the hairpin G4s with long (LU-G4 versus SU-G4 or LI-G4) and medium (MH-G4 versus MU-G4 or MI-G4) sized loops than hairpin G4s with short loops (SH-G4 versus SU-G4). We compared the predicted $\Delta G_h$ values of the hairpin with $T_m$ values of hairpin-G4s (Table 3) to examine the correlation between hairpin-G4 stability and hairpin stability. Among the model G4s, H-G4s with the lowest $\Delta G_h$ values showed the highest $T_m$ (SH-G4, MH-G4, and LH-G4), which is expected since the predicted hairpin structures of H-G4s contain perfect base pairing in the stem region (Supplementary Figure S1). However, very little correlation between $\Delta G_h$ and $T_m$ was found in I-G4s and U-G4s, and their $T_m$ values were lower than those of H-G4s. These results suggest that hairpin formation significantly contributes to the stability of hairpin-G4s only when the hairpin contains perfect base pairing, and hairpins with mismatches have lower propensity to stabilize hairpin-G4s. Comparison of the melting and cooling curves of the model G4s showed that the curves were mostly superimposable except in the case of LU-G4 and LH-G4 (Supplementary Figure S4), which show slight hysteresis of 4–5°C.

We also performed temperature-dependent CD measurement similar to the NMR melting studies to evaluate the impact of hairpin formation on the stability of hairpin-G4 structure (Supplementary Figure S5). Our studies showed that the characteristic G4 spectrum of hairpin-G4 (LH-G4) is retained until 60°C, followed by an abrupt decrease in ellipticity. In comparison, its long-loop G4 with unstructured loops (LU-G4), showed a gradual decrease in ellipticity and spectral shift at lower temperature compared to LH-G4. A similar trend was observed in the G4s containing medium (MH-G4 versus MU-G4) and short (SH-G4 versus SU-G4) loops. This indicates that the formation of the hairpin could possibly contribute to the higher thermal stability of hairpin-G4s in contrast to their counterparts with unstructured loops.

| Sample name | Loop sequence | Loop length | Predicted* $\Delta G_h$ | Predicted* $T_m(37^\circ C)$ | Stem length | Hairpin loop length |
|-------------|---------------|-------------|-------------------------|-----------------------------|------------|---------------------|
| SU-G4       | TCTTCTTACATAT | 13          | -0.21                   | 39.6                        | 2          | 5                   |
| SI-1-G4     | TCTGCGGCAGTT  | 13          | 0.09                    | 35.9                        | 2          | 5                   |
| SI-2-G4     | TCTTCGCGACTT  | 13          | 1.38                    | -3.4                        | 1          | 5                   |
| MH-G4       | TTTCAGTATGCGCATACTGACAT | 23 | -5.14                   | 63.9                        | 7          | 5                   |
| MI-1-G4     | TTTCAGAATTGCGCATACTGACAT | 23 | -2.7                    | 54.8                        | 5          | 9                   |
| MI-2-G4     | TTTCAGTTAGGCACACTGACAT | 23 | -1.21                   | 47.7                        | 5          | 7                   |
| LH-G4       | TTTCGATCTGTCTGGCAGACTATACCTGACAT | 33 | -9.94                   | 69.7                        | 15         | 12                  |
| LI-1-G4     | TTTCGATCTGTCTGGCAGACTATACCTGACAT | 33 | -3.23                   | 51.2                        | 10         | 5                   |
| LI-2-G4     | TTTCGATCTGTCTGGCAGACTATACCTGACAT | 33 | -1.38                   | 49.4                        | 4          | 4                   |
| SU-G4       | TCTTCTTACATAT | 13          | No folding              | Determined                  | 0          | 0                   |
| MU-G4       | TCTTCTTATTTATTTCTTTACTACA | 23 | 2.94                    | -47.9                       | 1          | 9                   |
| LU-G4       | TCTTTCTTATTTATTTCTTTACTTTACTTTTCTTT | 23 | 2.63                    | -21.5                       | 2          | 11                  |

*Predictions were carried out using UNAFold DNA-folding form server with parameters 100 mM Na+ and 1 mM Mg2+ at 37°C.
Figure 2. Hairpin-G4s can form stable G4s with higher stability than the mutants with unstructured loops. (A) CD spectra of short (S), medium (M) and long (L) loop-G4s with perfect hairpin loop (H-G4s), intermediate hairpin loop (I1 and I2-G4s) and unstructured loop (U-G4s) showed characteristic G4 peaks at 240 and 260 nm. (B) CD thermal melting analysis showed that G4s with hairpin loops (H/I-G4s) are more stable than the G4 with unstructured loop (U-G4s) in varying Mg2+ concentrations. LI-2-G4 was extremely stable at 10 mM MgCl2, and hence the melting temperature could not be determined. All measurements were taken in 10 mM HEPES buffer (pH 7.5) containing 100 mM KCl and 1 mM MgCl2. The ramp rate for CD melting experiments was set to 0.2 °C/min. Error bars represent ± SD.

Table 3. Comparison between unstructured loop and hairpin-forming loop $\Delta G_h$ and hairpin-G4 and unstructured-loop mutant CD $T_m$

| Sample name | Predicted loop $\Delta G_h$ | CD $T_m$ (°C) |
|-------------|----------------------------|---------------|
| SU-G4       | No folding                 | 69.48 ± 1.17  |
| SH-G4       | −0.21                      | 71.39 ± 1.37  |
| SI-1-G4     | 0.09                       | 69.42 ± 0.14  |
| SI-2-G4     | 1.38                       | 68.67 ± 0.09  |
| MU-G4       | 2.94                       | 65.34 ± 0.03  |
| MH-G4       | −5.14                      | 69.71 ± 0.04  |
| MI-1-G4     | −2.7                       | 66.11 ± 0.28  |
| MI-2-G4     | −1.21                      | 63.27 ± 0.08  |
| LU-G4       | 2.63                       | 62.97 ± 0.29  |
| LH-G4       | −9.94                      | 68.80 ± 0.28  |
| LI-1-G4     | −3.23                      | 63.55 ± 0.50  |
| LI-2-G4     | −1.38                      | 66.26 ± 0.01  |

CD experiments were performed in buffer containing 100 mM KCl and 1 mM MgCl2 in 10 mM HEPES (pH 7.5). $\Delta G_h$ was predicted using UN-AFold DNA-folding form server with parameters 100 mM Na+ and 1 mM Mg2+ at 37°C. The ramp rate of 0.2°C/min was used for CD melting.

Effect of magnesium Ions on hairpin-G4 stability

Hairpin-G4s have both G4 and a hairpin which contains a dsDNA stem region (Figure 1B). Since Mg2+ can stabilize dsDNA (27) as well as G4 (22,23), we tested the effect of Mg2+ on the stability of the hairpin-G4s by measuring the thermal stability of hairpin-G4s in the presence of 1 and 10 mM MgCl2. As expected, we found that the hairpin-G4 stability represented by the $T_m$ values was directly proportional to the concentration of Mg2+ in the buffer (Figure 2B and Supplementary Figure S3). However, the correlation between $T_m$ and $\Delta G_h$ was consistently observed even in the presence of Mg2+. For example, H-G4s showed the highest stability, while the stability of I-G4 and U-G4 were similar regardless of Mg2+ (Figure 2B). Interestingly, in the case of U-G4s, thermal stability was inversely proportional to the length (SU-G4 ≥ MU-G4 > LU-G4), which could be explained by the longer unstructured loop that enhances G4 instability. This stability change became more obvious in the presence of Mg2+ (SU-G4 > MU-G4 > LU-G4) (Figure 2B), which is consistent with the previous reports in which Mg2+ increases the flexibility of a single-stranded DNA loop (28,29) and becomes detrimental to G4 stability (30). In the case of the C-MYC G-quadruplex (pu22myc), we did not observe any drastic change in thermal stability upon addition of 1 mM MgCl2 (Supplementary Figure S6). However, under 10 mM MgCl2, there was significant stabilization of pu22myc (Supplementary Figure S6), which corroborates previous reports that showed stabilization of promoter quadruplexes under high concentrations of Mg2+ (31). This demonstrates that although Mg2+ stabilizes the G4 structure itself, the contribution of Mg2+ to the stability of the hairpin region of the hairpin-G4 is more significant, especially in the case of MH-G4 and LH-G4 (Figure 2B).

Effect of length of the stem-flanking region on hairpin-G4 stability

The stem-flanking region of the hairpin-forming loop connects the hairpin structure with G4. Therefore, the intriguing question is whether the length of the stem-flanking
region influences the stability of hairpin-G4s. Answering this question is necessary to identify the stable hairpin-G4s in the genome. In previous reports, the effect of the stem-flanking region was evaluated using a model hairpin-G4 with a hairpin-forming loop with a 6-bp stem (13). In this study, there was no significant effect of the flanking region on thermal stability of the hairpin-G4. We extended this study to hairpins with a longer stem (13-bp) by further systematic evaluation of the effect of the flanking region on the stability of hairpin-G4 with a longer loop. For this purpose, we introduced new model hairpin-G4s that contain one (L₁H-G₄), two (L₂TH-G₄), or three (L₃TTTH-G₄) thymines on both ends of the hairpin stem, only on the 5’ end (L₅TH-H₄, L₅TH-H₄, L₅TTTH-H₄), or only on the 3’ end (L₃TH-H₄, L₃TTTH-H₄, L₃TTTH-H₄) (Table 1). The CD spectra showed no significant deviation from the parallel G4 structure in all cases (Figure 3A). We evaluated the Tₘ from CD melting curves and found that there was no significant impact of the stem flanking region on the thermal stability of the hairpin-G4 structure (Supplementary Figure S7A). The maximum ΔTₘ between LH-G₄ and LH-G₄ with extended flanking regions (L₁H-G₄/ L₂TH-H₄/ L₃TTTH-H₄ and their 5’ and 3’ counterparts) were only ~2.5°C in the presence of 1mM Mg²⁺ (Figure 3B). The maximum ΔTₘ between the equally distributed (L₅TH-H₄/ L₅TH-H₄/ L₃TTTH-H₄) and unequally distributed (L₅TH-H₄/L₅TH-H₄, L₅TH-H₄/L₅TH-H₄) and L₃TTTH-H₄/L₃TTTH-H₄) flanking regions was only ~2.8°C in the presence of 1 mM Mg²⁺ (Figure 3B), indicating that the distribution of the flanking nucleotides in the hairpin does not significantly affect the stability of the hairpin-G4.

Effect of hairpin-forming loop position on hairpin-G4 stability

In previous studies, the effect of the hairpin-forming loop was only evaluated by placing it in the centre of the G₄-forming sequence (13,15). For more comprehensive understanding of the stability and function of hairpin-G₄s, we checked whether the position of the hairpin-forming loop has any effect on G₄ formation and stability by evaluating the CD spectra and Tₘ of the model G₄s with a hairpin at alternative positions in comparison with LH-G₄, in which the hairpin is present in the third loop. For this purpose, we designed two long-loop G₄s, LH₁-G₄ and LH₂-G₄, which have a hairpin at the first and second loop positions, respectively (Figure 3C and D). Additionally, we made two more long-loop G₄s that have unstructured loops at the first and second positions (LU₁-G₄, and LU₂-G₄, respectively). The CD spectra showed that they all formed the parallel G₄ regardless of the position of the loop (Figure 3C). The thermal melting curves (Supplementary Figure S7B) demonstrated that the difference between Tₘ values of G₄s with loops at different positions (LH-G₄ vs. LH₁-G₄ or LH₂-G₄; LU-G₄ vs. LU₁-G₄ or LU₂-G₄) was near ~2°C regardless of the formation of hairpin in the loop (Figure 3D). These results indicated that irrespective of the position of the unstructured or hairpin loop, the G₄ structure was well maintained with similar stability. Importantly, this widens the scope for genome-wide searches for G₄s with a hairpin-forming loop.

Detection of hairpin-G4s using SYBR Green I fluorescence

Several fluorescent molecules have been developed as probes for the evaluation of G₄ formation and differentiating G₄ topologies (32,33). However, for the study of the hairpin in the hairpin-G₄s, it is necessary to find a probe for the detection of hairpin formation. SYBR Green I (SG) is a DNA-binding dye that shows increased fluorescence upon binding to double stranded DNA (dsDNA) (34,35). Since SG preferentially binds to the dsDNA in the stem region of the hairpin but has a weak binding affinity to G₄s (36,37), we speculated that SG can be used as a probe to detect hairpin formation in the long-loop hairpin-G₄s (Figure 4A). To test this possibility, we examined the fluorescence enhancement upon binding of SG to the hairpin-G₄. The SG fluorescence analysis revealed that the fluorescence from H-G₄s was higher than that from U-G₄s (Figure 4B). To evaluate the influence of SG binding to G₄ on the fluorescence, we analysed the fluorescence enhancement upon SG binding to G₄ present in the C-MYC promoter region (pu22myc), which does not contain any long loop or hairpin (38,39). Since pu22myc showed negligible fluorescence enhancement upon SG treatment (Figure 4B), we concluded that fluorescence enhancement observed in LH-G₄ mostly originated from the binding of SG to the hairpin region. Additionally, we observed that the fluorescence enhancement was proportional to the number of base pairs in the hairpin stem (Supplementary Figure S1). Accordingly, the difference in fluorescence from G₄s containing hairpins and unstructured loops was mostly prominent in LH-G₄ (Figure 4B), since LH-G₄ with a 12 bp stem could accommodate more SG dye. The same rule applied when the fluorescence from G₄s that contain the same loop length with different hairpin forming tendency was compared (Figure 4B). The fluorescence order was as follows: LH-G₄ > LI₁-G₄ > LI₂-G₄ > LU-G₄ and MH-G₄ > MI₁-G₄ > MI₂-G₄ > MU-G₄ (Figure 4B). However, among G₄s with shorter loops (S-G₄), a significant difference in fluorescence was not observed among SH-G₄, SU-G₄, SI₂-G₄, and SI₁-G₄ (Figure 4B), which can be explained by the low binding affinity of SG to the short loop, since at least 3–4 bp is necessary for SG binding to dsDNA (35). Overall, our data indicates that SG is as an effective probe for hairpin-G₄s, although this cannot contribute to identifying the nature of G₄s except confirm the presence of hairpin.

Evaluation of hairpin-G₄ structures by native gel electrophoresis

To obtain structural information of hairpin-G₄s, we evaluated the migration of the model hairpin-G₄s using electrophoretic mobility shift assays (EMSA) by assuming that the compact DNA structures migrate faster than unstructured or linear DNA (40). We assumed that the mobility of H-G₄s in EMSA gels would be faster than those of U-G₄s and I-G₄s because the loops in U-G₄s and I-G₄s would be more relaxed and less compact than the loops in H-G₄s. As expected, the migration order was H-G₄ > I-G₄ > U-G₄, regardless of the loop length (Figure 5). In addition, most model hairpin-G₄s formed the intramolecular structure (Figure 5A), although a small population of intermolecular G₄s with lower mobility than intramolecular
G4s was observed. The difference in mobility indicated that the dynamic nature of the unstructured loop significantly affects the overall structure of hairpin-G4s.

Evaluation of hairpin-G4 formation and stability by ¹H NMR

From the CD analysis, we found that H-G4s have higher stability than U-G4s, and thus we further examined the effect of the hairpin on the stabilization of the G4 structure by investigating H-G4s and U-G4s using ¹H NMR. In the NMR analysis of H-G4, peaks in both the 10–12.5 ppm region, representing imino protons of guanines in Hoogsteen base pairing (41–43), and the 12.5–15 ppm region, corresponding to imino protons arising from the Watson-Crick base pairing of duplex DNA in the hairpin stem region (44), were observed (Figure 6), confirming the formation of hairpins in the loop region of the long-loop G4s. Interestingly, there were 2, 5, and 12 peaks in the 12.5–14.0 ppm region of the spectra of SH-G4, MH-G4 and LH-G4, respectively, which represents the number of base pairs in the stem region of the hairpin (Figure 6). This also showed that the number of base pairs in the stem region predicted by the UNAFold program is well supported by the experimental evidence. Consistently, analysis of U-G4 NMR spectra revealed no NMR peaks in the low field region (12.5–14.0 ppm), suggesting the absence of hairpin formation in U-G4s (Figure 6). However, regardless of the length and structure of the loop region of the model G4s, all long-loop G4s showed 12 peaks between 10 and 12.5 ppm, confirming the formation of G4 structure with Hoogsteen base pairs (Figure 6).

Once we confirmed the formation of the hairpin structure in the long-loop G4 using NMR, we further evaluated the impact of hairpin formation on the stability of the G4 structure by examining the temperature-dependent conformational change of long-loop G4s. For this purpose, we monitored the NMR peaks of SH-G4, SU-G4, MU-G4, MH-G4, LH-G4 and LU-G4 in the temperature between 20 and 80°C (Figure 6), which enables us to understand the unfolding mechanism of H-G4s in comparison with U-G4s. In the case of LH-G4, the peaks for the Watson-Crick base pairing between 12.5 to 14.0 ppm decreased first as the temperature increased (Figure 6), followed by the decrease in imino proton peaks of G4 (10–12.5 ppm). Therefore, LH-G4 peaks disappeared completely at 80°C. However, in the case of LU-G4, a gradual disappearance of the imino proton peaks of G4 was observed even at temperatures lower than 70°C and all peaks disappeared completely at 80°C. This disappearance of G4-associated imino proton peaks at
Figure 4. SYBR Green is an effective probe to identify the presence of hairpin in putative hairpin-G4s. (A) Schematic representation of the SYBR Green fluorescence assay. SYBR Green binds to the dsDNA stem region of the hairpin in G4s, which results in fluorescence enhancement. However, in G4s without hairpins, the fluorescence enhancement is not observed due to the absence of SYBR Green binding to the DNA. (B) SYBR Green fluorescence spectra for model hairpin-G4s (H/I-G4s) and unstructured-loop mutant (U-G4) and conventional G4 (pu22myc) show highest fluorescence enhancement for hairpin-G4s, confirming hairpin formation.

higher versus lower temperatures was also seen in the middle (MH-G4 versus MU-G4) and short loop (SH-G4 versus SU-G4) G4s (Figure 6). Consequently, the initial disappearance of the 12.5–14 ppm Watson-Crick base pair peaks followed by disappearance of 10–12.5 ppm Hoogsteen base pair peaks in H-G4s indicates that the hairpin structure contributes to stabilizing the G4 structure.

Mining of hairpin-G4s motifs from human gene promoters

Although previous studies proposed that the loop length of 10nt is most stable for G4 formation, our studies on the long-loop G4s, especially LH-G4s, showed that (i) G4s with a loop longer than 10 nt can still be stabilized when the loop forms a hairpin with perfect base pairing in the stem region, (ii) the position of the hairpin-loop does not affect the stability of the hairpin-G4 and (iii) the stem-flanking region has no effect on hairpin-G4 stability. These observations provide biophysical information required to predict the long-loop G4s with high stability in the genome and led us to question whether hairpin-G4s, like conventional G4s, present in the promoter can affect gene expression (14). We mined promoter regions within −500 to +100 bp from the transcription start site (TSS) in the human genome to identify 25,707 G4s with loop lengths between 1–40nt by applying the schema provided in the methods (Step 1, Supplementary Table S3). After overlapping with previous experimentally verified G4s from G4Seq (14), we
Among the predicted hairpin-G4s, we selected the representative candidate G4s present in the promoter regions of MCM4, MSI1, TMCC3, NRBPI and CHST1 with hairpin-forming loop lengths of 21, 27, 23, 27 and 28 nt, respectively, as candidates for further functional study, considering the biological importance of these genes and their hairpin-forming propensity (Step 5, Materials and Methods). For instance, MCM4 is implicated in oesophageal carcinoma (45), in, MSI1 and NRBPI are implicated in colorectal cancer (46), and TMCC3 is implicated in breast cancer (47), whereas CHST1 is reported to be associated with spondyloepiphyseal dysplasia (48). Additionally, the hairpin-G4s from these promoters are expected to form stable hairpins based on the predicted structure and ΔGh values (–7.29, –4.96, –3.98, –5.33 and –6.72 kcal/mol, respectively) (Supplementary Figure S8A).

To investigate the biophysical properties, we analysed their CD spectra and thermal melting and cooling characteristics. The CD spectra of the candidate G4s revealed that they form parallel G4s as evidenced by the peak at 260 nm and trough at 240 nm (Supplementary Figure S8B). We mutated the hairpin-forming loop region in the candidate hairpin-G4s to disrupt the hairpin and generated unstructured-loop-G4s with the same loop length as the hairpin-G4s (Table 1). The CD spectra of the unstructured-loop mutants revealed that G4 formation was still observed (Supplementary Figure S8B). Thermal melting analysis revealed that the Tm of the hairpin-G4 candidates was higher than their corresponding unstructured-loop-G4s except in the case of TMCC3 (Supplementary Figure S8C and Table S1). The ΔTm between the hairpin-G4 candidates and their unstructured-loop mutants at 1 mM MgCl2 were 13.88, 14.20, 7.37 and 5.99°C for CHST1, MCM4, MSI1 and NRBPI, respectively, indicating that the hairpin-G4 candidates were much more stable than their mutants (Supplementary Figure S8C and Table S1). In the case of TMCC3-G4 and its mutant G4, both showed exceptionally high thermal stability and therefore the Tm could not be accurately determined (Supplementary Figures S8 and S9). Analysis of the melting and cooling curves of the candidate G4s (Supplementary Figure S10) indicated that the curves were almost reversible, signifying stable structure formation.

Native PAGE analysis of the candidate hairpin-G4s (Figure 5B) revealed that the G4s formed by the sequences were predominantly intramolecular as evidenced by their faster mobility. However, the presence of intermolecular G4s are also considered especially in the case of CHST1 and TMCC3-G4s by judging their slower mobility. Consistently, the presence of intermolecular G4 populations has been previously observed on native PAGE in many studies on functionally important G4s such as CMYC-G4 (39), VEGF-G4 (49) and HOX11-G4 (50). Interestingly, it is found that the levels of intra- and intermolecular G4s varies depending on the hairpin formation (Figure 5). This confirmed the advantage of using native gel for identification of hairpin formation. The CD spectra analysis. However, it is noteworthy that only the intramolecular structure is biologically relevant since the intermolecular G4s cannot be formed within promoter regions in cellulo and thus it does not affect reporter activity.

We also performed the SYBR Green fluorescence assay to probe for hairpin formation and found that the hairpin-G4s displayed higher fluorescence than the mutants (Supplementary Figure S11), confirming hairpin formation in the candidate hairpin-G4s; this was also confirmed by 1H NMR spectra. While both candidate hairpin-G4s and their
Figure 6. Hairpin-G4s show characteristics of hairpin and G4 formation and are more stable than their unstructured loop mutants. In $^1$H NMR spectra analysis, hairpin-G4s (SH-G4, MH-G4, and LH-G4) show peaks for hairpin formation between 12.5–15 ppm, whereas G4s with unstructured loops (SU-G4, MU-G4 and LU-G4) do not show peaks for hairpin formation. Both H-G4s and U-G4s display peaks for G4 formation between 10–12.5 ppm. Temperature-dependent NMR spectra of H-G4s and U-G4s shows that H-G4s are more stable than U-G4s because the peaks for G4 start to disappear at higher temperatures. All spectra were recorded in 10 mM HEPES pH 7.5, 100 mM KCl, 1 mM MgCl$_2$ at 25°C.

The functional impact of hairpin-G4s present in the genome was evaluated using luciferase reporter assays in HEK293T cells. To measure the cis-regulatory effect of the hairpin-G4, the candidate hairpin-G4s and their mutants were inserted in the region upstream of the SV40 promoter of a luciferase reporter vector, and luciferase activity was measured (Figure 7A–C). We found that MCM4 ($P \leq 0.05$), TMCC3 ($P \leq 0.001$) and CHST1 ($P \leq 0.05$)-G4s reduced reporter expression whereas MSI1- and NRBP1-G4s did not show significant increase in reporter expression compared with their respective mutants (Figure 7C). The only difference between the candidates and their mutants is the mutation of the hairpin-forming loop that disrupts hairpin formation. Therefore, the difference between the hairpin-G4s and the mutant-G4s in reporter activity can be attributed to the presence of a hairpin structure.

Previous studies showed that G4 formation in the 5′UTR region on the antisense strand can affect transcription as a result of RNA polymerase II stalling and following transcription attenuation (19). To investigate whether a similar effect exists in the case of hairpin-G4s, we inserted the candidate hairpin-G4 and mutant-G4 sequences in the 5′UTR on the antisense strand of the reporter plasmid (Figure 7D). Reporter activity was significantly increased in CHST1 ($P \leq 0.01$), MCM4 ($P \leq 0.05$), and NRBP1 ($P \leq 0.01$)-G4s, compared to the activities in their respective mutants, while TMCC3-G4 and MSI1-G4 did not affect reporter expression significantly (Figure 7F). This suggests that hairpin-formation in the loop can have different effects on gene expression depending on its propensity to stabilize the hairpin-G4.

To explain this different effect on reporter expression, we performed a reporter assay using the representative model G4s (LH-G4 and LU-G4) in the sense and antisense strands of the 5′UTR and promoter regions (Figure 7). Our reporter assay was based on a previous study where the
Figure 7. The effect of hairpin formation in the candidate G4s present in the human gene promoters on the promoter activity. (A) Schema showing reporter construction for analyzing cis-regulatory effect of hairpin-G4s on luciferase reporter activity. (B) Cis-regulatory effect of the model hairpin-G4 and their mutants with the unstructured loop on the luciferase reporter activity. (C) Cis-regulatory effect of the candidate-G4 containing hairpin and their mutants with the unstructured loop on the luciferase reporter activity. (D) Reporter activity was constructed in the same way in Figure 7A except inserting the G4 sequences into the 5’UTR region. (E) The model G4s (LU-G4 and LH-G4) have been tested as controls in both sense and antisense strand. (F) Reporter activity was also analyzed for candidate-G4 containing hairpin and their mutants in the 5’UTR. All data was normalized to the respective control G4 containing unstructured loop. In all experiments, luciferase expression was quantified 24 hours after transfection. Error bars represent standard deviations of three individual trials (N = 3). Significance was calculated using one-sample t-test (*p ≤ 0.05; **p ≤ 0.01; ***p ≤ 0.001).

G4 was cloned at different positions (5’UTR/ Promoter) in the sense or antisense strand, and the reporter expression was analyzed (19). We had cloned the G4s in the antisense strand because the candidate-G4s we chose from the genome were present on the antisense strand. We also checked the impact of hairpin-G4 formation on the sense strand by cloning the model G4s LH-G4 and LU-G4 in the sense strand in both promoter as well as 5’UTR positions (Figure 7B and 7E). The results suggest the hairpin-G4 (LH-G4) showed the higher expression activity compared to the G4 with unstructured loop (LU-G4) when they present in the promoter (Figure 7B), while LH-G4 showed the reduced activity when present in the 5’UTR regardless of their presence in sense or antisense strand.
Our data is similar to those observed in previous studies which showed an increase in reporter expression when the G4 was present in the promoter (25). In the case of the 5′UTR, our data matches the results obtained by the previous report which features a similar reporter plasmid and experiment (19). According to the reference study, it is plausible that the reporter expression may decrease due to the formation of G4 in the 5′UTR of the transcript (19). Overall, current results using the model G4s are opposite to the effect of candidates G4s. Therefore, it is expected that an additional variable might affect the reporter activity.

We observed that the effect of the model hairpin-G4s on reporter expression is different from that of the candidates G4s present in the genome, especially when it was inserted in the 5′UTR. Therefore, we considered the possibility of formation of multiple or alternative quadruplex structures by the guanine tracts as reported in the previous studies (39,51–53). To prove this hypothesis, we have designed MCM4-G4<sub>NGT</sub> (where NGT represents No Guanine Tract) in a manner that the GC% of the long loop is retained while the sequence in the stem region of the hairpin in MCM4-G4 was changed to remove the GGG tract (Supplementary Figure S13A). We confirmed that MCM4-G4<sub>NGT</sub> could also form stable parallel G4 structure similar to MCM4-G4 (Supplementary Figure S13B and S13C) and contain the hairpin structure (Supplementary Figure S13D). Then we compared the effect of MCM4-G4<sub>NGT</sub> on the reporter activity with that of MCM4-MutG4 (Supplementary Figure S13E). While the gene expression increased when MCM4-G4 was inserted in the 5′UTR in comparison to the MCM4-MutG4 (Figure 7C), reporter with MCM4-G4<sub>NGT</sub> showed the reduced reporter activity compared with the MCM4-MutG4 (Supplementary Figure S13E), similar to the trend found in the model sequence LH-G4. Since MCM4-G4<sub>NGT</sub> and LH-G4 do not contain a GGG tract in the long loop, the chance of forming mutually exclusive G4 formation is less. This clearly indicates that the presence of a GGG tract in the long-loop leads to the formation of mutually exclusive G4s, which result in a different impact on expression as proposed previously (39,52). In addition, we also observed a decrease in the expression when MCM4-G4<sub>NGT</sub> cloned in the promoter region compared to MCM4-MutG4 (Supplementary Figure S13E), which is similar to the decrease in expression observed in the case of MCM4-G4 cloned in the promoter region, although the magnitude of repression was higher in MCM4-G4<sub>NGT</sub> suggesting that the potential formation of mutually exclusive G4 or the presence of multiple conformations could influence the impact of H-G4 formation on expression. Further studies are needed to explore their impact on the structure, stability, and functional effects.

**Impact of hairpin-G4 formation on transcription fidelity in vitro**

Considering the different impact of candidate hairpin-G4s from the genome observed in reporter assays, we also wanted to confirm their ability to suppress transcription in vitro. In this regard, the RNA polymerase stop assay has been previously employed to ascertain the impact of G4 formation and stability on transcription fidelity (20). Briefly, if the stable G4 is formed, the transcript can either be slipped (longer transcripts) or arrested (shorter transcript). Additionally, transcriptional pausing can happen leading to the formation of lower amounts of the full-length transcripts. Significantly stable G4s produce higher amount of arrested transcripts at 35nt, which is the distance from the polymerase binding site to the G4-forming region. However, if the G4 is not stable, polymerase stalling does not happen and higher amounts of the full-length transcript are obtained. We performed this assay using the candidate hairpin-G4s and the model hairpin-G4s (LH/LU/LI-G4s). Our results on the model hairpin-G4s (Figure 8A) indicated the distinct reduction of full-length transcript and increase in the level of 35nt arrested transcript in LH-G4 compared with LU-G4 and LI-G4 (LU-G4 > LI-1-G4 > LI-2-G4 > LU-G4), correlating well with the expected stabilities of the designed hairpin-G4s. We also observed a similar trend in the case of the genomic candidates (Figure 8B), indicating the ability of the hairpin-G4s to perturb transcription by influencing the progression of RNA polymerase.

**DISCUSSION**

The potential roles of promoter G4s in gene regulation in cancer has been well established from previous studies such as those on C-MYC, BCL2, KRAS and C-KIT promoters (54). Furthermore, much experimental and informatic evidence has implicated the role of G4 formation in many cellular processes and disease pathogenesis (55,56). Therefore, identification of G4s in the genome and functional annotation of their roles are necessary for advanced studies of the genome. Although the optimal loop length was initially predicted to be 3–7 nt, later studies determined that longer loops can also form G4s (57). However, the stability of the G4 is inversely proportional to the length of the loop (58). Further structural and functional studies on G4s revealed that G4 formation is not simply determined by the presence of G-tract and the loop length but is also affected by many parameters such as loop length, loop composition, and the loop structure (3,4,58,59). We hypothesized that long-loop G4s with a loop longer than 20 nt can still form stable G4s depending on the stability of the loop. We further speculated that the presence of the hairpin structure in the long loop can make the loop less flexible, thus increasing the stability of the G4 structure. Indeed, analysis of T<sub>m</sub> from CD experiments on G4s with a hairpin loop (H-G4s) and G4s containing unstructured long loops (U-G4s) indicated that H-G4s were more stable than U-G4s in all instances. However, we also noticed that the H-G4 structure is not as stable as the canonical G4 with short loops, which may be due to the effect of the loop length as well as the effect of sudden hairpin unwinding that disrupts the G4 structure, as evidenced by NMR melting.

It is well known that the stability of the G4 decreases when the loop length increases. This is primarily due to the increased flexibility of longer loops. Corroborating this, when we compared the CD T<sub>m</sub> of the model hairpin-G4s (Table 3), we found that the T<sub>m</sub> of U-G4s decreased as the loop length increased (SU-G4 > MU-G4 > LU-G4). However, when we compared the CD T<sub>m</sub> of hairpin-G4s (Table 3), we found that there was no significant impact as the loop
Figure 8. Hairpin-G4s can act as transcriptional barriers. T7 polymerase stop assay using model G4s (A), and candidate G4s (wild type and mutant containing unstructured loop) (B). Large amounts of arrested (A) and slipped transcript (S), and less amount of full length transcript (FL) are observed in G4s with stable hairpin compared to the counterparts containing the unstructured loop. ‘M’ represents the size marker.

length increased, indicating that the formation of the hairpin structure in the loop reduced the impact of loop flexibility on the G4 structure (SH-G4∼MH-G4∼LH-G4). A similar result was also seen in the case of intermediate G4s (I-G4s), indicating that less stable stem-loops could also reduce the impact of loop length on G4 structure.

There are a number of factors that govern hairpin stability, such as salt concentration, loop length, loop composition, closing base pair (CBP), and presence of mismatches in the stem region. Higher salt concentration (60), short pyrimidine-rich loops (61–64), C:G/G:C closing base pairs (65), and absence of mismatches in the stem region are conditions that favour hairpin formation. However, most studies have been performed on hairpins with a homogeneous loop sequence such as TTTT, GGGG, CCCC and so on (66–69). Heterogeneous loop regions have not been systematically studied, so it is only possible to speculate on the influence of these parameters on the stability of hairpins with heterogeneous base composition in their loop regions.

The importance of loop distribution in G4s has not been well studied; the role of loop position was only recently analysed with loops of 3 nt length composed of thymines (59). Although the position of the loop was found to have a significant impact on the topology of the G4, it was difficult to establish direct correlations between the loop position and topology of the G4. In the condition in which two loops are of equal length and the third loop is longer, the study observed a maximum difference of ∼4°C in the average $T_m$. Our study also did not find any significant difference in the $T_m$ of H-G4s when the position of the hairpin-forming loop was changed (Figure 3A), which expands the scope for prediction of hairpin-G4s in the human genome...
by abolishing any bias with respect to loop position and facilitates the discovery of new gene targets for G4-mediated regulation.

The CD and NMR experiments both showed that the sequences can form parallel G4 structure. Indeed, the imino proton peaks between the unstructured loop G4 (U-G4) and hairpin-G4 (H-G4) were shifted, indicating that the structure of G4 in minutiae are changed, but the overall structure of the G4s in both cases seem to remain parallel confirmed by CD spectra at various temperature ranges (20–80 degree, Supplementary Figure S5). In addition, since NMR spectra of H-G4s did not show significant chemical shift changes after disappearance of hairpin peaks, hairpin does not seem to affect the large conformational change of G4. Therefore, it is expected that binding of hairpin to the side of G4 might contribute to the G4 stability. However, in this case, binding must be not strong enough to induce the conformational change of G4. Alternatively, this can be explained by that destabilizing effect of long loop is less in the case of hairpin-G4 compared to the unstructured-loop G4 due to the limited conformational flexibility of hairpin. Further study with high resolution structure will answer this question.

Non-canonical long-looped G4s are of immense importance considering their prevalence in the genome, which was previously demonstrated in G4-seq (14). The G4-forming sequences were shown to be enriched in regulatory regions such as promoters. From current hairpin-G4 mining studies in the human genome, we found that the gene promoters including important oncogenes and tumour suppressor genes such as NOTCH1, RET and ELL contain many potential hairpin-G4 forming sequences. We applied the parameters for hairpin-G4 stability prediction to the putative hairpin-G4 sequences obtained from the genome and chose to work on CHST1, MCM4, MSI1, NRPB1, and TMCC3 genes for our study because of their higher hairpin-forming propensity and biological significance in cancer and other diseases.

While reporter assays with the candidates TMCC3, CHST1 and MCM4 showed statistically significant changes (P < 0.05) in activities compared with their unstructured loop controls (Figure 7A), we could not find a direct correlation between the thermal stability of the candidate hairpin-G4s with the magnitude of expression change in reporter assays because the cis-regulatory effect might be mediated by the binding of various transcription factors, as evidenced by previous reports (70,71). The generalization of the effect of promoter G4 formation on gene expression and the impact on downstream processes is further complicated by the association of expression levels with diseases. For example, high levels of CHST1 and NRPB1 are associated with inflammation and cancer progression (72,73). Interestingly, NRPB1 was implicated as an oncogene in prostate cancer and a tumour suppressor in lung, breast, and colorectal cancers. In our analysis, the presence of the hairpin-G4 in the promoter of CHST1 reduced promoter activity in the reporter system, whereas the hairpin-G4 in the NRPB1 promoter increased reporter expression (Figure 7A). Concomitantly, MCM4, MSI1 and TMCC3 are oncogenes implicated in adenocarcinoma, cervical cancer, glioblastomas, leukaemia, and breast cancers (45,47,74). Therefore, low expression of these genes is greatly preferred. Our studies show that the promoter hairpin-G4s from MCM4 and TMCC3 genes from MSI1 promoter upregulated reporter activity, but the hairpin-G4 in the MSI1 promoter upregulated reporter activity (Figure 7A). Extrapolation of our reporter assay results implicates hairpin-G4s in both gene upregulation as well as downregulation. This indicates that in the genomic context, the activity of hairpin-G4s may be influenced by factors that can stabilize or destabilize the hairpin-G4s, such as supercoiling and transcription factor binding. Consistent with these results, genome-wide transcription studies of the human cytomegalovirus also revealed context-dependent roles of G4 in gene regulation (70). Individual promoter-level studies may facilitate further elucidation of the exact roles of hairpin-G4s in gene regulation.

We also evaluated the impact of hairpin-G4 formation on polymerase stalling (Figure 8) and transcription/translation (Figure 7). Although the model hairpin-G4s and candidate G4s always showed higher polymerase stalling activity than the corresponding G4s with the unstructured long loop (Figure 8), in the case of our reporter assay the candidate G4s showed upregulation when cloned in the 5′UTR. It is expected that the expression would decrease when the G4 was formed in the antisense because it can block polymerase progression (19). Closer examination of the sequence and secondary structure of the long loops in the candidate G4s revealed that they contained a GGG tract. It has been previously reported that the presence of multiple GGG tracts can lead to the formation of mutually exclusive G4s, possibly showing multiple conformations (39,51–53).

We hypothesized that the increase in expression in candidate G4s could be a result of mutually exclusive G4 formation, and we tested this hypothesis by performing the reporter assay by cloning the model LH-G4 and a mutated candidate G4 MCM-G4NGT, both without any GGG-tracks in the long loop, in the 5′UTR. Interestingly, both LH-G4 and MCM-G4NGT showed a significant decrease in reporter expression, unlike the candidate G4s with the GGG tract that showed increase in expression. The decrease in expression is similar to the previous study where they showed that stable G4s in the 5′UTR can stall polymerase progression (19). We also tested the strand-specific activity by cloning LH-G4 in the sense strand, and found that it can also decrease transcription, possibly due to the formation of hairpin-G4 in the 5′UTR region of the RNA as observed in a previous study (19).

Our strategy for selection of candidates considered the biological significance of the gene as well as the propensity for hairpin formation in the loop region. When we predict the propensity using the ΔGth value, it is evident that candidates with higher G–C bonding in the stem region will show the highest propensity for hairpin formation. Therefore, possibility of the formation of mutually exclusive G4s must be evaluated as a factor influencing the impact of hairpin-G4 formation on gene expression while working on hairpin-G4s from the human genome in future studies.

The hairpin itself is also a critical DNA secondary structure that is shown to exist in DNA as cruciform DNA (16,75). Hairpins are shown to play critical roles in transcription regulation by modifying the chromatin and also by
interacting with specific proteins that can bind to the hairpin structure (75). These proteins include key transcription factors such as p53 (76) and p73 (77), architectural proteins such as DEK (78), and tumour suppressors such as BRCA1 and PARP-1 (16). Considering the structure-specific recognition of these proteins, it is plausible that these factors can recognize the hairpin region of the hairpin-G4 and contribute to additional regulation, i.e., by preventing the binding of G4-specific transcription factors or proteins, by providing additional stability to the hairpin-G4 structure, or by directly impacting the local DNA topology. Therefore, there are multifarious ways in which both the hairpin as well as G4 regions of the hairpin-G4 can impact gene regulation.

CONCLUSION

We have demonstrated that stable hairpin-G4 structures can be formed even in sequences with long loops depending on the structural property of the loop. We also proved that the formation of hairpins results in increased stability of the hairpin-G4 irrespective of the position of the loop or the length of the stem-flanking region. Based on these observations, we identified five G4s that contain hairpin forming long loops. We further demonstrated that hairpin-G4s have a regulatory role in gene expression by examining the roles of G4-specific transcription factors or proteins, by providing additional stability to the hairpin-G4 structure, or by directly impacting the local DNA topology. Therefore, there are multifarious ways in which both the hairpin as well as G4 regions of the hairpin-G4 can impact gene regulation.

DATA AVAILABILITY

Derived data supporting the findings of this study are available from the corresponding author (K.K.) on request.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Author contributions: K.K.K. and S.R. designed the study, S.R. and M.R. performed the experiments, N.P., A.G. and S.R. performed the bioinformatics analysis and G4 prediction. K.K.K. and S.R. wrote the manuscript.

FUNDING

This work was supported by Samsung Science & Technology Foundation (SSTF-BA1301-01) and National Research Foundation of Korea funded by the Ministry of Science and ICT [2020R1A4A1018019, 2021A2C3011644] to K.K.K., and the Ministry of Education [2019R111A1A01060394] to S.R. Funding for open access charge: National Research Foundation of Korea.

Conflict of interest statement. None declared.

REFERENCES

1. Cang,X.H., Spoer,J and Cheatham,T.E. (2011) Insight into G-DNA structural polymorphism and folding from sequence and loop connectivity through free energy analysis. J. Am. Chem. Soc., 133, 14270–14279.
2. Karsiosisis,A.I., Hessari,N.M., Novellino,E., Spada,G.P., Randazzo,A. and Webba da Silva,M. (2011) Topological characterization of nucleic acid G-quadruplexes by UV absorption and circular dichroism. Angew. Chem. Int. Ed. Engl., 50, 10645–10648.
3. Tippiana,R., Xiao,W. and Myong,S. (2014) G-quadruplex conformation and dynamics are determined by loop length and sequence. Nucleic Acids Res., 42, 8106–8114.
4. Hazel,P., Huppert,J., Balasubramanian,S. and Neidle,S. (2004) Loop-length-dependent folding of G-quadruplexes. J. Am. Chem. Soc., 126, 16405–16415.
5. Webba da Silva,M. (2007) Geometric formalism for DNA quadruplex folding. Chemistry, 13, 9738–9745.
6. Guedin,A., De Cian,A., Gros,J., Lacroix,L. and Mergny,J.L. (2008) Sequence effects in single-base loops for quadruplexes. Biochimie, 90, 686–696.
7. Bedrat,A., Lacroix,L. and Mergny,J.L. (2016) Re-evaluation of G-quadruplex propensity with G4Hunter. Nucleic Acids Res., 44, 1746–1759.
8. Kikin,O., D’Antonio,L. and Bagg,P.S. (2006) QGRS mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. Nucleic Acids Res., 34, W676–W682.
9. Huppert,J.L. and Balasubramanian,S. (2005) Prevalence of quadruplexes in the human genome. Nucleic Acids Res., 33, 2908–2916.
10. Agrawal,P., Lin,C., Mathad,R.I., Carver,M. and Yang,D. (2014) The major G-quadruplex formed in the human BCL-2 proximal promoter adopts a parallel structure with a 13-nt loop in K+ solution. J. Am. Chem. Soc., 136, 1750–1753.
11. Jodoin,R., Bauer,L., Garant,J.M., Mahdi Laaref,A., Phaneuf,F. and Huppert,J.L. (2009) G-terminal folding of G-quadruplexes possessing a long central loop. RNA, 20, 1129–1141.
12. Amrane,S., Adrian,M., Heddi,B., Serero,A., Nicolas,A., Mergny,J.L. and Phan,A.T. (2012) Formation of pearl-necklace monomorphic G-quadruplexes in the human CEBB2 minisatellite. J. Am. Chem. Soc., 134, 5807–5816.
13. Lim,K.W., Khong,Z.J. and Phan,A.T. (2014) Thermal stability of DNA quadruplex-duplex hybrids. Biochemistry, 53, 247–257.
14. Chambers,V.S., Marsico,G., Boutell,J.M., Di Antonio,M., Smith,G.P. and Balasubramanian,S. (2015) High-throughput sequencing of DNA G-quadruplex structures in the human genome. Nat. Biotechnol., 33, 877–881.
15. Lim,K.W., Jenaroenpun,P., Low,Z.J., Khong,Z.J., Ng,Y.S., Kuznetsov,V.A. and Phan,A.T. (2015) Duplex stem-loop-containing quadruplex motifs in the human genome: a combined genomic and structural study. Nucleic Acids Res., 43, 5630–5646.
16. Brazda,V., Laister,R.C., Jagelska,E.B. and Arrowsmith,C. (2011) Cruciform structures are a common DNA feature important for regulating biological processes. BMC Mol. Biol., 12, 33.
17. Dong,F., Allawi,H.T., Anderson,T., Neri,B.P. and Lyamichev,V.I. (2001) Secondary structure prediction and structure-specific sequence analysis of single-stranded DNA. Nucleic Acids Res., 29, 3248–3257.
18. Tulpan,D., Andronescu,M. and Leger,S. (2010) Free energy estimation of short DNA duplex hybridizations. BMC Bioinformatics, 11, 105.
23. Kypr, J., Kejnovska, I., Renciu, D. and Vorlickova, M. (2009) Circular dichroism and conformational polymorphism of DNA. *Nucleic Acids Res.*, 37, 1713–1725.

24. Randazzo, A., Spada, G.P. and da Silva, M.W. (2013) In: Chaires, J.B. and Graves, D. (eds). *Quadruple Nucleic Acids*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 67–86.

25. Biswas, B., Kandpal, M. and Vivekanandan, P. (2017) A G-quadruplex motif in an envelope gene promoter regulates transcription and virion secretion in HBV genotype B. *Nucleic Acids Res.*, 45, 11268–11280.

26. Rachwal, P.A. and Fox, K.R. (2007) Quadruplex melting. *Methods*, 43, 291–301.

27. Owczarzy, R., Moreira, B.G., You, Y., Behlke, M.A. and Walder, J.A. (2010) Selective recognition of oncogene promoter G-quadruplexes by Mg2+. *Biochem. Biophys. Res. Commun.*, 402, 614–618.

28. Kreig, A., Calvert, J., Sanoica, J., Cullum, E., Tippana, R. and Myong, S. (2015) G-quadruplex formation in double strand DNA probed by NMM and CV fluorescence. *Nucleic Acids Res.*, 43, 7961–7970.

29. Owczarzy, R., Moreira, B.G., You, Y., Behlke, M.A. and Walder, J.A. (2010) Selective recognition of oncogene promoter G-quadruplexes by Mg2+. *Biochem. Biophys. Res. Commun.*, 402, 614–618.

30. Rachwal, P.A., Brown, T. and Fox, K.R. (2007) Sequence effects of single base loops in intramolecular quadruplex DNA. *FEBS Lett.*, 581, 1657–1660.

31. Yan, Y.Y., Lin, J., Ou, T.M., Tan, J.H., Li, D., Gu, L.Q. and Huang, Z.S. (2010) Selective recognition of oncoregene promoter G-quadruplexes by Mg2+. *Biochem. Biophys. Res. Commun.*, 402, 614–618.

32. Bram, A.I., Pavlovic, R., McGivney, J.B., Casas-Finet, J.R., Bishop, E.S., Stroose, R.J., Schenerman, M.A. and Geddes, C.D. (2012) SYBR Green I: fluorescence properties and interaction with DNA. *J. Fluoresc.*, 22, 1189–1199.

33. Nicolaidis, J.M., Miller, S.T., Jeffrey, P.D., Barrett, S.P., Rabin, R.P., Lawton, T.J. and Yatsky, L.A. (2012) Optimized end-stacking provides specificity of N-methyl mesoporphyrin IX for human telomeric G-quadruplex DNA. *J. Am. Chem. Soc.*, 134, 20446–20456.

34. Zipper, H., Brunner, H., Bernhagen, J. and Vithazum, F. (2004) Investigations on DNA intercalation and surface binding by SYBR Green I, its structure determination and methodological implications. *Nucleic Acids Res.*, 32, e103.

35. Dragan, A.I., Pavlovic, R., McGivney, J.B., Casas-Finet, J.R., Bishop, E.S., Stroose, R.J., Schenerman, M.A. and Geddes, C.D. (2012) SYBR Green I: fluorescence properties and interaction with DNA. *J. Fluoresc.*, 22, 1189–1199.

36. Xu, H., Gao, S., Yang, Q., Pan, D., Wang, L. and Fan, C. (2010) Amplified fluorescent recognition of g-quadruplex folding with a cationic conjugated polymer and DNA intercalator. *ACS Appl. Mater. Interfaces*, 2, 3211–3216.

37. Zhan, S., Wu, Y., Luo, Y., Liu, L., He, L., Xing, H. and Zhou, P. (2014) Label-free fluorescent sensor for lead ion detection based on lead(II)-stabilized G-quadruplex formation. *Anal. Chem.*, 462, 19–25.

38. Ambrus, A., Chen, D., Dai, J., Jones, R.A. and Yang, D. (2005) Solution structure of the biologically relevant G-quadruplex element in the human c-MYC promoter. Implications for G-quadruplex stabilization. *Biochemistry*, 44, 2048–2058.

39. Siddiqui-Jain, A., Grand, C.J., Bearss, D.J. and Hurley, L.H. (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc. Natl. Acad. Sci. U.S.A.*, 99, 11593–11598.

40. Bryan, T.M. and Baumann, P. (2011) G-quadruplexes: from guanine gels to chemotherapeutics. *Mol. Biotechnol.*, 49, 198–208.

41. Dai, J., Chen, D., Jones, R.A., Hurley, L.H. and Yang, D. (2006) NMR solution structure of the major G-quadruplex structure formed in the human BCL2 promoter region. *Nucleic Acids Res.*, 34, 5133–5144.

42. Adrian, M., Heddi, B. and Phan, A.T. (2012) NMR spectroscopy of G-quadruplexes. *Methods*, 57, 11–24.

43. Aboul-ele, F., Murchie, A.I. and Lilley, D.M. (1992) NMR study of parallel-stranded tetraplex formation by the hexaexonucleotide d(TG4T). *Nature*, 360, 280–282.

44. Rink, L.J. and Tinoco, I. Jr (1991) A proton NMR study of DNA dumb-bell structure with hairpin loops of only two nucleotides: d(CACGGTGTGGC TTCGCGCA). *Nucleic Acids Res.*, 19, 3695–3700.

45. Choy, B., LaLonde, A., Que, J., Wu, T. and Zhou, Z. (2016) MCM4 and MCM7, potential novel proliferation markers, significantly correlated with Ki-67, Bmi1, and cyclin E expression in esophageal adenocarcinoma, squamous cell carcinoma, and precursor lesions. *Hum. Pathol.*, 57, 126–135.
68. Kuznetsov, S.V., Ren, C.C., Woodson, S.A. and Ansari, A. (2008) Loop dependence of the stability and dynamics of nucleic acid hairpins. *Nucleic Acids Res.*, 36, 1098–1112.

69. Shen, Y.Q., Kuznetsov, S.V. and Ansari, A. (2001) Loop dependence of the dynamics of DNA hairpins. *J. Phys. Chem. B*, 105, 12202–12211.

70. Mishra, S.K., Tawani, A., Mishra, A. and Kumar, A. (2016) G4IPDB: a database for G-quadruplex structure forming nucleic acid interacting proteins. *Sci. Rep.*, 6, 38144.

71. Zhang, C., Liu, H.H., Zheng, K.W., Hao, Y.H. and Tan, Z. (2013) DNA G-quadruplex formation in response to remote downstream transcription activity: long-range sensing and signal transducing in DNA double helix. *Nucleic Acids Res.*, 41, 7144–7152.

72. Liao, Y., Yang, Z.H., Huang, J.T., Chen, H., Xiang, J., Li, S.M., Chen, C.Y., He, X., Lin, F., Yang, Z.L. et al. (2018) Nuclear receptor binding protein 1 correlates with better prognosis and induces caspase-dependent intrinsic apoptosis through the JNK signalling pathway in colorectal cancer. *Cell Death. Dis.*, 9, 436.

73. Li, X., Tu, L., Murphy, P.G., Kadono, T., Steeber, D.A. and Tedder, T.F. (2001) CHST1 and CHST2 sulfotransferase expression by vascular endothelial cells regulates shear-resistant leukocyte rolling via L-selectin. *J. Leukoc. Biol.*, 69, 565–574.

74. Das, M., Prasad, S.B., Yadav, S.S., Govardhan, H.B., Pandey, L.K., Singh, S., Pradhan, S. and Narayan, G. (2013) Over expression of minichromosome maintenance genes is clinically correlated to cervical carcinogenesis. *PLoS One*, 8, e69607.

75. Ward, G.K., Shihab-el-Deen, A., Zannis-Hadjopoulos, M. and Price, G.B. (1991) DNA cruciforms and the nuclear supporting structure. *Exp. Cell Res.*, 195, 92–98.

76. Coufal, J., Jagelska, E.B., Liao, J.C. and Brazda, V. (2013) Preferential binding of p53 tumor suppressor to p21 promoter sites that contain inverted repeats capable of forming cruciform structure. *Biochem. Biophys. Res. Commun.*, 441, 83–88.

77. Cechova, J., Coufal, J., Jagelska, E.B., Fojta, M. and Brazda, V. (2018) p73, like its p53 homolog, shows preference for inverted repeats forming cruciforms. *PLoS One*, 13, e0195835.

78. Waldmann, T., Baack, M., Richter, N. and Gruss, C. (2003) Structure-specific binding of the proto-oncogene protein DEK to DNA. *Nucleic Acids Res.*, 31, 7003–7010.