Published online 12 April 2013 *Nucleic Acids Research, 2013, Vol. 41, No. 10* ***5533–5541***
doi:10.1093/nar/gkt264

Co-transcriptional formation of DNA:RNA hybrid G-quadruplex and potential function as constitutional *cis* element for transcription control

Ke-wei Zheng, Shan Xiao, Jia-quan Liu, Jia-yu Zhang, Yu-hua Hao and Zheng Tan*

State Key Laboratory of Biomembrane and Membrane Biotechnology, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, P. R. China

Received December 28, 2012; Revised February 16, 2013; Accepted March 24, 2013

ABSTRACT

G-quadruplex formation in genomic DNA is considered to regulate transcription. Previous investigations almost exclusively focused on intramolecular G-quadruplexes formed by DNA carrying four or more G-tracts, and structure formation has rarely been studied in physiologically relevant processes. Here, we report an almost entirely neglected, but actually much more prevalent form of G-quadruplexes, DNA:RNA hybrid G-quadruplexes (HQ) that forms in transcription. HQ formation requires as few as two G-tracts instead of four on a non-template DNA strand. Potential HQ sequences (PHQS) are present in >97% of human genes, with an average of 73 PHQSs per gene. HQ modulates transcription under both *in vitro* and *in vivo* conditions. Transcriptomal analysis of human tissues implies that maximal gene expression may be limited by the number of PHQS in genes. These features suggest that HQs may play fundamental roles in transcription regulation and other transcription-mediated processes.

INTRODUCTION

Besides the conventional Watson–Crick double helix, DNA can also adopt other forms of higher order structures, such as G-quadruplexes. G-quadruplexes are four-stranded structures (Figure 1A) formed by guanine-rich (G-rich) nucleic acids. A G-quadruplex is characterized by a stack of planar G-quartets (Figure 1B), each of which comprises four guanines connected by eight Hoogsteen hydrogen bonds (1). Interest in the biological significance of G-quadruplexes was initially evoked by telomeric DNA that forms intramolecular G-quadruplex and inhibits its extension by telomerase (2), a reverse transcriptase expressed in cancerous but not in normal somatic cells. Thus, targeting G-quadruplexes has been considered as a promising therapeutic strategy against cancer and other diseases (3,4). In the past few years, bioinformatic analyses have revealed large numbers of potential intramolecular G-quadruplex sequences in the genomes of various organisms (5,6), ranging from bacteria to animals. It was predicted that G-quadruplexes might modulate transcription and translation because of their enrichment near transcription and translation start sites (7).

For many years, interest in G-quadruplex has drawn attention to the question of whether G-quadruplexes are actually present in cells. Using engineered antibody, a recent work explicitly provides substantive evidence for the presence of G-quadruplex structures in the genome of mammalian cells (8). Genomic DNA undergoes dynamic changes in structural organization during transcription and replication. Information on the formation of G-quadruplexes in these processes and their structural forms is important for understanding the physiological function of G-quadruplexes. G-quadruplex formation requires four G-tracts (Figure 1A). A G-quadruplex can form intramolecularly or intermolecularly depending on the number of G-tracts available in participating strands. For example, an intramolecular G-quadruplex requires a strand that carries at least four tandem G-tracts. A dimeric intermolecular G-quadruplex can form as long as the two participating strands can supply a sum of four G-tracts (9). To date, studies on G-quadruplexes of genomic sources have mostly been focused on intramolecular G-quadruplexes, and the biogenesis of G-quadruplexes in physiological processes in cells remains largely elusive.

In this work, we report the discovery of a unique G-quadruplex structure, a DNA:RNA hybrid G-quadruplex (HQ) (Figure 1C), that forms during transcription by G-tracts from both the non-template DNA strand and the nascent RNA transcript. Recently, the

*To whom correspondence should be addressed. Tel: +86 10 6480 7259; Fax: +86 10 6480 7099; Email: z.tan@ioz.ac.cn or tanclswu@public.wh.hb.cn

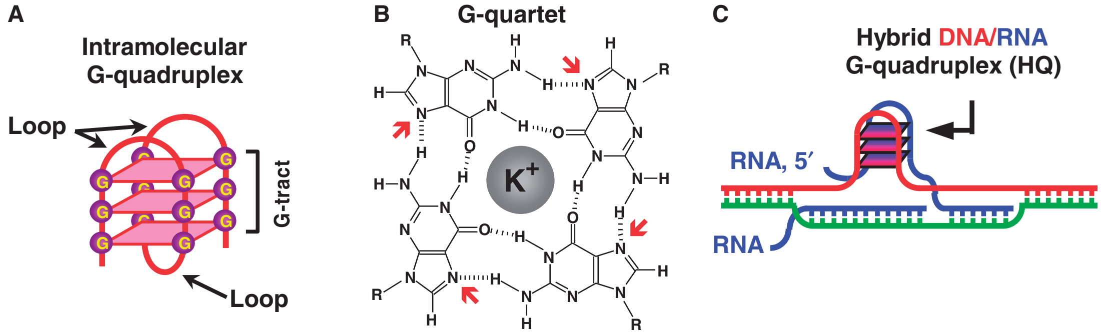

Figure 1. Structure of (**A**) an intramolecular G-quadruplex composed of three G-quartet layers with four G-tracts connected by three loops, (**B**) a G-quartet with four guanines connected by eight Hoogsteen hydrogen bonds (dashed lines) and (**C**) a DNA:RNA hybrid G-quadruplex (HQ) with two G-tracts from the non-template DNA strand and two from an RNA transcript. Arrow in (B) indicates the 7-nitrogen.

formation of HQ in the *in vitro* transcription of DNA carrying the CSB II motif from mitochondrial genome was suggested (10). In the work, the transcription of a plasmid and a linear DNA generated RNA fragments of ~50 nt in size that was resistant to RNase A digestion and that disappeared when the guanine was replaced by the 7-deaza guanine analog either in the template DNA or RNA. This observation was considered to reflect a formation of HQ. However, the identity of these ~50-nt fragments and the structures associated with them were not clarified. Because the size of the fragments is much larger than the CSB II G-core sequence (14 nt) that could in principle be protected in an HQ from RNase A digestion, the formation of HQ needs to be examined with more stringent criteria. Using a reconstituted T7 transcription model, here, we provide detailed chemical and biochemical analyses of the co-transcriptional formation of HQ. To illustrate the biological implications of HQ, we also show that HQ modulates transcription under both *in vitro* and *in vivo* conditions, and the occurrence frequency of potential HQ sequence (PHQS) in genes correlates with the transcriptomal profiles in human tissues. Collectively, these results suggest that PHQSs are abundant in genome, and HQs are a major physiological form of G-quadruplexes that forms in transcription. HQs may function as general *cis* control elements for the regulation of transcription.

MATERIALS AND METHODS

Double-stranded DNA and plasmids

Double-stranded DNA (dsDNA) carrying a T7 promoter and the indicated gene sequences was prepared by overlap extension polymerase chain reaction, as described previously (11). dsDNA with single mutation was annealed by synthetic oligonucleotides. A fragment of the NRAS gene, or one of several corresponding mutants, was inserted into the pGL3-control plasmid at the HindIII/NcoI restriction site.

In vitro transcription

Transcription was carried out essentially as described previously (11), with 0.1 µM FAM-labeled dsDNA in a total volume of 100 µl at 37°C for 1.5 h in transcription buffer, which contained 40 mM Tris–HCl (pH 7.9), 8 mM $MgCl_2$, 10 mM dithiothreitol (DTT), 2 mM spermidine, 50 mM KCl or LiCl, 200 U T7 RNA polymerase (Fermentas, MBI), 0.4 U inorganic pyrophosphatase (Fermentas, MBI) and 2 mM NTP (0.4 mM GMP supplied when using dzGTP), in the presence or absence of 40% (w/v) PEG 200. The reaction was stopped with the addition of 1/25 vol of 0.5 M ethylene diamine tetraacetic acid disodium ($EDTA-Na_2$). For dimethyl sulfate (DMS) footprinting, native gel electrophoresis and photocleavage footprinting, the samples were further treated with 0.4 mg/ml of RNase A (Fermentas, MBI), then 0.6 mg/ml proteinase K (TAKARA, Dalian) at 37°C for 1 h each.

Analysis of RNA transcripts

Transcription was performed as described earlier in the text, except that 20 ng/µl of plasmid DNA was used as the transcription template, 1 U/µl of RiboLock RNase inhibitor (Fermentas, MBI) was added and 10% of the UTP was substituted with fluorescein-UTP (Fermentas, MBI). After transcription, samples were treated with 0.04 U/µl of DNase I (Fermentas, MBI) at 37°C for 15 min, and the digestion was stopped with phenol/chloroform extraction. RNA products were denatured and resolved on a denaturing 8% polyacrylamide gel.

DMS footprinting

DMS footprinting was performed as described previously (11).

Native gel electrophoresis

Samples were electrophoresed at 4°C, 8 V/cm, on an 8% polyacrylamide gel that contained 150 mM KCl and 40% (w/v) PEG 200. Electrophoresis was run in 1× tris-borate-EDTA (TBE) buffer that contained 150 mM KCl (11).

Photo-cross-linking

Transcription was carried out as in the other experiments, but with the addition of 4-thio-uridine-5′-triphosphate (TriLink BioTechnologies, USA) in place of the usual UTP concentration. After transcription, 50 µl of sample was transferred to a 24-well microtiter plate (Greiner

Bio-One, Germany), placed on ice, in a UVP CL-1000 Ultraviolet Cross-linker (UVP, USA) (12) and irradiated for 20 min with 365-nm ultraviolet (UV) light at a distance of 4–5 cm. DNA was recovered with phenol/chloroform extraction and ethanol precipitation. Next, primer extension was performed to identify the cross-linked sites with the 5′-FAM-CCAGCCTGCGGCGAGTG primer and Thermo Sequenase from the Cy5 Dye Terminator Cycle Sequencing Kit (GE Healthcare, USA). Guanine and adenine markers were prepared by primer extension in the presence of ddCTP or ddTTP, respectively, in a 1/100 molar ratio to dCTP or dTTP, respectively.

Ligand-induced photocleavage footprinting

Transcription was conducted as described earlier in the text but in the absence of DTT. Photocleavage footprinting was conducted as described previously (13). After phenol/chloroform extraction (twice) and ethanol precipitation, 10 µg of plasmid DNA was dissolved in 1× Tris-EDTA (TE) buffer and digested with 0.4 mg/ml of RNase A (Fermentas, MBI) at 37°C for 15 min. The NTPs in the samples were removed with a desalting column from the mini Quick Spin Oligo Columns (Roche, Germany). The plasmid was digested with 50 U SalI (Fermentas, MBI) at 37°C for 1 h, followed by heating at 65°C for 10 min. The digested DNA was purified with the desalting column of the mini Quick Spin Oligo Columns (Roche, Germany). Purified DNA was labeled at the recessive 3′-end in a fill-in extension reaction at 37°C for 30 min with the klenow exo− polymerase (Fermentas, MBI) and 1 µM fluorescein-12-dUTP (Fermentas, MBI). The reaction was terminated by adding EDTA to a concentration of 20 mM. For DNA precipitation, each 60-µl sample received 30 µl of 7.5 M ammonium acetate, 2 µl of glycogen (10 mg/ml) and 300 µl of 100% ethanol, was thoroughly mixed and was left overnight at −20°C. Precipitated DNA was dissolved, denatured and resolved on a denaturing 12% polyacrylamide gel (13). The gel was scanned on a Typhoon 9400 (GE Healthcare, America) imager and processed with ImageQuant 5.2 software.

In vivo gene expression of luciferase reporter

HEK293 cells were grown to 60–80% confluence in 12-well plates before transfection. The transfection mixture contained 0.2 µg of modified reporter vector, pGL3-Control, which contained the human *NRAS* G-core or mutant sequences (Figure 4A, scheme), and 0.02 µg of the internal control vector, pRL-CMV. Transfection was performed with Lipofectamine 2000 (Invitrogen). Transfected cells were cultured in the presence or absence of 2 µM Zn-TTAPc. Cell lysates were prepared at 24 h after transfection. Lysates were assayed for luciferase activity with the Dual-Luciferase Reporter Assay System (Promega) on a Multi-Plate Reader (Biotek, USA).

Computational analysis of PHQS in genes in human genome

Protein-coding sequences in FASTA format were downloaded from the Ensembl 68 database with the BioMart (http://www.ensembl.org/). PHQS was identified with a home-made perl script that searched for $G_{\geq 3}(N_{1-7}G_{\geq 3})_{\geq 1}$, where G denoted guanine and N denoted any nucleotide, including G. The identified motifs were then grouped into four categories, denoted 1G, 2G, 3G and 4G+, respectively, according to the number of G-tracts they contain. The script is implemented using the Active PERL 5.14.2. The script is included in the Supplementary Data File as an attachment. Transcriptomal data on human genes were obtained from two sources: (i) the GNF Atlas 2 track, which covered 44 775 transcripts in 79 tissues (14); this was downloaded from the UCSC genome browser; and (ii) the Supplementary Table S2 from Castle *et al.* (15), which covered 33 220 transcripts in 11 tissues. The maximal expression value of each gene in different tissues was retrieved along with the gene ID with a home-made perl script. The occurrence of PHQS in each corresponding gene was analyzed as described earlier in the text.

RESULTS AND DISCUSSION

Co-transcriptional HQ formation in linear dsDNA

We reported recently that G-quadruplexes could form in the non-template strand of dsDNA during *in vitro* transcription with T7 RNA polymerase (11). We soon noticed that, among the four G-tracts in the NRAS oncogene sequence, one G-tract seemed unprotected from cleavage in the DMS footprinting (11), a technique for detecting guanines in the G-quartet of G-quadruplex (16). A G-quadruplex requires four G-tracts to form (Figure 1A). From that, we reasoned that the G-quadruplex assembly might involve G-tracts from both DNA and the transcribed RNA to meet this requirement.

To test this possibility, we conducted transcription in DNA that carried the NRAS sequence (Figure 2A, top scheme). In our previous work (11), transcription was allowed to proceed for two G-tracts by supplying only two NTPs. To determine whether the failure to protect all the four DNA G-tracts was specific to the limited transcription, both partial and full-length transcriptions were carried out. After transcription was terminated with EDTA, samples were treated with RNase A to digest the RNA; then protease K was added to remove the polymerase before a structural analysis. RNase H was not used here because it would be inactivated by EDTA-mediated depletion of Mg^{2+}. RNase A can digest single-stranded RNA, either free or hybridized with DNA, in a 0–100 mM salt solution (product instruction and Supplementary Figure S1). A same DNA carrying an intramolecular DNA G-quadruplex was prepared by heating and cooling (11) and used as a reference. This reference had four G-tracts that were protected from cleavage during DMS footprinting as expected (Figure 2A, lane 2, red arrowheads). Partial transcription with ATP and GTP resulted in protection of only three, instead of four, consecutive G-tracts from the 5′-end of the non-template strand (Figure 2A, lane 4, blue arrowheads), consistent with previous observations (11). In contrast, all four G-tracts were protected when the DNA underwent full

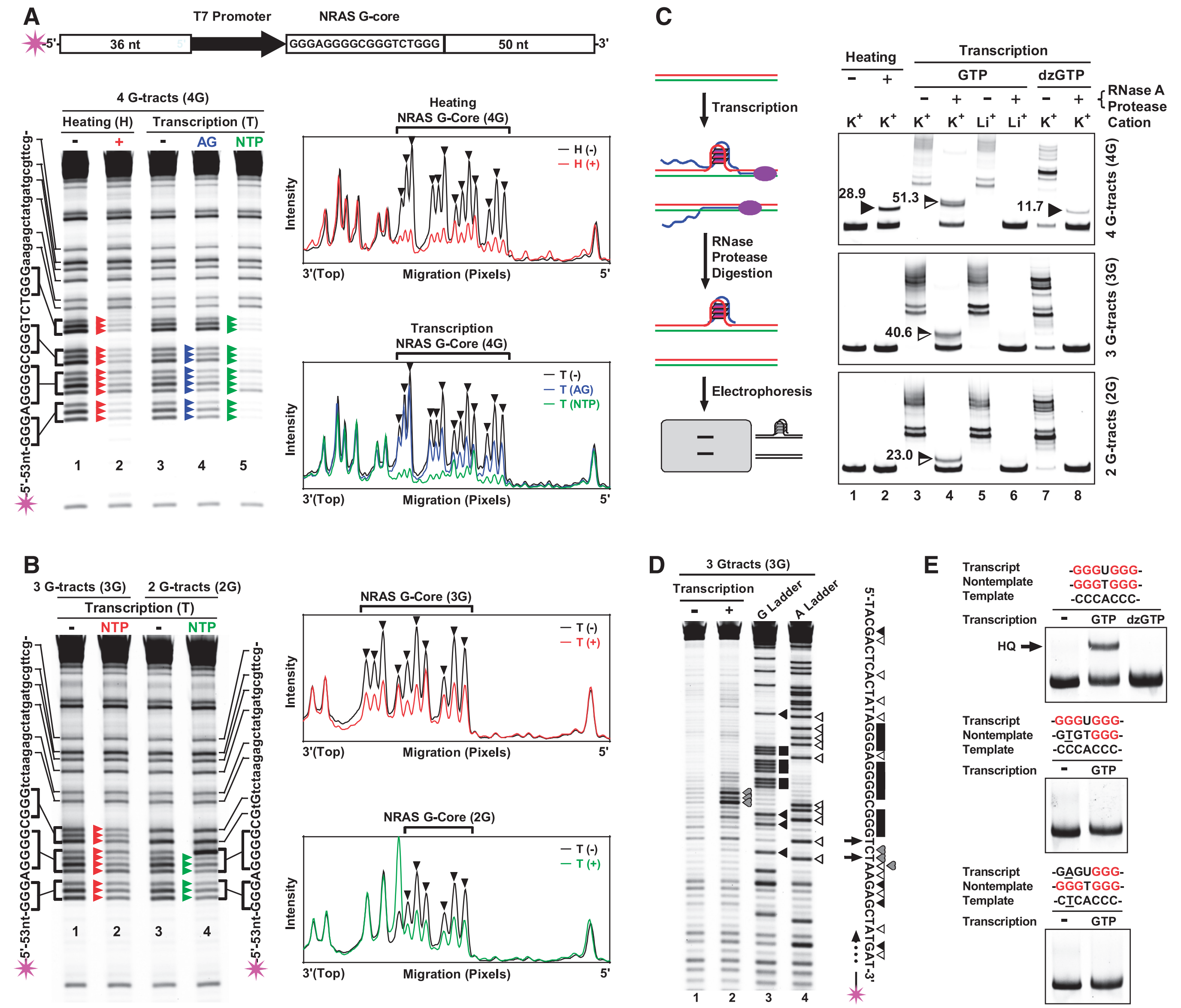

Figure 2. HQ formation during transcription of linear dsDNA that contained a G-core motif from the NRAS gene on the non-template strand. (**A, B**) HQ detection by DMS footprinting in DNA that carried (A) four or (B) three and two G-tracts. DNA was labeled at the 5′-end of the non-template strand with a FAM (asterisk), then subjected either to heat denaturation/renaturation (H) or transcription (T) with ATP and GTP (AG) or all four NTPs (NTP) and followed by RNase A and protease K digestion. Footprinting cleavage fragments were resolved by denaturing gel electrophoresis (left) and digitized (right) for comparison. Arrowheads indicate protected guanines. (**C**) HQ detection by native gel electrophoresis. DNA carrying four (top), three (middle) or two (bottom) G-tracts was heated or transcribed with either GTP or dzGTP and the other three NTPs as in (A, B). After RNase A and protease K digestion, the DNA was resolved on a native gel. Filled and half-filled arrowheads indicate intramolecular DNA G-quadruplexes and intermolecular DNA:RNA hybrid G-quadruplexes (HQ), respectively, and their amounts as the percentage of total DNA. (**D**) Primer extension detection of photo-cross-linking at the non-template strand by 4-thio-uridine incorporated into the RNA transcript. The 5′-FAM labeled primer was annealed to the 3′-end of the non-template of non-transcribed (N, lane 1) or transcribed (T, lane 2) DNA, followed by extension with DNA sequenase. Extension products were resolved by denaturing gel electrophoresis. G and A ladders were obtained by primer extension with ddCTP and ddTTP, respectively. Symbols on the right sides of the lanes and non-template sequence indicate the cross-linking sites (gray heart), G-tracts (black bar), guanines (filled triangle) and adenines (open triangle). Arrows on the left side of the non-template sequence indicate the sites of 4-thio-uridine incorporation. (**E**) HQ formation requires four G-tracts. Transcription and HQ detection were carried out as in (C) using dsDNA carrying the indicated G/C-tracts without or with a single mutation (nucleotide underlined) at the middle of a single G- or C-tract. Transcriptions in (A–E) were all conducted in solution containing 50 mM K^+ and 40% (w/v) PEG 200.

transcription (Figure 2A, lane 5, green arrowheads). These results strongly suggested the formation of HQ, at least during the partial transcription, despite that the structure formed in the full transcription was not clear. The protection to the G-tracts in the partial transcription was weaker than that in the full transcription. This implied that more DNA participated in the formation of G-quadruplex when more G-tracts from RNA were available. To verify HQ formation, we reduced the number of G-tracts in the DNA to three (3G) or two (2G) to prevent the formation of non-hybrid intramolecular G-quadruplexes. Surprisingly, the G-tracts remained protected (Figure 2B, lanes 2 and 4, red and green arrowheads). This indicated that G-quadruplexes could form in these two DNAs, even though they beard less than four G-tracts. Therefore, we inferred that the structure might have recruited G-tracts from the RNA transcript.

HQ formation was further analyzed by native gel electrophoresis. In this assay, dsDNA that carries a G-quadruplex migrates slower than the same DNA without a G-quadruplex (11). Figure 2C shows the results obtained from DNAs that carried four (4G), three (3G) or two (2G) G-tracts. As expected (11), the 4G DNA heated in a K^+ solution formed intramolecular G-quadruplexes, as indicated by the appearance of a slow-migrating band (Figure 2C, lane 2, top gel). Without RNase and protease treatment, the transcribed DNA appeared as a few super-shifted bands (Figure 2C, lane 3, 5 and 7) that might reflect complex DNA/RNA/protein interactions. When such interactions were eliminated by degradation with RNase and protease, the 4G DNA also showed a slow-migrating band (Figure 2C, lane 4 top gel) that, however, was more retarded than the band containing an intramolecular G-quadruplex prepared by heating (Figure 2C, lane 2 top gel). This additional retardation could be interpreted as most likely because of the presence of an RNA fragment with the DNA in an HQ structure; this interpretation was supported by its resistance to RNase A digestion (Supplementary Figure S2). In contrast, Li^+ does not stabilize G-quadruplex (17); therefore, the G-quadruplex band did not appear when transcription was conducted in a Li^+ solution (Figure 2C, lane 6, top gel).

To examine possible participation of RNA transcript in the HQ formation in the 4G DNA, we carried out transcriptions in which GTP was substituted with 7-deaza-GTP (dzGTP). In dzGTP, the 7-nitrogen (Figure 1B), essential for forming a G-quartet (18), is replaced with a carbon, which cannot form a hydrogen bond. Therefore, RNA participation in G-quadruplex formation was inhibited. Under this condition, we observed a slowly migrating band (Figure 2C, lane 8, top gel) that showed the same migration as the heated DNA that contained an intramolecular G-quadruplex (Figure 2C, lane 2, top gel). This indicated that an intramolecular DNA quadruplex formed without the participation of RNA. The intensity of this band was much weaker than the HQ band (Figure 2C, lane 4, top gel); this implied that the formation of the intramolecular DNA G-quadruplex might be suppressed, and that the formation of the HQ might have been favored in the transcription with GTP. One possible reason for this could be that folding four G-tracts from two sources would experience weaker sequence constraint than from a single strand. In consistence of this, the co-transcriptional formation of HQ was also implied with many other sequences that carried four G-tracts from different genes (Supplementary Figure S3).

In principle, co-transcriptional HQ formation requires at least two G-tracts on the non-template strand. In agreement with the protection from DMS footprinting (Figure 2B), transcription of the 3G and 2G DNAs resulted in a K^+-dependent slowly migrating bands (Figure 2C, middle and bottom gels, lane 4 versus lane 1). These bands should be represented as HQs, because, in theory, intramolecular DNA G-quadruplexes could not form with these sequences. Substituting GTP with dzGTP to prevent the participation of RNA totally abolished G-quadruplex formation (Figure 2C, lane 8, middle and bottom gels); this indicated that the RNA transcript was required for the G-quadruplex formation. As expected, DNA carrying one G-tract did not show a retarded band because it was unable to form a G-quadruplex of any kind (Supplementary Figure S4).

To verify the participation of the RNA transcript in HQ formation with DNAs that have fewer than four G-tracts, 4-thio-uridine was incorporated into the RNA during transcription, then UV-activated to cross-link nucleotides in the non-template DNA strand (12). Formation of an HQ (Figure 1C) should bring the participating G-tracts in the non-template DNA and those in the RNA transcript into close proximity. We anticipated that this would permit interstrand cross-linking between the two strands, which could be detected with primer extension on the non-template strand (12). Figure 2D shows the results for the 3G DNA, which required at least one G-tract from RNA to form a G-quadruplex. The three contiguous high-density bands in the transcribed sample (Figure 2D, lane 2) showed that the three nucleotides (CTA) in the non-template strand just downstream of the G-core were clearly cross-linked, presumably to the two 4-thio-uridines just downstream of the G-core in the RNA transcript. This result indicated that the RNA was in close proximity to the non-template DNA strand at the cross-linked sites near the HQ.

In the aforementioned experiments, we proved that the formation of HQ involved G-tracts from both the non-template DNA strand and RNA transcript. To confirm the four-stranded architecture of an HQ, we next show that preventing one G-tract from participation abolishes HQ formation. We first transcribed a dsDNA carrying a GGGTGGG motif in the non-template strand with a fully matched CCCACCC motif in the template strand. In this case, transcription generated RNA bearing a GGGUGGG motif that can be used to form HQ. The transcription produced expected HQ as indicated by the appearance of a retarded band (Figure 2E, top gel). When a single mutation was introduced to one G-tract in the non-template strand without changing the template, HQ formation was not observed, even though the RNA transcript still had two G-tracts (Figure 2E, middle gel). This result clearly shows that the participation of the non-template DNA strand was required. Similarly, a single mutation in one G-tract in the template strand resulted in an availability of two G-tracts from the non-template and one G-tract from an RNA. This also abolished HQ formation (Figure 2E, bottom gel), showing the participation of RNA was necessary. Together, these results confirm that the assembly of the HQ was four-stranded and required a joint participation of G-tracts from both the non-template DNA strand and RNA transcript.

Co-transcriptional HQ formation in plasmid and suppression of transcription

To better understand HQ formation and its effect on transcription, the NRAS sequence and mutants were inserted into a pGL3 plasmid downstream of a T7 promoter (Figure 3A) and upstream of a T7 terminator. In this case, G-quadruplex formation was detected by a ligand-induced photocleavage footprinting technique that we recently developed (13). This technique uses tetrakis (2-trimethylaminoethylethanol) phthalocyaninato zinc tetraiodine (Zn-TTAPc), a ligand that binds G-quadruplex with high affinity and selectivity (19–25). UV irradiation of this ligand causes preferential cleavage of the guanines in G-quadruplexes (13). Figure 3B shows that HQs were clearly formed, as indicated by the cleavage bands in the region corresponding to the NRAS G-core with two or three G-tracts (2G and 3G). Moreover, the cleavage increased with increasing numbers of G-tracts, and the highest degree of cleavage was observed in the 4G DNA. This implied that the HQ formation was enhanced with increases in the number of G-tracts. In transcriptions of linear dsDNA (Figure 2), polyethylene glycol (PEG) was added to stabilize G-quadruplex formation in dsDNA, which facilitated detection (11). In the plasmids, G-quadruplex formation was detected in both the absence and presence of PEG (Figure 3B). The ability to sustain the G-quadruplex without PEG may be explained by the superhelicity in the plasmids, which facilitates G-quadruplex formation (26).

Figure 3C shows the effect of HQ on transcription in the absence and presence of Zn-TTAPc, which stabilizes G-quadruplexes (19,20). In the presence of GTP, which permitted the formation of HQ, the number of transcripts that reached the T7 terminator sequence decreased with increases in the number of G-tracts (Figure 3C, top panel). When dzGTP was used in place of GTP to prevent HQ formation, transcription was rescued. This suggested that the inhibition was dependent on the formation of HQ structures. Transcription was more significantly inhibited (Figure 3C, bottom panel) in the presence of Zn-TTAPc because of its stabilization of G-quadruplexes (19,20). When GTP was replaced with dzGTP in the presence of Zn-TTAPc, clear inhibition was observed for plasmids that carried three or four G-tracts (Figure 3C, bottom panel). This contrasted with the lack of inhibition observed with dzGTP in the absence of Zn-TTAPc

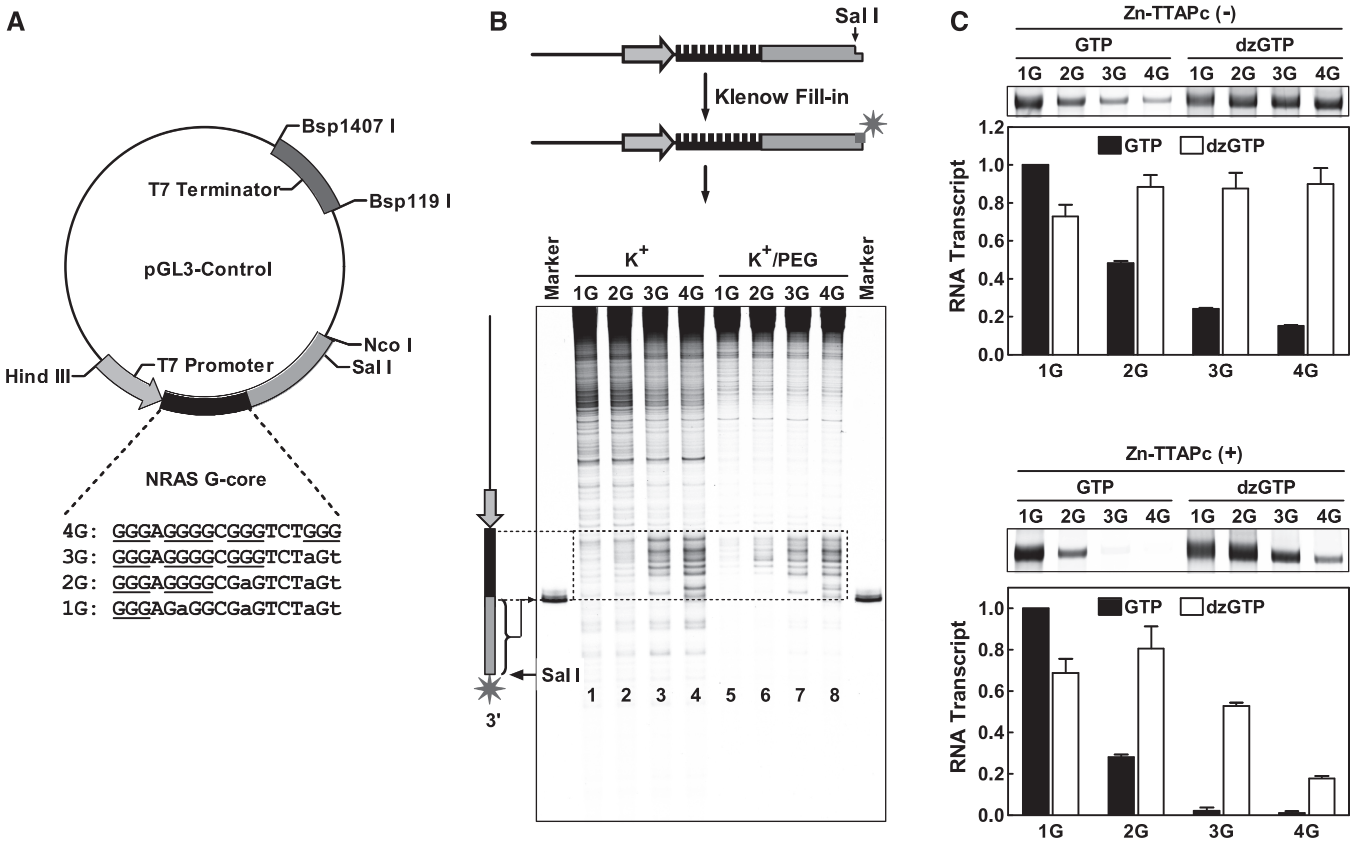

Figure 3. HQ formation during transcription of a supercoiled plasmid that contained a G-core motif from the NRAS gene and its effect on transcription. (**A**) Structure of the plasmid. (**B**) HQ detection by ligand-induced photocleavage footprinting. The plasmids, transcribed in a K^+ solution in the absence or presence of PEG 200, were subjected to photocleavage, then cut at the SalI restriction site and filled-in at the 3′-end with fluorescein-dUTP. Cleavage fragments were resolved by denaturing gel electrophoresis. The marker was a 3′-fluorescein labeled single-stranded synthetic DNA equivalent to the fragment between the SalI site and the 3′-end of the NRAS G-core (gray bar). (**C**) Inhibition of transcription by HQ in the absence (top panel) or presence (bottom panel) of 2 μM G-quadruplex-stabilizer Zn-TTAPc. The transcription was conducted with either GTP or dzGTP and the other three NTPs. Transcripts were analyzed with denaturing gel electrophoresis. The RNA bands that terminated at the T7 terminator were normalized to the same band in the 1G sample produced with GTP.

(Figure 3C, top panel). We speculated that the Zn-TTAPc might stabilize the formation of HQ by recruiting three DNA and one RNA G-tracts. Under those conditions, the HQ would lose only one hydrogen bond out of eight for each G-quartet (Figure 1B).

HQ inhibits reporter gene expression in transfection plasmids in human cells

Experimental studies on specific genes, such as the c-kit, bcl-2, c-myc and kras, have shown that intramolecular G-quadruplexes in promoter regions affect transcription (27–30). In particular, such G-quadruplexes have been shown to hinder translocation of helicase (31) and cause transcription arrest (32) when located in front of a moving protein. It has been proposed that G-quadruplexes may act as a steric block to repress transcription (33). Our *in vitro* results given in Figure 3C suggest that HQ may also have similar function to transcription. Based on this, we further investigated how HQs might affect transcription under *in vivo* conditions.

We inserted sequences from NRAS that carried one to four G-tracts (1G–4G) into a pGL3 luciferase reporter vector, at 17-nt downstream of a SV40 promoter (Figure 4A, right scheme). The vector was then transfected into HEK293 cells, and reporter gene expression was analyzed. A control plasmid was co-transfected to calibrate the variation in transfection efficiency, variation in cell lysis and lysate stability and other experimental factors. The results showed that the 2G–4G sequences, which are able to form HQ, inhibited luciferase expression (Figure 4A, left panel). Moreover, this inhibition was intensified in the presence of the G-quadruplex-stabilizing ligand, Zn-TTAPc. This supported the involvement of G-quadruplex in the inhibition of reporter expression. However, the degree of inhibition was less significant than that observed *in vitro* (Figure 3C). This might be attributed to the presence of other factors in cells, like G-quadruplex-resolving proteins (34), regulation at the translation level and other regulatory activities. In fact, the *in vivo* results may have reflected a combination of different effects from many possible sources. Increasing the number of G-tracts from two to three led to increased inhibition (Figure 4A, 2G versus 3G). This correlated with an increased probability of HQ formation (Figures 2C and 3B). However, this trend did not extend to the 4G sequence. The decreased potency of the 4G sequence here may be explained by the possibility of intramolecular G-quadruplex formation. In the *in vitro* results in Figure 3C, the 4G DNA produced little inhibition in transcription with dzGTP, in which HQ formation was prevented, but intramolecular DNA G-quadruplex formation was permitted. These results (Figures 3C and 4A) suggested that the HQ might be more potent than the non-hybrid intramolecular DNA G-quadruplex for inhibiting transcription, but further study is required for confirmation. Nevertheless, the inhibition of the reporter gene expression implied that HQ could function in eukaryotic transcription *in vivo*.

Prevalence of PHQS in human genome and its correlation with transcriptomal profiles of tissues

To survey the presence of PHQS in genome, we carried out a computational search for PHQS in protein-coding

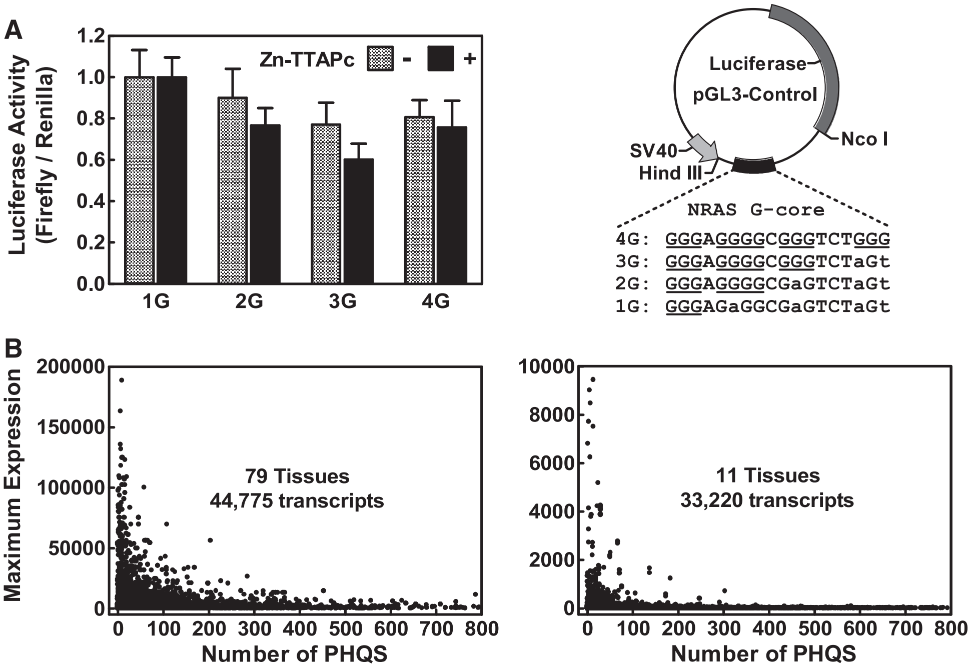

Figure 4. Effect of PHQS on gene expression in human cells and tissues. (**A**) Comparison of luciferase reporter activities among vectors that carried one to four G-tracts (1G, 2G, 3G and 4G) from the NRAS gene and were transfected into HEK293 cells. Luciferase activity was first normalized to the internal control; then it was expressed relative to that of the 1G sample. Data are mean ±S.D of two independent experiments, where each transfection was carried out in triplet. (**B**) Correlations between the maximal expression value and the number of PHQSs in genes in 79 and 11 human tissues, respectively. The maximal expression value for each gene among the indicated number of tissues was plotted against the number of PHQSs in the gene.

genes in the human genome using a home-made perl script. The search algorithm found all G-core motifs that matched the pattern $G_{\geq 3}(N_{1\text{–}7}G_{\geq 3})_{\geq 1}$; that is, two or more G-tracts of three or more guanines connected by a loop of 1–7 nt. The resulting PHQS motifs were then grouped into four categories, designated 1G, 2G, 3G and 4G+, in which the PHQS contains 1, 2, 3 and ≥4 G-tracts, respectively (Supplementary Figure S5). Table 1 summarizes the results from 22 058 human protein-coding genes. PHQSs were found in >97% of genes, with an average of >73 PHQSs per gene.

Given the effect of HQ observed in the aforementioned *in vitro* transcription and *in vivo* expression assays, we next explored the relationship between PHQS and *in vivo* expression of human genes with an analysis (Figure 4B) of two independent sets of human transcriptome profiles. One transcriptome covered 44 775 transcripts in 79 tissues (14) and the other covered 33 220 transcripts in 11 tissues (15). The expression of these genes in tissues is optimized for the *in vivo* status of each tissue under complex regulation at different levels. Therefore, we did not expect a correlation between the PHQS and the expression level of all genes. However, when we plotted the maximal expression value of each gene against the number of PHQSs in the gene, we found that the genes with high-level RNA expression carried low numbers of PHQSs. Conversely, the genes with large numbers of PHQSs always exhibited low expression. This finding suggested that transcription was inhibited by the PHQS and was consistent with our *in vitro* (Figure 3C) and *in vivo* (Figure 4A) results. We speculate that the number of PHQSs may determine or limit the expression potential of a gene.

CONCLUSION

Our discovery of co-transcriptionally formed HQ structures reveals a previously neglected, but an abundant form of G-quadruplexes in genome. The results in this work and those in our ongoing studies (data not shown) show that HQ readily forms in transcription of DNA with two or more G-tracts of three or more continuous guanines on the non-template strand (Figures 2 and 3 and Supplementary Figure S3). This is a general phenomenon rather than a situation associated to a specific sequence. Synthetic G-rich DNA and RNA oligomer forms HQ when they are mixed and incubated together (9,10), or delivered together into cells (35). With all the techniques available to us to identify the structure, our results show that HQ also forms in a dynamic physiological process. This fact establishes a connection between G-quadruplex formation and transcription, a fundamental event inside cells. The effect of HQ on transcription also demonstrates the biological relevance and significance of G-quadruplex structures. Our additional bioinformatic analyses show that PHQS began to enrich in the genomes of amphibians and became constitutional in genes in warm-blooded animals. They are preferentially concentrated in the immediate 1000-nt region downstream of transcription start sites on the non-template strand, suggestive of a positive selection for the formation and function of HQ in transcription. These results will be presented in a separate work.

Our results show that the formation of an HQ can directly suppress transcription (Figures 3C and 4A). This effect was intrinsic to transcription because it involved only the transcribed DNA, RNA and polymerase, without the participation of third-party factors. Therefore, it is built-in or encoded in a gene directly, and it is executed only when the host gene is transcribed. This results in a simple, efficient and economical autologous control of gene expression. The genome-wide presence of PHQS in human genes implied that the mechanism was not limited to specific genes or pathways. We hypothesize that HQs may provide a general, primary *cis* control at the root level of transcription to limit the expression potential of a host gene. This hypothesis was supported by our results from the *in vivo* gene expression (Figure 4A) and transcriptomal analysis (Figure 4B).

Table 1. PHQS found in the non-template strand of 22 058 human protein-coding genes

G-tracts	All	1G[a]	2G	3G	4G+
Positive sequences	21 437	7261	21 348	19 463	16 652
Per cent of positive sequences	97.18	32.92	96.78	88.24	75.49
Total PHQS found	1 613 107	13 937	1 251 030	245 194	102 946
Mean PHQS/sequence	73.13	0.63	56.72	11.12	4.67

[a] Those motifs that consists of one long tract of contiguous guanines and satisfies the pattern $G_{\geq 3}(N_{1\text{–}7}G_{\geq 3})_{\geq 1}$. Each of them can be clarified into one or more of the other four categories. For example, a G_7 can be regarded as a 2G sequence by taking the G in the middle as loop.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1, Supplementary Figures 1–5 and Supplementary perl source code.

FUNDING

Ministry of Science and Technology of China [2013CB530802, 2012CB720601 and 2010CB945300]; National Science Foundation of China [30970617 and 21072189]. Funding for open access charge: Ministry of Science and Technology of China [2012CB720601].

Conflict of interest statement. None declared.

REFERENCES

1. Burge,S., Parkinson,G.N., Hazel,P., Todd,A.K. and Neidle,S. (2006) Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res.*, **34**, 5402–5415.
2. Neidle,S. and Parkinson,G. (2002) Telomere maintenance as a target for anticancer drug discovery. *Nat. Rev. Drug Discov.*, **1**, 383–393.
3. Collie,G.W. and Parkinson,G.N. (2011) The application of DNA and RNA G-quadruplexes to therapeutic medicines. *Chem. Soc. Rev.*, **40**, 5867–5892.

4. Balasubramanian,S., Hurley,L.H. and Neidle,S. (2011) Targeting G-quadruplexes in gene promoters: a novel anticancer strategy? *Nat. Rev. Drug Discov.*, **10**, 261–275.
5. Todd,A.K. (2007) Bioinformatics approaches to quadruplex sequence location. *Methods*, **43**, 246–251.
6. Huppert,J.L. (2010) Structure, location and interactions of G-quadruplexes. *FEBS J.*, **277**, 3452–3458.
7. Huppert,J.L. and Balasubramanian,S. (2007) G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.*, **35**, 406–413.
8. Biffi,G., Tannahill,D., McCafferty,J. and Balasubramanian,S. (2013) Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat. Chem.*, **5**, 182–185.
9. Xu,Y., Kimura,T. and Komiyama,M. (2008) Human telomere RNA and DNA form an intermolecular G-quadruplex. *Nucleic Acids Symp. Ser.*, 169–170.
10. Wanrooij,P.H., Uhler,J.P., Shi,Y., Westerlund,F., Falkenberg,M. and Gustafsson,C.M. (2012) A hybrid G-quadruplex structure formed between RNA and DNA explains the extraordinary stability of the mitochondrial R-loop. *Nucleic Acids Res.*, **40**, 10334–10344.
11. Zheng,K.W., Chen,Z., Hao,Y.H. and Tan,Z. (2010) Molecular crowding creates an essential environment for the formation of stable G-quadruplexes in long double-stranded DNA. *Nucleic Acids Res.*, **38**, 327–338.
12. Harris,M.E. and Christian,E.L. (2009) RNA crosslinking methods. *Methods Enzymol.*, **468**, 127–146.
13. Zheng,K.W., Zhang,D., Zhang,L.X., Hao,Y.H., Zhou,X. and Tan,Z. (2011) Dissecting the strand folding orientation and formation of G-quadruplexes in single- and double-stranded nucleic acids by ligand-induced photocleavage footprinting. *J. Am. Chem. Soc.*, **133**, 1475–1483.
14. Su,A.I., Wiltshire,T., Batalov,S., Lapp,H., Ching,K.A., Block,D., Zhang,J., Soden,R., Hayakawa,M., Kreiman,G. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
15. Castle,J.C., Armour,C.D., Lower,M., Haynor,D., Biery,M., Bouzek,H., Chen,R., Jackson,S., Johnson,J.M., Rohl,C.A. *et al.* (2010) Digital genome-wide ncRNA expression, including SnoRNAs, across 11 human tissues using polyA-neutral amplification. *PLoS One*, **5**, e11779.
16. Sun,D. and Hurley,L.H. (2010) Biochemical techniques for the characterization of G-quadruplex structures: EMSA, DMS footprinting, and DNA polymerase stop assay. *Methods Mol. Biol.*, **608**, 65–79.
17. Hardin,C.C., Perry,A.G. and White,K. (2000) Thermodynamic and kinetic characterization of the dissociation and assembly of quadruplex nucleic acids. *Biopolymers*, **56**, 147–194.
18. Fletcher,T.M., Sun,D., Salazar,M. and Hurley,L.H. (1998) Effect of DNA secondary structure on human telomerase activity. *Biochemistry*, **37**, 5536–5541.
19. Ren,L., Zhang,A., Huang,J., Wang,P., Weng,X., Zhang,L., Liang,F., Tan,Z. and Zhou,X. (2007) Quaternary ammonium zinc phthalocyanine: inhibiting telomerase by stabilizing G quadruplexes and inducing G-quadruplex structure transition and formation. *Chembiochem*, **8**, 775–780.
20. Zhang,L., Huang,J., Ren,L., Bai,M., Wu,L., Zhai,B. and Zhou,X. (2008) Synthesis and evaluation of cationic phthalocyanine derivatives as potential inhibitors of telomerase. *Bioorg. Med. Chem.*, **16**, 303–312.
21. Yaku,H., Murashima,T., Miyoshi,D. and Sugimoto,N. (2012) Specific binding of anionic porphyrin and phthalocyanine to the G-quadruplex with a variety of *in vitro* and *in vivo* applications. *Molecules*, **17**, 10586–10613.
22. Yaku,H., Fujimoto,T., Murashima,T., Miyoshi,D. and Sugimoto,N. (2012) Phthalocyanines: a new class of G-quadruplex-ligands with many potential applications. *Chem. Commun.*, **48**, 6203–6216.
23. Yaku,H., Murashima,T., Miyoshi,D. and Sugimoto,N. (2010) Anionic phthalocyanines targeting G-quadruplexes and inhibiting telomerase activity in the presence of excessive DNA duplexes. *Chem. Commun.*, **46**, 5740–5742.
24. Alzeer,J. and Luedtke,N.W. (2010) pH-mediated fluorescence and G-quadruplex binding of amido phthalocyanines. *Biochemistry*, **49**, 4339–4348.
25. Alzeer,J., Vummidi,B.R., Roth,P.J. and Luedtke,N.W. (2009) Guanidinium-modified phthalocyanines as high-affinity G-quadruplex fluorescent probes and transcriptional regulators. *Angew. Chem. Int. Ed. Engl.*, **48**, 9362–9365.
26. Sun,D. and Hurley,L.H. (2009) The importance of negative superhelicity in inducing the formation of G-quadruplex and i-motif structures in the c-Myc promoter: implications for drug targeting and control of gene expression. *J. Med. Chem.*, **52**, 2863–2874.
27. McLuckie,K.I., Waller,Z.A., Sanders,D.A., Alves,D., Rodriguez,R., Dash,J., McKenzie,G.J., Venkitaraman,A.R. and Balasubramanian,S. (2011) G-quadruplex-binding benzo[a]phenoxazines down-regulate c-KIT expression in human gastric carcinoma cells. *J. Am. Chem. Soc.*, **133**, 2658–2663.
28. Wang,X.D., Ou,T.M., Lu,Y.J., Li,Z., Xu,Z., Xi,C., Tan,J.H., Huang,S.L., An,L.K., Li,D. *et al.* (2010) Turning off transcription of the bcl-2 gene by stabilizing the bcl-2 promoter quadruplex with quindoline derivatives. *J. Med. Chem.*, **53**, 4390–4398.
29. Tian,M., Zhang,X., Li,Y., Ju,Y., Xiang,J., Zhao,C. and Tang,Y. (2010) Inducement of G-quadruplex DNA forming and down-regulation of oncogene c-myc by bile acid-amino acid conjugate-BAA. *Nucleosides Nucleotides Nucleic Acids*, **29**, 190–199.
30. Cogoi,S. and Xodo,L.E. (2006) G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription. *Nucleic Acids Res.*, **34**, 2536–2549.
31. Liu,J.Q., Chen,C.Y., Xue,Y., Hao,Y.H. and Tan,Z. (2010) G-quadruplex hinders translocation of BLM helicase on DNA: a real-time fluorescence spectroscopic unwinding study and comparison with duplex substrates. *J. Am. Chem. Soc.*, **132**, 10521–10527.
32. Broxson,C., Beckett,J. and Tornaletti,S. (2011) Transcription arrest by a G quadruplex forming-trinucleotide repeat sequence from the human c-myb gene. *Biochemistry*, **50**, 4162–4172.
33. Huppert,J.L. (2008) Four-stranded nucleic acids: structure, function and targeting of G-quadruplexes. *Chem. Soc. Rev.*, **37**, 1375–1384.
34. Paeschke,K., Capra,J.A. and Zakian,V.A. (2011) DNA replication through G-quadruplex motifs is promoted by the *Saccharomyces cerevisiae* Pif1 DNA helicase. *Cell*, **145**, 678–691.
35. Xu,Y., Ishizuka,T., Yang,J., Ito,K., Katada,H., Komiyama,M. and Hayashi,T. (2012) Oligonucleotide models of telomeric DNA and RNA form a Hybrid G-quadruplex structure as a potential component of telomeres. *J. Biol. Chem.*, **287**, 41787–41796.