11706–11720 *Nucleic Acids Research, 2020, Vol. 48, No. 20* *Published online 12 October 2020*
doi: 10.1093/nar/gkaa841

Detection of genomic G-quadruplexes in living cells using a small artificial protein

Ke-wei Zheng[1,2,†], Jia-yu Zhang[1,3,*,†], Yi-de He[1,4,†], Jia-yuan Gong[1], Cui-jiao Wen[1], Juan-nan Chen[2], Yu-hua Hao[1], Yong Zhao[4] and Zheng Tan[1,5,*]

[1]State Key Laboratory of Membrane Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, P.R. China, [2]School of Pharmaceutical Sciences (Shenzhen), Sun Yat-Sen University, Guangzhou 510275, P.R. China, [3]CAS Key Laboratory for Biomedical Effects of Nanomaterials and Nanosafety, Multidisciplinary Research Division, Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, P.R. China, [4]School of Life Sciences, Sun Yat-Sen University, Guangzhou 510006, P.R. China and [5]Center for Healthy Aging, Changzhi Medical College, Changzhi 046000, Shanxi, P.R. China

Received June 13, 2020; Revised September 16, 2020; Editorial Decision September 18, 2020; Accepted September 19, 2020

ABSTRACT

G-quadruplex (G4) structures formed by guanine-rich nucleic acids are implicated in essential physiological and pathological processes and serve as important drug targets. The genome-wide detection of G4s in living cells is important for exploring the functional role of G4s but has not yet been achieved due to the lack of a suitable G4 probe. Here we report an artificial 6.7 kDa G4 probe (G4P) protein that binds G4s with high affinity and specificity. We used it to capture G4s in living human, mouse, and chicken cells with the ChIP-Seq technique, yielding genome-wide landscape as well as details on the positions, frequencies, and sequence identities of G4 formation in these cells. Our results indicate that transcription is accompanied by a robust formation of G4s in genes. In human cells, we detected up to >123 000 G4P peaks, of which >1/3 had a fold increase of ≥5 and were present in >60% promoters and ∼70% genes. Being much smaller than a scFv antibody (27 kDa) or even a nanobody (12–15 kDa), we expect that the G4P may find diverse applications in biology, medicine, and molecular devices as a G4 affinity agent.

INTRODUCTION

G-quadruplexes (G4s) are four-stranded secondary structures formed by guanine-rich nucleic acids. Putative G-quadruplex forming sequences (PQSs) are abundant in the genomes of animal cells with particular enrichment near transcription start sites (TSSs), implying an essential role of G4s in transcription (1–5). Accordingly, they are emerging as a new class of drug targets for pharmaceutical applications. As such, the detection and quantitation of G4s in genomes with sequence identity are indispensable for exploring the biological and pathological function of G4s. The formation of G4s in cells becomes the molecular basis to justify their physiological role. Over the past two decades, small molecule ligands (6–19) and proteins (20–22) capable of interacting with G4s have been used to report the existence of G4s in cells (for a recent review, see (23)). Of these tools, many small ligands detect G4s in living cells by the fluorescence produced when they bind to targets but do to provide information on the identity and quantity of the targets. On the other hand, detection of G4s with native G4-interacting proteins may experience non-specificities since they may also interact with non-G4 components in cells. There are chances for them to be brought to targets indirectly or subject to complex interactions, resulting in non-specificity or impeded recognition.

Recently, G4s have been detected in chemically fixed human cells by immunostaining (24) or immunoprecipitation (25,26) with a G4 antibody. In these applications, the binding of antibodies to the G4s occurs in a non-native cellular environment after a series of treatments. It is not known how G4s are preserved or affected during the permeabilization, staining, fragmentation and/or DNA purification (27–29) in the duplex genome DNA in which the hybridization of two complementary DNA strands competes against the G4s. A recent study has shown that G4 DNA in the non-transcribed strand of R-loop is lost rapidly upon removal of RNA using RNAse H (30) even in an environment in which DNA hybridization is significantly weakened and G4 simultaneously stabilized (31). For these reasons, G4 identification in living cells is desired as an alternative option for the G4 community. Unfortunately, antibodies are not suitable

*To whom correspondence should be addressed. Tel: +86 355 3030833; Fax: +86 355 3030833; Email: z.tan@ioz.ac.cn
Correspondence may also be addressed to Jia-yu Zhang. Email: zhangjy86@ihep.ac.cn
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

for living cells because they do not permeate into cells. Besides, the reducing environment of the cytoplasm of living cells is not compatible with the formation of the disulfide bonds required for maintaining the tertiary structure of antibodies (32).

To overcome the aforementioned difficulties, we engineered a 64 amino acids (64-aa) 6.7 kDa protein based on a 23-aa segment (RHAU23) of the G4-binding domain of the RHAU (also named DHX36 or G4R1), an RNA helicase able to bind and unwind G4s (33). This G4 probe (G4P) comprises only two small G4-binding domains linked by a flexible and optimized linker in between. Such a simple composition minimizes potential non-specific interactions with other proteins. Moreover, the synergy between two binding domains dramatically improves affinity and selectivity towards G4s. Expression of the G4P in cells followed by chromatin immunoprecipitation (termed G4P-ChIP) allowed us to capture G4s in living human, mouse, and chicken cells through the ChIP-Seq technique, revealing genome-wide landscape and details on the locations, frequencies, and sequence identities of G4 formation in these cells.

MATERIALS AND METHODS

Plasmid construction for G4P-ChIP

DNA coding the RHAU23-(GTGSGA)-3xFLAG (RHAU 23) or RHAU23-(GTGSGA)3-RHAU23-(GTGSGA)-3x FLAG were synthesized by Generay Biotechnology (Shanghai, China) and inserted into pET28b between the Nde I and EcoR I sites to obtain pET28b-RHAU23 and pET28b-G4P. The former coding sequence was also inserted into pIRES2-EGFP between the Nhe I and EcoR I sites to obtain pG4P-IRES2-EGFP. The DNA fragment containing a nuclear localization signal (NLS) of the SV40 large antigen (PKKKRKV) was synthesized by Sangon (Shanghai, China) and inserted into pG4P-IRES2-EGFP at the Nhe I site to obtain plasmid pNLS-G4P-IRES2-EGFP. The construction of AAVS1 donor and PX330 plasmids for knock-in was as described (34). Briefly, the DNA fragment for expressing NLS-G4P and eGFP was amplified from the pNLS-G4P-IRES2-EGFP and inserted into the AAVS1 donor plasmid between the Spe I and Sal I sites. The AAVS1 loci specific guide RNA sequence (5′-GTCACCAATCCT GTCCCTAG-3′) was designed by the online CRISPR tool (crispr.mit.edu) and inserted into the PX330 plasmid between two Bbs I sites (35,36).

Cell lines and cell culture conditions

A549, NCI-H1975, 293T, HeLa-S3, 3T3 and DF-1 cells were kindly provided by Stem Cell Bank, Chinese Academy of Sciences. Cells were grown in DMEM supplemented with 10% FBS and 1 × penicillin-streptomycin.

Transient transfection and knock-in of G4P

For transient transfection, cells were cultured in 15 cm dishes to 70–80% confluence and transfected with 30 μg plasmid pNLS-G4P-IRES2-EGFP using lipofectamine 3000 (Thermo Scientific) according to the manufacturer's instructions. Cells were cultured for an additional 24 h before harvesting. For G4P knock-in, AAVS1 donor and PX330 plasmid containing G4P and AAVS1 gRNA were co-transfected into 293T cells using lipofectamine 2000 (Thermo Scientific). After 24 h, GFP positive single cell was sorted by a flow cytometer (MoFlo XDP, Beckman) into a 96-well plate. Cells were cultured for two weeks and the cell lines with stable expression of G4P were verified by PCR, western blot, and immunofluorescence.

Recombinant G4P and RHAU23

The plasmid pET28b-G4P and pET28b-RHAU23 were transformed into the E.coli strain BL21 (DE3). Cells were grown in LB medium supplemented with 0.05 mg/ml kanamycin at 37°C for 3 h. G4P and RHAU23 expression were induced with 1 mM isopropyl thiogalactoside for 4 h at 37°C. The two peptides were purified using the Capturem HIS-tagged Purification Miniprep Kit (Takara, Dalian) and diluted in buffer containing 20 mM Tris–HCl (pH 7.4), 150 mM NaCl, 0.1 mM EDTA, 50% glycerol and stored at −20°C.

Circular dichroism (CD) spectroscopy

CD spectroscopy was conducted as described (37). The incubation of 1 μM dsDNA or ssDNA (Supplementary Tables S1 and S3) with G4P was carried out at RT and time interval with a 1:1 protein/DNA ratio.

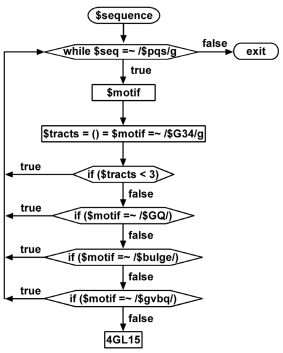
Electrophoretic mobility shift assay (EMSA)

DNAs (Supplementary Tables S1 and S3) were dissolved at 10 nM in a buffer containing 20 mM Tris–HCl (pH 7.4), 75 mM KCl, 1 mM EDTA,0.4 mg/ml BSA, without or with (in case of dsDNA) 30% (w/v) PEG 200, denatured at 95°C for 5 min, and slowly cooled down to 25°C. DNA was then incubated with G4P of the indicating concentration at 4°C or the indicated temperature for 1 h. Samples were resolved on 12% non-denaturing polyacrylamide gel containing 75 mM KCl in the absence or presence (in case of dsDNA) of 40% (w/v) PEG200 at 4°C for 2 h in 1× TBE buffer containing 75 mM KCl. DNA was visualized by the FAM dye covalently labeled at the 5′ end of the DNA on a Chemi-Doc MP (Bio-Rad) and digitized using the Image Quant 5.2 software. Dissociation constant was determined by fitting the fractional bound DNA (Y) in a 1:1 stoichiometry against the free G4P concentration (X) to the equation: $Y = F1 \times X/(K_{d1} + X) + (1 - F1) \times X/(K_{d2} + X)$, where $F1$ denoted the fraction of the subpopulation associated with the K_{d1}.

Identification of PQS motifs in genomes

Chromosome files (hg19 for human, mm9 for mouse, and gal5 for chicken) in fasta format were downloaded from the UCSC website (http://genome.ucsc.edu/). Global searching of four types of PQS motifs (i.e. 4G, 4GL15, Bulge and GVBQ) was performed using home-made Perl scripts. 4G PQS motifs were identified using a regular expression G{3,}(.{1, 7}?G{3, }){3,} as previously described (38,39). Identification of the other three PQSs involved a first round of search to find motifs with the

desired feature. Each motif found was then subjected to several additional rounds of pattern matching to remove unwanted motifs. A flow chart and the variables for the identification of 4GL15 PQS is shown below as an example.




```
my $L7 = ".{1,7}?";
my $G3 = "G{3,}";
my $G34 = "G{3,4}";
my $bulgeG = "GG[^G]GG?|GG?[^G]GG";
my $GQ = "$G3($L7$G3){3,}";
my $nGLX = "[^G].{0,6}?";
my $LXnG = ".{0,6}?[^G]";
my $flankGL = "(GG?[^G]|($bulgeG)$LXnG)";
my $flankGR = "([^G]GG?|$nGLX($bulgeG))";
my $longLoopG = "$G3.{1,15}?$G3";
my $longLoop1G =              "($longLoopG)($L7$G3){2,}";
my $longLoop2G = "($G3$L7){3,}($longLoopG)($L7$G3){1,}";
my $longLoop3G = "($G3$L7){3,}($longLoopG)";

my $bulge =
"(($bulgeG)($L7$G3){3}|($G3$L7)($bulgeG)($L7$G3){2}|($G3$L7
){2}($bulgeG)($L7$G3)|($G3$L7){3}($bulgeG)}";

my $gvbq =
"(GG($nGLX$G3|G{4})($L7$G3|G{4}){2}|($G3$LXnG|G{4})GG($nGLX
$G3|G{4})($L7$G3|G{4})|($G3$L7|G{4})($G3$LXnG|G{4})GG($nGLX
$G3|G{4})|($G3$L7|G{4}){2}($G3$LXnG|G{4})GG)";

my $longLoop = "($longLoop1G|$longLoop2G|$longLoop3G)";

my $pqs = "(?<!G)($flankGL)?($longLoop)($flankGR)?(?!G)";
```

Pull down of G4-containing plasmid with G4P

Plasmids containing the indicated PQS or a mutant control sequence (Supplementary Table S2) on the non-template strand were constructed as described (40). Transcription was carried out in a total volume of 50 μl at 37°C for 1 h in transcription buffer containing 40 mM Tris–HCl (pH 7.9 at 25°C), 8 mM $MgCl_2$, 10 mM DTT, 2 mM spermidine, 50 mM KCl, 200 U T7 RNA Polymerase (Thermo Scientific), 40 U RNase inhibitor (Thermo Scientific), 2 mM NTP, 0.5 μg pEGFP-N1 plasmid, and 0.5 μg PQS or control plasmid. The reaction was stopped by addition of EDTA to 16 mM. 1 μl of samples were used as input and diluted in 50 μl water. The remaining samples were mixed with 0.2 μM G4P and incubated at 4°C for 30 min.

For the pull-down assay, 12 μl anti-FLAG M2 magnetic beads (Sigma-Aldrich) were washed three times with binding buffer (20 mM Tris–HCl, pH 7.4, 150 mM KCl, 1 mM EDTA and 0.5% Triton X-100) and then resuspended in 200 μl binding buffer. The beads were then incubated with the DNA samples on a rotating mixer at 4°C for 2 h. After the removal of the supernatant, the beads were washed ten times with the binding buffer accompanied with three times of transfers to new tubes. The precipitated plasmids were released from the beads by incubating with 0.3 mg/ml 3xFLAG peptide (Sigma-Aldrich) in a 50 μl binding buffer at 4°C for 0.5 hours. The released plasmid and the input samples were treated with RNase A at 37°C for 15 min and proteinase K at 60°C for 20 min, followed by inactivation at 95°C for 10 min.

qPCR was conducted on the CFX Connect thermocycler (Bio-Rad) for the PQS or control plasmid using a PCR primer pair of 5′-TCCGACACTATGCCATCCTG A-3′ and 5′-TGCAGTCGTTTCGTATCGTTGA-3′. The EGFP gene on the pEGFP-N1 plasmid was amplified using a primer pair of 5′-GCAGAAGAACGGCATCAAGG-3′ and 5′-CGGACTGGGTGCTCAGGTAG-3′. After calibrating with the internal control, the enrichment was calculated using input DNA as a reference.

G4P-ChIP library construction

Approximately 0.5–1 × 10^8 transiently or stably transfected cells expressing G4P were crosslinked with 1% formaldehyde for 20 min at room temperature. Fixation was quenched by 0.125 M glycine for 15 min. The fixed cells were washed twice with PBS, suspended in NP-40 buffer (10 mM Tris–HCl pH 7.4, 150 mM NaCl, 0.5% NP-40 and 2 mM AEBSF) and incubated on ice for 10 min. After centrifugation at 800 × g for 5 min, the cell pellet was resuspended in a CHAPS buffer (20 mM Tris-HCl pH 7.4, 0.5 mM EGTA, 50 mM NaCl, 0.5% CHAPS, 10% glycerol and 2 mM AEBSF). The suspension was incubated on ice for 30 min and centrifuged at 800 × g for 5 min. The pellet was resuspended in 1 ml 1× dsDNase digestion buffer supplied with 50 μl dsDNase (Invitrogen, EN0771) and incubated at 37°C for 20 min with constant agitation. A final concentration of 20 mM EDTA was added to terminate the reaction. The samples were pelleted by centrifugation at 15 000 × g at 4°C and the resulting supernatant was collected and incubated on ice. The pellet was resuspended in 500 μl wash buffer (150 mM NaCl, 10 mM Tris–HCl, pH 7.4, 0.1 mM EDTA, 0.5% Triton X-100) and sonicated for 30–60 s in an ice-cold water bath. After centrifugation at 15 000 × g for 5 min, the supernatant of chromatin fragment was collected and combined with the supernatant from the previous step.

For library preparation, 50 μl of anti-FLAG M2 magnetic beads (Sigma-Aldrich) were washed with washing buffer (10 mM Tris–HCl, pH 8.0, 150 mM NaCl and 0.5% Triton X-100) and blocked in the same buffer containing 75 μg/ml single-stranded sperm DNA and 1 mg/ml BSA. 1% chromatin fragment was saved as input and the remaining was incubated with blocked anti-FLAG magnetic beads in rotation at 4°C for 3 h. The beads were sequentially washed ten times with washing buffer and transferred to new tubes three times. The chromatin was eluted with 300 μg/ml 3xFLAG peptide (Sigma-Aldrich) at 4°C for 1 h. The eluted chromatin and the input samples were incubated with proteinase K at 65°C overnight. After sequential RNase A and proteinase K digestion, DNA fragment was cleaned by extraction with phenol:chloroform:isoamyl alcohol, followed by ethanol precipitation. Libraries were constructed from the recovered DNA fragment using the NEBNext Ultra II DNA LibraryPrep Kit from Illumina (NEB) according to the manufacturer's instructions. The next-generation sequencing was performed with Illumina HiSeq X Ten by Genewiz (Suzhou, China).

ChIP-Seq data analysis

Clean paired-end sequencing data in fastq format were mapped to the human genome (hg19) using the Bowtie2 software (41) with the sensitive-local preset and –no-unal, –no-discordant, –no-mixed parameters. Mapped reads were written to bam files after being filtered by the Samtools view (42) to remove low-quality alignments with the parameter –q 20 and by Samtools rmdup to remove duplicates. Reads enrichment was calculated using the Deeptools (43) plotEnrichment with the bam files. Reads bam files were also processed by the Deeptools bamCompare to produce bigwig coverage files in subtract or ratio mode and normalized to RPKM. Profiles and heatmaps of reads were gener-

ated from the bigwig files using the Deeptools computeMatrix followed by plotProfile and plotHeatmap, respectively, with region bed files derived from the NCBI RefSeq bed file downloaded from the UCSC website (http://genome.ucsc.edu/) unless otherwise indicated. Coordinate duplicates in the bed files were removed. Peaks of reads enrichment were identified with the macs2 software (44) using –*q*value 0.001, –keep-dup 1, and default values for the other parameters. ChIP-Seq data from public repositories were downloaded from the GEO (https://www.ncbi.nlm.nih.gov/geo/) or Encode (https://www.encodeproject.org/) database and processed as described above whenever applicable.

RNA-Seq

Total RNA from cells expressing G4P was prepared using the SV Total RNA Isolation System (Promega). mRNA was purified from total RNA using the NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB). RNA library was constructed using the NEBNext Ultra II Directional RNA Library Prep Kit from Illumina (NEB) according to the manufacturer's instructions. The next-generation sequencing was performed with Illumina HiSeq X Ten by Genewiz (Suzhou, China).

ChIP-qPCR

Primers for amplifying PQS-proximal and distal regions were given in Supplementary Table S5. qPCR reaction was performed using the GoTaq qPCR Master Mix (Promega) and qTOWER 2.2. The cycling condition was 95°C for 20 s followed by 45 cycles of 30 s at 95°C and 30 s at 60°C. The enrichment of the genomic locus in the chip sample relative to the input was calculated using double delta Ct analysis with a PQS negative region as references.

Expression of nucleolin

Expression and purification of nucleolin (NCL) were conducted as described (45).

RESULTS

G4P binds canonical and non-canonical G4s *in vitro*

Our G4P core is a 64-aa engineered protein (Figure 1A) composed of two identical 23-aa G4-binding domains (RHAU23) from the RHAU protein that has a preference to bind parallel G4s with a dissociation constant (K_d) of ∼1 μM (33). The RHAU23 contained a 13-aa RHAU-specific motif that functions as a major determinant for the affinity and specificity toward G4s in the original RHAU (46). According to the solution structure of the RHAU23-G4 complex, each of the two terminal guanine quartets (G-quartets) of a G4 can bind an RHAU23 (33). Therefore, the two RHAU23 units in the G4P are expected to clamp onto the two terminal-G-quartets of a G4, reinforcing the binding through a synergy between two binding domains (Figure 1B). We added a HIS and a 3xFLAG tag at the N- and C-terminal of the G4P core, respectively (Figure 1A). This 113-aa G4P was expressed in *E. coli* and purified by affinity chromatography using the HIS tag.

Besides the canonical G4s, there are three types of well-characterized non-canonical G4s, i.e. G4s with one loop of 8–15 nucleotides (4GL15) (47), G-vacancy-bearing G4s (GVBQ) (48), G4s with a bulge of one non-G nucleotide (Bulge) (49), respectively. We first assessed the binding activity of the G4P to several representative G4s (Supplementary Table S1) by the electrophoretic mobility shift assay (EMSA) (33). For all the G4 DNAs tested, effective mobility shift was observed above the original DNA band in positive correlation to G4P concentration (Figure 1C and Supplementary Figure S1A, arrowhead, gels), which is in contrast to the RHAU23 that largely failed to bind the G4s (Supplementary Figure S1B). The tags did not affect the binding of G4P to G4s (Supplementary Figure S1C). Similar to the RHAU23, the G4P did not bind the non-G4 DNAs, including i-motif sequences, yeast tRNA, single-stranded DNA (ssDNA), and double-stranded DNA (dsDNA) (Figure 1D, Supplementary Figure S2).

In some cases, we noticed that the G4P did not bind all the DNAs in a sample even at saturating concentration (Figure 1C and Supplementary Figure S1A). For example, G4P bound only 20% of the C9orf72 and Tel DNAs. Because these DNAs have been reported to adopt an antiparallel (50) and mixed parallel/antiparallel (51) conformation, respectively, that is not favored by the original RHAU23 (33), there was a possibility that the 20% of the DNAs instead adopted a conformation recognized by the G4P. For the C9orf72, a formation of hairpin structures in equilibrium with the G4s (52) might also contributed to the reduced fraction of G4P binding. These results indicated that the conformations for a given DNA might be heterogeneous and a fraction of the G4s could still be recognized by the G4P. Therefore, we fitted the EMSA data to a two-population binding model for simplicity to derive the dissociation constant K_d. High-affinity K_d at low-nM was obtained for each G4 DNA with various fractional quantities (Figure 1C, Supplementary Figure S1A). The results also suggested that G4P could bind more than one folding conformations with distinctive affinities as demonstrated by the PDGFR-β G4s that exhibited two distinct K_d values.

We next examined G4 binding under a more physiologically relevant condition to pull down transcriptionally generated G4 in supercoiled plasmids (Figure 1E, Supplementary Table S2). The plasmid accommodated a CSTB, c-MYC, or a mutated motif on the non-template or the template strand. Transcription with T7 RNA polymerase efficiently induces a formation of G4 on the non-template but marginally on the template strand (53). Accordingly, the G4s on the non-template strand led to a high enrichment of the corresponding plasmid. A marginal enrichment of the plasmid with the mutant motif on the non-template strand might be attributed to the possible formation of a weak DNA:RNA hybrid G4 involving the G_2 and G_4 tracts from the DNA and RNA transcript (38,39).

The formation of the G4 in the plasmid was unlikely induced by the G4P because G4 was not detected when the CSTB motif was placed on the template strand, which we previously described as a strand-biased G4 formation in transcription (53). Furthermore, G4 was not detected by circular dichroism (CD) spectroscopy (Figure 2A) and EMSA (Figure 2B, lanes 3 and 5) when a CSTB-containing

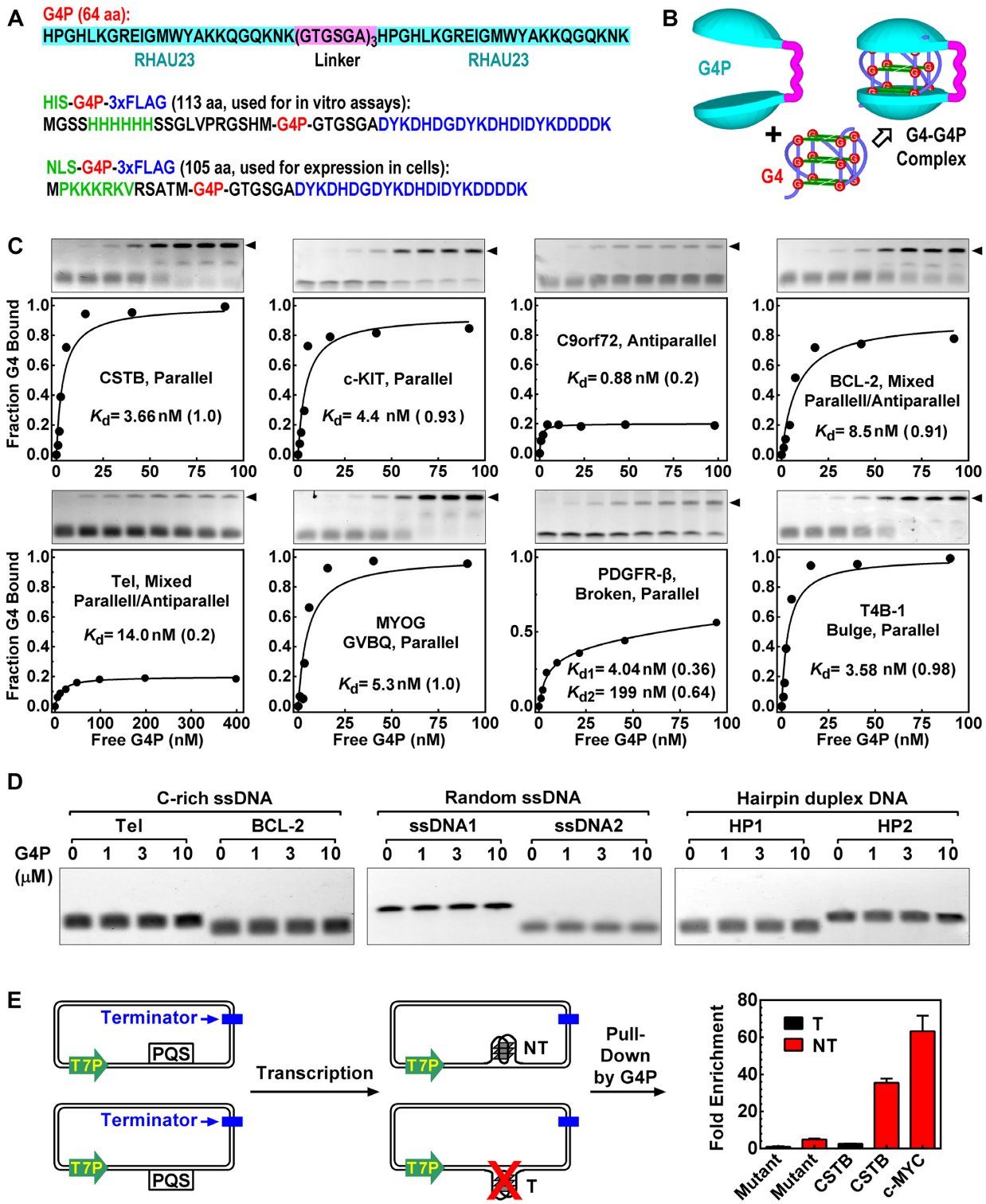


Figure 1. G4P binds G4s *in vitro*. (**A**) The amino acid sequence of G4P. The one with a nuclear localization sequence (NLS) was used for detecting G4s in cells. The one with a HIS tag was for *in vitro* assays. A 3xFLAG at the C-terminal was for binding with the anti-FLAG antibody. (**B**) Anticipated clamping-binding of a G4 by a G4P. (**C**) Dissociation constant K_d of G4P to G4s of different folding conformation determined by EMSA. K_d values were obtained by fitting the data of each DNA (Supplementary Table S1) to a model assuming the existence of two subpopulations of G4s. The number in parenthesis indicates the fraction of the G4 subpopulation associated with the corresponding K_d. One K_d is given when the fitting produced two identical K_d values or the other is >1 M. (**D**) G4P does not bind non-G4 DNA. Details of DNA are in Supplementary Table S1. (**E**) Pull-down of transcriptionally generated G4 in plasmids by G4P. A PQS or mutant motif (Supplementary Table S2) was placed on either the non-template (NT) or template (T) strand downstream of a T7 promoter (T7P) in a plasmid. The plasmid transcribed with T7 RNA polymerase was incubated with G4P. G4P-G4 complex was then pulled-down in the presence of an internal reference plasmid using anti-FLAG beads and G4-bearing plasmid was quantitated by qPCR relative to the internal plasmid. Fold enrichment was normalized to the mutant motif on the template strand.

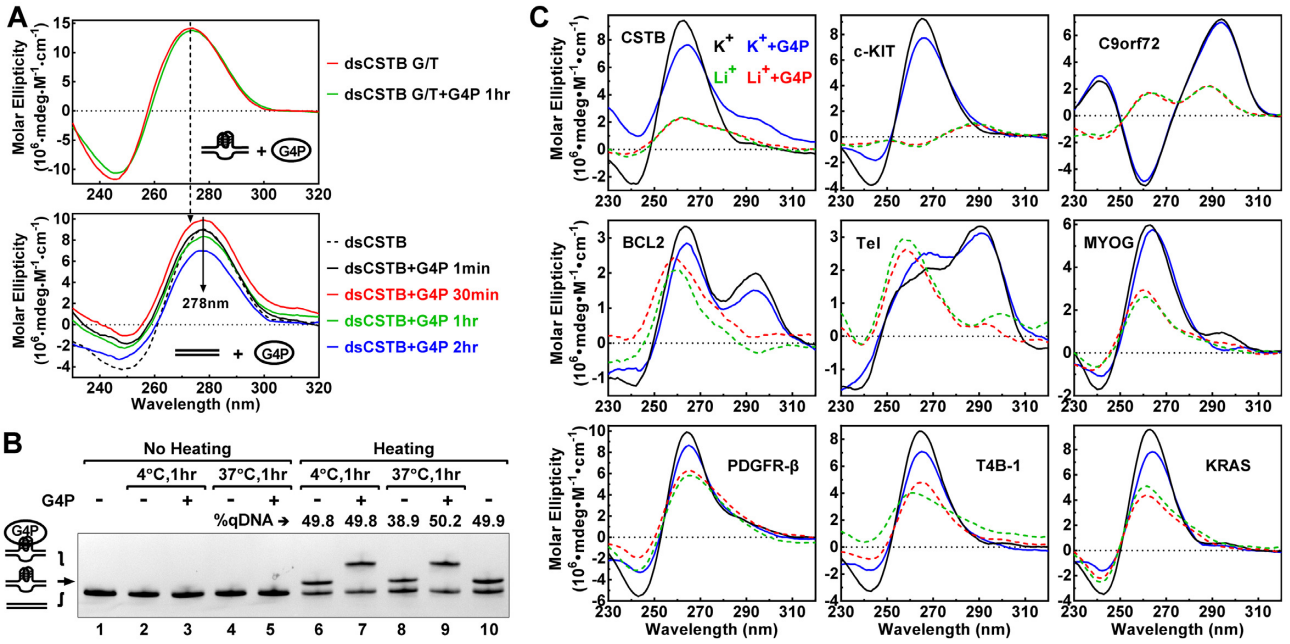


Figure 2. G4P did not induce a formation of G4 in dsDNA and ssDNA. (**A**) Circular dichroism (CD) spectra of a CSTB PQS-containing dsDNA in the absence and presence of G4P at 1:1 molar ratio to DNA. The DNA strands in the top panel were not complementary to each other within the PQS region to enable G4 formation. The DNA strands (Supplementary Table S3) in the bottom panel were fully complementary to each other. They were incubated with or without G4P for the indicated time at RT. (**B**) The dsDNA used was heated to generate G4 and then incubated without or with G4P at 4 or 37°C, respectively, for 1 hour before EMSA. At 4°C, the percentage of G4-bearing DNA (qDNA) remained unchanged when the DNA was left free or bound by G4P (lanes 6 and 7 versus 10). At 37°C, the percentage of qDNA decreased by ∼20% when the DNA was left free but remain unchanged when bound by G4P (lanes 8 and 9 versus 10). (**C**) CD spectra of G4 structures in the absence or presence of G4P. ssDNAs (Supplementary Table S1) were incubated without or with G4P at 1:1 molar ratio at RT for one hour in 150 mM Li^+ or K^+ solution. G4s are stabilized by K^+ but not by Li^+.

dsDNA (Supplementary Table S3) was incubated with G4P. Instead, the G4P only captured the G4s that already existed (Figure 2B, lanes 7 and 9 versus 3 and 5). G4s are characterized by peaks at unique wavelength in CD spectra depending on their folding topology. When we incubated G4P with single-stranded PQS motifs in a Li^+ solution, a condition that does not support G4 formation, the CD spectra of the DNAs remained virtually unchanged and remained different from the corresponding spectra in the K^+ solution (Figure 2C, dashed curves), indicating that the G4P did not induce a formation of G4 in these ssDNAs. The failure to observe the induction of G4 in both the ssDNA and dsDNA was in agreement with the fact that the G4P strictly recognized G4s. We did not expect an induction of G4 by the G4P without it recognizing a non-G4 target. Notably, the G4s formed in the K^+ solution maintained their folding topology after they were bound by G4Ps (Figure 2C, solid curves), which ensured that the recognition remained even in the G4P-G4 complexes.

G4P recognizes G4s in living cells

To detect G4s in living cells, we added a nuclear localization sequence (NLS) and a 3xFLAG tag at the N- and C-terminal of the G4P core, respectively, without compromising the binding to G4s (Supplementary Figure S1C). We then expressed this 105-aa G4P in cultured human A549 cells by transfection with a plasmid. The expressed G4P was destined to nuclei as intended (Supplementary Figure S3). ChIP-Seq was performed on the G4P–G4 complex (Figure 3A). The G4P reads were mapped to the human genome using the Bowtie2 (41) and high-quality reads ($-q = 20$) were subsequently obtained using the Samtools (42) and then their distribution on genome analyzed with the Deeptools (43). The binding of the G4P to G4s was first demonstrated by the enrichment of G4P reads mapped to the canonical PQS motifs defined by a consensus of $G_{\geq 3}(N_{1-7}G_{\geq 3})_{\geq 3}$ (38,54) (Figure 3B). We term this type of PQSs 4G PQSs to distinguish them from the non-canonical ones. The enrichment disappeared when the coordinates of the 4G PQSs were shuffled to random 4G PQS-free locations using the Bedtools ShuffleBed (55) and the reads around these non-PQS regions counted. Since PQSs are concentrated around TSSs where DNA is more accessible to fragmentation than other regions, the input reads at real 4G PQSs was greater than at the shuffled. The binding of G4P to G4s was further illustrated by a peak when the G4P reads were profiled around the center of the 4G PQSs, which appeared only in the G4P-ChIP but not so in the input sample (Figure 3C). The tiny peak in the input could be attributed to a greater openness of the binding site than the other regions, as 'features of open chromatin' in ChIP-Seq (56). Similarly, shuffling the coordinates of the 4G PQSs also removed the difference between the two samples. As a routine independent validation, we observed an enrichment of G4P at the 4G PQSs by ChIP-qPCR (Supplementary Figure S4).

To find out where G4 forms in the genome, we surveyed G4 formation in the four functional regions, i.e. TSS± 2 kb (promoter), intragenic, TES ± 2 kb, and intergenic regions (Figure 3D), where TSS and TES represented transcription

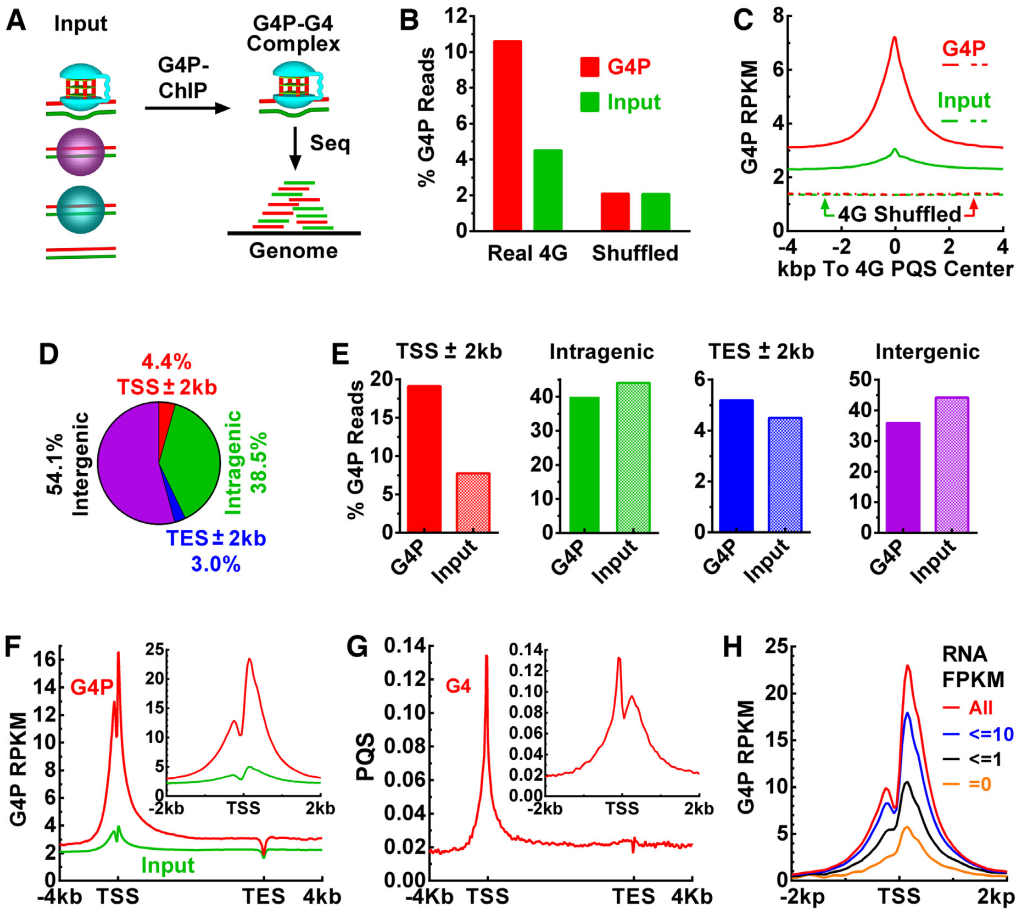


Figure 3. G4P binds genomic G4s in human A549 cells. (**A**) G4P-G4 complex was enriched from fragmented input DNA by chromatin immunoprecipitation (G4P-ChIP) and then subjected to sequencing (Seq) to identify G4 formation in the genome. (**B**) G4P reads at 4G PQSs as a percent of the total reads mapped to the genome. (**C**) Enrichment of G4P at 4G PQSs. (**D**) Percent nucleotides and (**E**) percent G4P reads in genomic regions. (**F**) G4P reads and (**G**) PQSs distribution across RefSeq genes and around TSSs (insert). (**H**) G4P reads distribution around TSS in genes of different RNA expression levels. The signal in (**F**) is normalized to RPKM and G4P signal in (**H**) is expressed as (G4P reads) – (input reads) and normalized to RPKM.

start and end site, respectively. The G4P in the TSS±2kb region counted >19% of the total mapped reads, more than twice the input in this region, which was in contrast to the other three regions (Figure 3E). This result indicated that the gene promoters were hot sites of G4 formation in the human genome. We further profiled G4P reads distribution in the RefSeq genes and found that G4s were concentrated at both sides of TSSs, with two peaks merging at the TSSs (Figure 3F). The distribution largely correlated with the presence of 4G PQSs across the TSSs (Figure 3G) and agreed with our previous *in vitro* studies in which transcription efficiently induced G4s at both sides of a TSS (37–39,45,48,53,57).

The distribution of the two features between TSS and TES were distorted due to the scaling on the gene bodies of different sizes. A better correlation was seen when the profiles were plotted across TSS only (Figure 3, F and G, inserts). It is noted that the extent of G4 formation was smaller at the upstream side than at the downstream side of the TSSs while the occurrence of PQS was the opposite. This phenomenon could be explained by the asymmetrical transcription activities crossing TSS. For example, synthesis of RNA occurs at the downstream side, generating an RNA:DNA hybrid structure known as R-loop (58), which has been shown to reinforce transcriptional G4 formation by suppressing the hybridization of the duplex DNA (30). *In vitro* transcriptional formation of G4s positively correlates with transcription activity (30,59). To find out the *in vivo* situation, we grouped TSSs according to their expression level and found greater G4 formation was associated with higher gene expression (Figure 3H).

G4P detects non-canonical G4s

As a natural deduction, more PQS motifs should lead to more G4 formation in a region. We thus plotted G4P profiles as a function of the number of 4G PQSs in the TSS±3kb regions. The reads density decreased as we gradually excluded the regions with a greater number of 4G PQS (Figure 4A). However, when all the 4G PQS-positive regions were excluded, the peak only dropped by ∼60% in the remaining ∼1/3 of the TSSs. This fact suggested that the canonical 4G PQSs did not account for all the G4s recognized by the G4P and other forms of G4s were also captured. We, therefore, searched for PQSs of 4GL15 (47), GVBQ (48), Bulge (49), respectively (Figure 4B). These

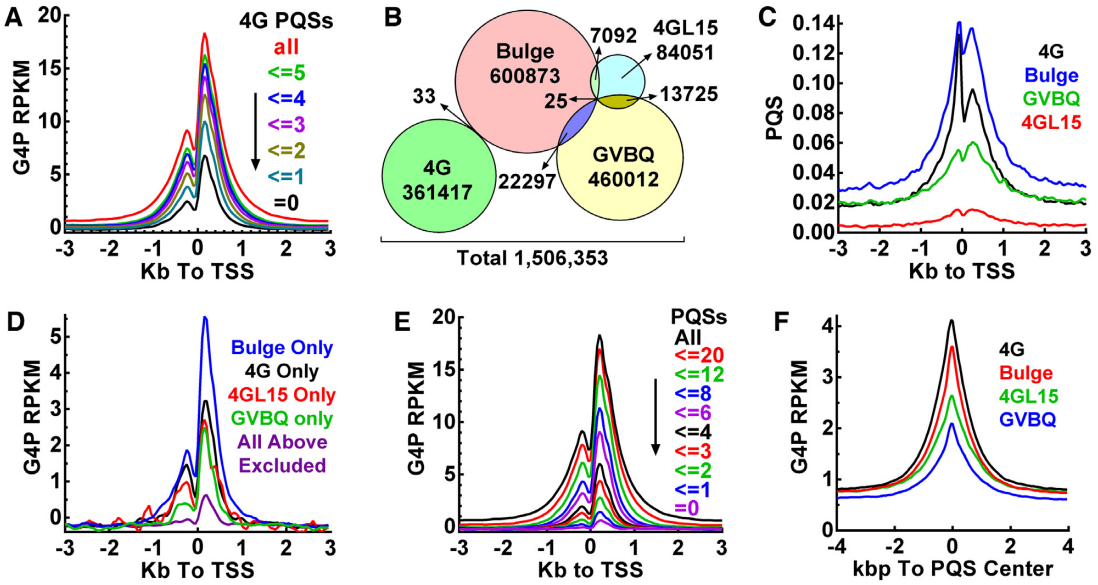


Figure 4. G4P binds canonical and non-canonical genomic G4s in human A549 cells. (**A**) Enrichment of G4P depends on the 4G PQS load. (**B**) Amount of four subtypes of PQSs in the human genome. Overlaps show motifs carrying more than one subtype of PQSs. (**C**) Distribution of 4G, 4GL15, GVBQ and Bulge PQSs around TSSs. (**D**) Distribution of G4P reads around TSSs with only the indicated subtype of PQS. (**E**) Enrichment of G4P depends on the load of four types of PQS. (**F**) Enrichment of G4P at subtypes of PQSs. G4P signal is expressed as (G4P reads) – (input reads) and normalized to RPKM. Numbers of PQS motifs within the TSS ± 3 kb region are given in panels A and E, respectively.

PQSs were similarly concentrated near TSSs as the 4G PQSs did (Figure 4C), suggesting that they were also evolutionally selected to form G4s and function in cells. Hence, they might all be targets of the G4P in cells as they were *in vitro* (Figure 1C and Supplementary Figure S1A).

To verify the detection of non-canonical G4s, we filtered out those TSSs carrying more than one subtype of PQSs in the TSS±3kb region and examined the contribution of each subtype. The result (Figure 4D) shows that G4P recognized each of them. When the four subtypes of PQSs were all excluded, the reads peak decreased to a very low level (Figure 4D, purple curve) that might represent recognition of other rare types of PQSs, for example, those with a loop larger than 15 nucleotides (47) or a bulge of more than one nucleotide (49). When the number of all the four subtypes of PQSs were counted, a positive correlation was also obtained between G4 formation and PQS load (Figure 4E). The enrichment of G4P at each subtype of PQS suggested that the ability of G4 formation followed the order of 4G>Bulge>4G15L>GVBQ (Figure 4F).

Overview and statistics of G4 formation in living cells

To overview the landscape of G4 formation in human genes, we generated a 3D heatmap of G4P around TSSs for the A549 cells. It showed that G4s formed in most of the genes near TSSs (Figure 5A) in the presence of PQSs (Figure 5B). A small fraction of TSSs showed an obvious enrichment of PQSs (Figure 5B, blue arrowhead) but with little G4P (Figure 5A), indicating an absence or low probability of G4 formation in a small fraction of the PQS motifs near the TSSs. We also expressed the G4P and detected G4s in cultured human NCI-H1975, HeLa-S3, 293T, mouse 3T3 and chicken DF-1 cells, respectively. G4 formation in these cell lines all displayed similar landscapes crossing TSSs (Supplementary Figure S5) and genes (Supplementary Figure S6A). It should be noted here that the TSSs in each heatmap were ordered independently to show an overview of a gradient of G4 formation at the promoter regions. The magnitude, number, and location of G4 formation could differ significantly in different cell lines (Supplementary Figures S6; S7, arrowheads) in a magnitude comparable to what was observed in fixed cells (25).

To identify individual G4-binding events, we performed peak calling on the G4P reads from the A549 cells using the MACS2 (44), which identified 123,274 G4P peaks (fold changes > 1.6) indicative of G4-G4P interactions. In Figure 5C, we present examples of G4 formation indicated by G4P peaks in association with the presence of different types of PQSs. More examples from other cell lines are given in Supplementary Figures S7–S9. From these examples, we found that G4s formed in many common loci regardless of cell lines (Supplementary Figure S7) and in some loci in a cell line-dependent manner (Supplementary Figure S7, arrowheads), which indicated that the G4P detected different G4 formation in different cell lines.

Among the 123 274 G4P peaks, 43 752 had a fold change ≥5 in signal relative to input (Figure 5D). The number of peaks in the four genomic regions dropped rapidly with an increase in the fold change of the peaks. Those peaks with high fold change should represent more stable G4s that form in higher frequency and/or last longer once formed. When the peaks in each genomic region were plotted as a percent of the total, those in the TSS±2kb region rapidly increased while those in the other three regions decreased (Figure 5E). This unique feature implied a more robust G4 formation in promoters driven by transcription. An obvious increment started at 4-fold, suggesting efficient G4 forma-

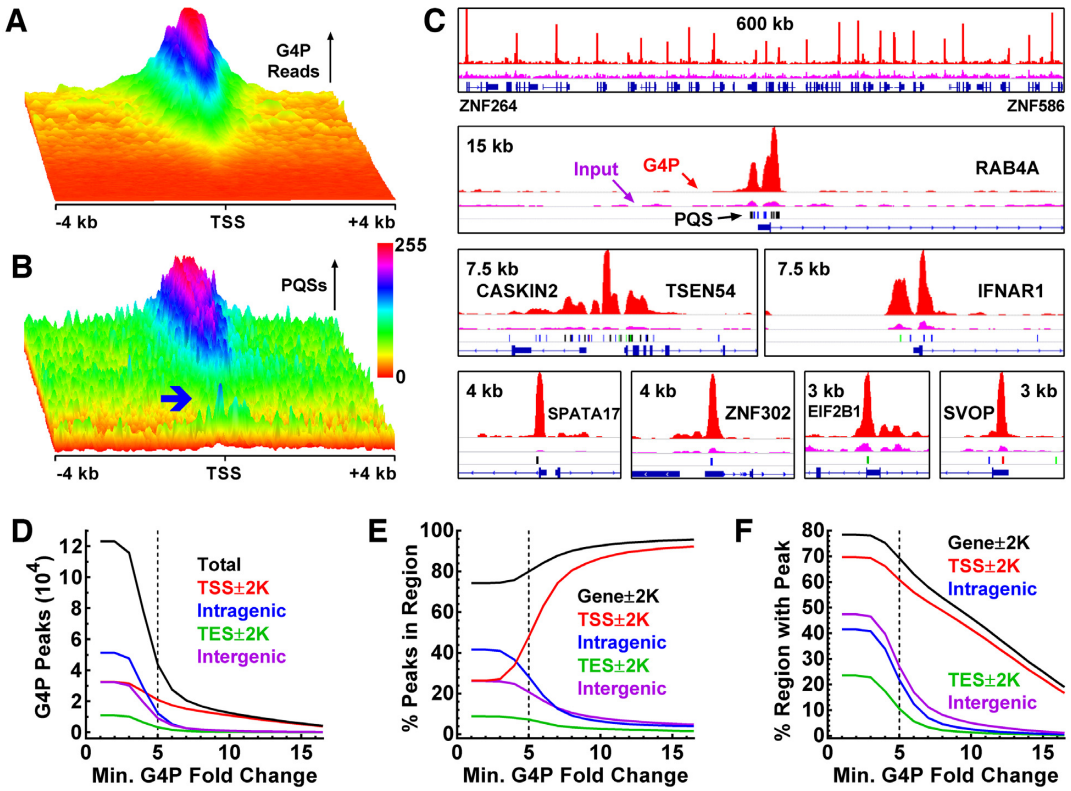


Figure 5. Overview and statistics of DNA G4 formation in human A549 cells. (**A**) G4P enrichment and (**B**) the presence of PQSs in the TSS±4kb regions of RefSeq genes showing G4 formed at where PQS existed. The two heatmaps were produced over the same TSS region file sorted in descending order on the maximum of G4P reads. The blue arrowhead shows a small fraction of PQS motifs with a low probability of G4 formation. Values in each heatmap were normalized independently to the range of 0–255 to enhance the visual resolution. (**C**) Examples of G4 formation depicted by G4P peaks. Reads signals for G4P and input are in red and magenta, respectively. Color bars beneath the input lane indicate 4G (black), 4GL15 (red), Bulge (blue) and GVBQ (green) motifs, respectively. Genomic range and genes are indicated in each panel. (**D–F**) Statistics of G4 formation as a function of G4P peak fold change.

tion at this enrichment occurred in response to transcription. Taking ≥5-fold as a threshold, ∼80% of the G4s were associated with genes (Figure 5E, gene ± 2 kb, black curve). Owing to the concentration of PQS at promoters, G4s were detected in a much greater fraction in the TSS ± 2 kb region than in the other three ones (Figure 5F). Counting peaks of ≥5-fold change, G4s were detected in >60% promoters (TSS ± 2 kb) and ∼70% genes. A summary Supplementary Table S4 provides details on the G4P peaks regarding their genomic coordinates, fold change, PQS motifs covered, and genes associated. Taken together, our data revealed that G4s readily form in association with transcription in living animal cells and promoters are their main playgrounds.

Co-localization of G4P with native G4-binding proteins

The identification of G4P binding peaks allowed us to compare the binding activity of the G4P among the human cell lines and with several endogenous native proteins, namely FUS, TAF15, RBM14 and TARDBP, that have been reported to bind DNA and RNA G4s *in vitro* (60–62). Thus, we plotted distribution profiles of the G4P and the four G4-binding proteins around the G4P peak regions from the A549 cells (Figure 6A, top panels). We can see that the G4P-binding signal was also present in the other three human cell lines although their magnitudes varied (Figure 6A, top panels 1–4). In addition, binding signals were also found for the four native G4-binding proteins in these G4P-binding regions (Figure 6A, top panels 5–8), suggesting they might co-localized with the G4P at their binding sites. The formation of G4s in one DNA strand leaves its complementary cytosine-rich (C-rich) partner single-stranded. In correlation with this, the binding signal was also seen for two such proteins, i.e. PCBP1 and hnRNP K (Figure 6A, top panels 9–10) that recognize single-stranded C-rich DNA (63,64).

To obtain more insight regarding the co-localization of these proteins, we sorted the G4P peak region file of the A549 cell line on the mean G4P reads in descending order and plotted heatmaps of the three groups of proteins over this G4P-binding region file. The results turned out that all the binding activities generally followed a similar gradient (Figure 6A, bottom panels 1–10) in a positive correlation with the presence of PQS motifs (Figure 6A, bottom panel 11). Clearly, these results showed that the G4P bound G4s as the native proteins did in cells, for which representative examples are shown in Figure 6B. They also revealed that a common landscape of G4 formation was largely shared among the different cell lines (Figure 6A, bottom panels 1–4) although their magnitudes varied (Figure 6A, top panels 1–4). Collectively, these observations demonstrated that G4P detected G4s that could form natively in the cells.

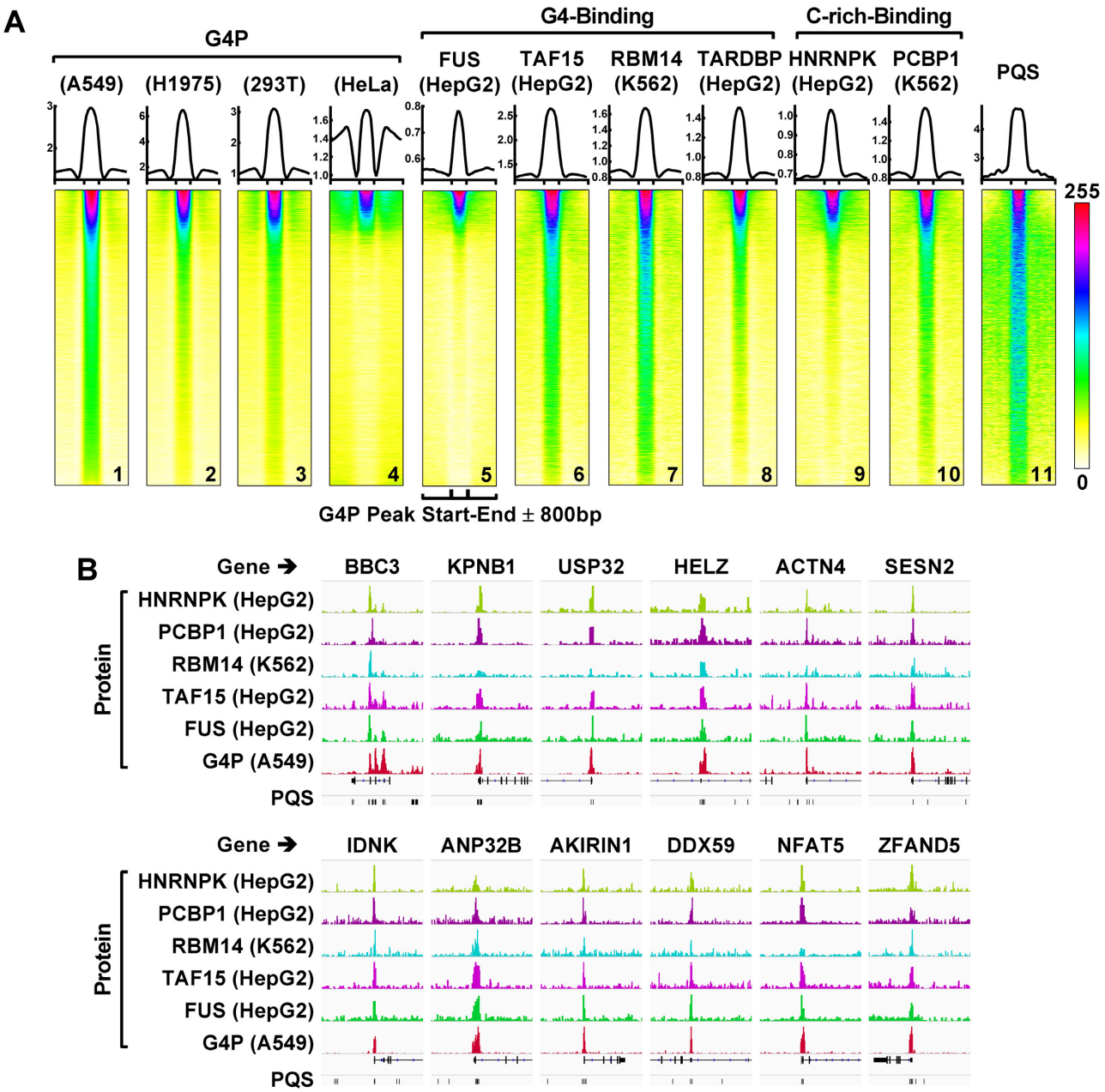


Figure 6. G4P binds genomic G4s in human cells as the indicated native G4-binding proteins do. (**A**) Profiles (top) and heatmaps (bottom) of proteins at the G4P peak ± 800 bp regions. The cell line is indicated in parenthesis. All heatmaps were produced over the G4P peak bed file of the A549 cells sorted on the mean G4P reads of the A549 cells in descending order. Values in each heatmap were normalized independently to the range of 0–255 to enhance the visual resolution. The bigwig files of the G4- and C-rich-binding proteins were from the Encode database (https://www.encodeproject.org/) with the following identifier numbers: ENCFF274FNN, ENCFF061RJA, ENCFF707TKH, ENCFF470CIC, ENCFF292JAR and ENCFF882OWS. (**B**) Examples of protein binding peaks at G4P regions.

G4s form in response to DNA negative supercoiling

RNA polymerases and other motor proteins, when moving along a DNA track, generate negative supercoiling torsion that under-twists a DNA duplex to break or weaken the inter-strand hydrogen bonds in the DNA. This torsion wave can propagate in a DNA helix for thousands of base pairs both *in vitro* and *in vivo* (45,65,66) and has been reported to efficiently trigger G4 formation during *in vitro* transcription (45,59). In agreement with this, the formation of G4s in living cells was associated with negative supercoiling (Figure 7A, top panel) in an inverse correlation with a higher degree of G4 formation in coincidence with a lower magnitude of supercoiling (Figure 7B, top panel). When the G4P reads and the negative supercoiling signal were plotted around G4P peak centers, the two features displayed almost identical profiles in a reversed direction (Figure 7A, bottom panel). The two sharp peaks at the G4P centers implied that G4s formed when the negative supercoiling reached the PQS motifs. A correlation was also obtained in the heatmaps (Figure 7B, bottom panel). These results argue against the possibility of G4 induction by the

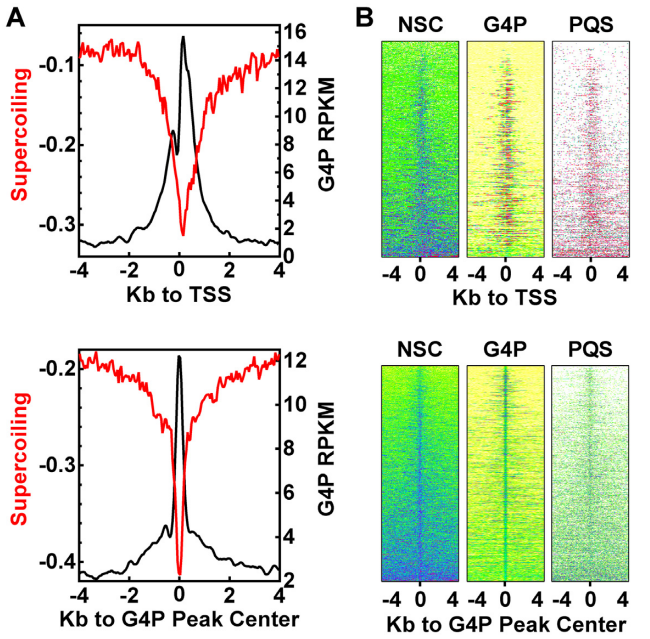


Figure 7. G4 formation coincides with negative supercoiling. (**A**) Profile of G4P (A549), negative supercoiling (NSC) around TSSs (top) and G4P peak centers (bottom). (**B**) Heatmap of G4P, NSC, and PQS within the ±4 kb regions around TSSs (top) and G4P peak centers (bottom) sorted on NSC signal. ChIP-chip data for NSC of Raji cells (GSE43751) in GFF format was downloaded from NCBI, converted to bigwig format, and processed as for that of the A549 cells.

G4P because the negative supercoiling occurred in the absence of G4P and could hardly be causally coupled with a G4P binding event in the cells expressing G4P.

Effect of G4P expression on transcription

Owing to the involvement of G4s in transcription, the expression of an exogenous G4-binding G4P in cells should affect gene expression. In return, the changes in gene expression brought by the G4P, if any, are expected to affect G4 formation in genes because of its dependence on transcription activity (Figure 3H). To evaluate the impact of G4P, we assessed the changes in gene expression by RNA-Seq (Supplementary Figure S11), revealing bi-directional changes in RNA level among genes. As shown in Supplementary Figure S11C, the changes were no more than 2-fold in ∼95% of the genes in the 293T cells with the G4P knock-in. For the plasmid-transfected cells, the RNA levels changed by less than 4-fold in >85% of the genes (Supplementary Figure S11), slightly greater than that in the 293T cells, which might be explained by a greater expression with the transfection. The lack of dependence of the variation in gene expression on the PQS load within the TSS±2kb regions agreed with a global impact on the transcription of the cells (Supplementary Figure S12).

In-living-cell versus in-fixed-cell detection

To briefly compare the in-living-cell and in-fixed-cell detection, we looked at the G4P reads distribution around TSS and the PQS coverage in the G4P peaks in all the cells we tested. The results revealed differences in three general aspects in comparison to the detection in fixed cells using the BG4 antibody (25) that also has nM level affinities to G4s as the G4P (51). First, the profiles of G4P reads all showed two peaks (Figure 8A) in correlation with the two peaks of PQSs (Figure 3G, insert) and the asymmetric transcription activity at TSS while those of the BG4 all appeared as a single symmetric peak centered at the TSSs (Figure 8B). Second, the enrichment peak of the G4P at the TSS was generally much higher than that of the BG4, meaning much more G4s were detected in living than in fixed cells. Third, the percentage of PQS-positive peaks detected by G4P (Figure 8C) was significantly greater than those detected by the BG4 (Figure 8D). These differences demonstrated significant improvements in resolution, sensitivity, specificity, and efficiency of the G4P in recognizing G4s. Despite the improvement of PQS coverage, ∼30% of the G4P peaks were PQS negative based on the four subtypes of PQS motifs (Figure 8C). Of these peaks, 63% contained motifs of two or three G-tracts that, according to our previous studies, might potentially form DNA:RNA hybrid G4s (30,37–40,57,67). When these hybrid G4 motifs were counted in, then 89% of the G4P peaks would be PQS-positive. This value could still be underestimated since motifs that could form, for example, intra-locked (68), Knot-like (69), zero-nucleotide loop (70), or other irregular (71) G4s were not considered. It should be noted here that the two approaches were conducted with different cell lines and different strategies, which might provide a more plausible explanation for the differences. For instance, the G4Ps in living cells could compete against native G4-binding proteins to bind G4s, thus detecting more G4s than the BG4s that were added to fixed cells and might only bind G4s not bound by native proteins. G4-protein complexes are crosslinked in fixed cells and such complexes may not be further recognized by G4-binding activities as exemplified in Supplementary Figure S10.

DISCUSSION

In summary, we developed a unique probe to capture G4 formation in living cells with the ability to reveal the locations, magnitudes, and sequence identities of G4s in a whole genome (Supplementary Table S4). Using this probe, we observed a robust formation of G4s in genes as a native event in transcription in animal cells (Figures 3–6), which establishes a molecular basis for the role of G4s in transcription. The positive correlation of G4 formation with transcription activity (Figure 3H) and PQS load (Figure 4, A and E) confers a capability for the G4-related regulation to be performed in response to the activity of transcription defined by the amount and type of PQS motifs. The similar landscapes of G4 formation shared among the different cell lines (Figure 5A, Supplementary Figure S5, and Figure 6) suggested the presence of a common, essential and likely constitutional mechanism involving G4s in transcription regulation.

Transcription is a driving force for G4 formation

Our previous studies have demonstrated that *in vitro* transcription efficiently induces G4 formation around a TSS

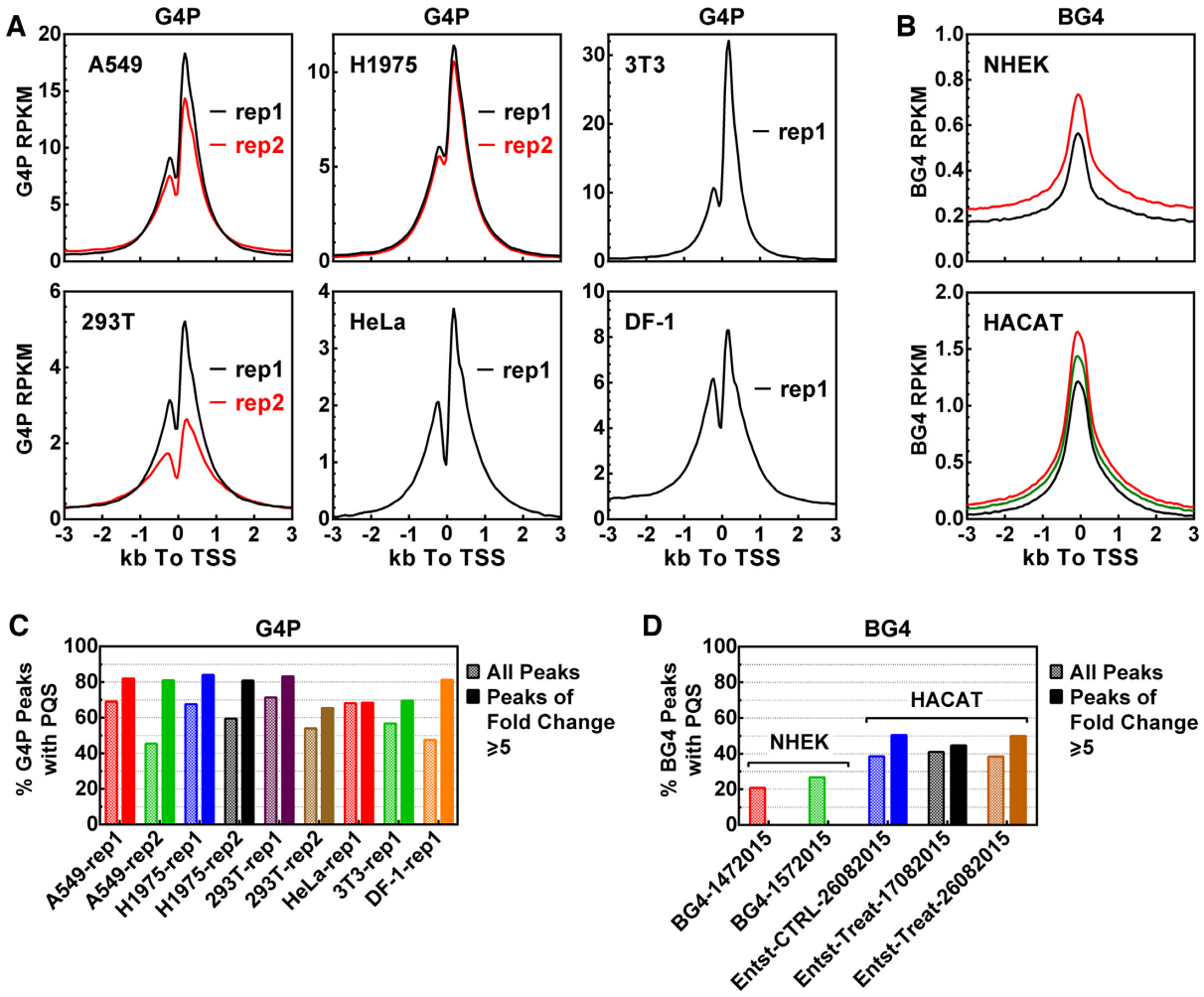


Figure 8. Comparison between detection of G4s in living cells using G4P and in fixed cells using BG4. Distribution of (**A**) G4P or (**B**) BG4 reads around TSSs. Percent of (**C**) G4P or (**D**) BG4 peaks that overlapped with at least one PQS by one or more nucleotides. Fastq and narrow peak files for BG4 were downloaded from GSE76688 in the GEO database (https://www.ncbi.nlm.nih.gov/geo/). Bigwig files were produced from the fastq files as were the files for G4P using the same parameters whenever applicable. Calculation of distribution and PQS overlapping were all conducted in the same way. The cell line is indicated under the X-axis or in the panel.

(37–40,45,57,72,73) by two distinctive mechanisms. At the upstream side, a G4 can form in response to the upward transmission of negative supercoiling waves (45,59) while at the downstream side a G4 can form when a PQS is approached by a transcription bubble (53). In correlation with the former mechanism, G4 formation in cells coincided closely with negative supercoiling (Figure 7). On either side, the degree of G4 formation is proportional to the activity of *in vitro* transcription (30,59). In agreement with this, G4 formation detected by G4P depended on the transcription activity of genes (Figure 3H). In particular, the R-loop formed between an RNA transcript and the non-template DNA strand slows down the unfolding of a G4 by preventing the annealing of a duplex (30). Owing to this enforcement, G4P revealed a greater formation of G4s at the downstream side of the TSSs than at the upstream side (Figure 3F) while the frequency of PQS occurrence is the opposite (Figure 3G).

Impact of G4P expression on cell metabolism and G4 formation

Introducing G4P into cells is expected to affect cellular metabolism in several ways. Firstly, G4P competes with native G4-binding proteins in binding genomic G4s. Secondly, the binding of G4P to genomic G4s slows down their unfolding if they are otherwise free. These may also occur to RNA G4s and affect related processes. Such G4-G4P interaction globally influenced the gene expression of the human cells (Supplementary Figures S11 and S12). Due to the dependence of G4 formation on transcription activity (Figure 3H), these changes in transcription activity should in return further affect the formation of G4s, altering the original landscape of G4 formation. In principle, the overall impact of G4P will depend on its binding activity relative to the endogenous interactions between G4s and the dozens of native G4-binding activities in human cells (74). Although it is a complicated situation, the overall effect may be evalu-

ated by the effect on transcription. Gene expression undergoes significant variation in native physiological processes, for example, cell cycle (75). Besides, it also exhibits intrinsic temporal fluctuations of up to >10-fold changes (76,77). Because the degree of variation in gene expression caused by G4P (Supplementary Figures S11 and S12) is generally comparable to those that can undergo in native physiological processes, we assume that the introduction of G4P might affect the magnitude of G4 formation unless a gene was completely switched on or off by the G4P.

We failed to observe induction of G4 formation by G4P in both dsDNA and ssDNA (Figure 2). The fact that G4 formed at some PQSs but not others in a cell line (Supplementary Figures S7–S9) is not explained by a mechanism of induction. The stand-biased formation of G4 suggested that transcription, not G4P, was the determinant of the formation of G4 in transcribed DNA (Figure 1E). This implication is further supported by the opposite polarity of G4 formation versus the PQS occurrence frequency across the TSSs (Figure 3F versus G). Like the original RHAU23, the G4P is highly selective in that it even does not bind G4s that are not in the correct conformation (Figure 1D, Supplementary Figure S2B), not to mention the unfolded species. It is unlikely that a G4P would turn a non-G4 species into a G4 without binding to the target. Transcription can efficiently induce G4 formation (30,37–40,45,57,59,72,73) even in a single round (59). With a high affinity to G4s, the G4P should preferentially interact with the large number of G4s that are already present rather than with the non-G4 species to induce them to form G4s with undetectable affinity. In support of this anticipation, the co-localization of G4P with the native G4-binding proteins and their correlation in magnitude (Figure 6) implied that the G4P faithfully detected G4s that could form natively in the cells, suggesting that capturing G4s already formed should be, at least, the major form of interaction.

G4P as a tool for G4 biology

As a G4 probe, the G4P possesses several key traits that outperform antibodies and native G4-binding proteins. In comparison, the G4P has a much smaller size, higher specificity, and affinity (K_d of low-nM). Most importantly, it overcomes the problem of impermeability and disulfide bonds associated with antibodies such that the G4P is readily applicable in living cells. With the removal of >90% of the amino acid residues from the original RHAU, the G4P is unlikely to interact with other proteins as the RHAU and other proteins might do, therefore, ensuring direct target recognition and specificity. While native G4-binding proteins can also locate G4 formation, they may result in false-positive detection because of potential interaction with other macromolecules. For instance, the hnRNP A2* protein can bind the RNA component of telomerase in addition to telomere DNA G4s (78). In theory, it may crosslink genomic DNA at where the RNA component of telomerase is synthesized when a cell is being fixed. On the other hand, the tiny size of the G4P helps it gains higher resolution and accessibility to G4s. The core G4P is only 6.7 kDa (spanning only two RHAU23s), being about half the size of a nanobody (12–15 kDa, the smallest antibody derivative). Even with the NLS and FLAG tags, it is merely 11 kDa, still smaller than a nanobody, not to mention the scFv (27 kDa) and regular antibody (∼150 kDa). Chemical probe has been used for detecting non-B DNA structures in living cells by ChIP-Seq (79) that covers a broader range of secondary structures including G4s. Our G4P is dedicated to G4s and expected to provide unique complementation to such existing tools.

The G4P also offers many opportunities to expand its functionality in various applications both *in vitro* and *in vivo* to study or modify the activity of G4s and related biological and medical processes. For instance, the fusion of G4P with fluorescent proteins may enable visualization of G4s and G4-related activities. Fusion with enzymes may enable modifications to protein/DNA/RNA or create desired reactions in association with G4s that would not occur natively. G4P-ChIP may be employed to identify proteins or processes that interact or associate with G4s. The 64-aa core G4P may be used as building blocks and find application in biological drugs, molecular nanodevices to confer G4 recognition capability. The core G4P can be synthesized in large quantities at low cost with a possibility of chemical modifications during or after synthesis. Functional groups or labels, such as radionuclei, quantum dots, nanoparticles, fluorescent dyes, etc., can be conjugated through the amine or carboxyl groups of the G4P. Collectively, all these beneficial attributes make the G4P a satisfactory G4 affinity agent.

DATA AVAILABILITY

Original sequencing (fastq) and processed (narrowPeak, bigwig) G4P-ChIP datasets, and PQS motifs (bigbed) have been deposited in and can be downloaded from the NCBI Gene Expression Omnibus (GEO) under accession code GSE133379.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Natural Science Foundation of China (NSFC) [21672212, 21432008, 21708042, 21602220] and China Postdoctoral Science Foundation [2019M653167]. Funding for open access charge: NSFC.
Conflict of interest statement. None declared.

REFERENCES

1. Rhodes,D. and Lipps,H.J. (2015) G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res.*, **43**, 8627–8637.
2. Neidle,S. (2017) Quadruplex nucleic acids as targets for anticancer therapeutics. *Nat. Rev. Chem.*, **1**, 0041.
3. Asamitsu,S., Obata,S., Yu,Z., Bando,T. and Sugiyama,H. (2019) Recent progress of targeted G-quadruplex-preferred ligands toward cancer therapy. *Molecules*, **24**, 429.
4. Ruggiero,E. and Richter,S.N. (2018) G-quadruplexes and G-quadruplex ligands: targets and tools in antiviral therapy. *Nucleic Acids Res.*, **46**, 3270–3283.
5. Bessi,I., Wirmer-Bartoschek,J., Dash,J. and Schwalbe,H. (2017) In: Webb,G.A. (ed). *Modern Magnetic Resonance*. Springer International Publishing, Cham, pp. 1–22.

6. Rodriguez,R., Miller,K.M., Forment,J.V., Bradshaw,C.R., Nikan,M., Britton,S., Oelschlaegel,T., Xhemalce,B., Balasubramanian,S. and Jackson,S.P. (2012) Small-molecule-induced DNA damage identifies alternative DNA structures in human genes. *Nat. Chem. Biol.*, **8**, 301–310.
7. Tseng,T.Y., Chien,C.H., Chu,J.F., Huang,W.C., Lin,M.Y., Chang,C.C. and Chang,T.C. (2013) Fluorescent probe for visualizing guanine-quadruplex DNA by fluorescence lifetime imaging microscopy. *J. Biomed. Opt.*, **18**, 101309.
8. Zhang,S., Sun,H., Wang,L., Liu,Y., Chen,H., Li,Q., Guan,A., Liu,M. and Tang,Y. (2018) Real-time monitoring of DNA G-quadruplexes in living cells with a small-molecule fluorescent probe. *Nucleic Acids Res.*, **46**, 7522–7532.
9. Moruno-Manchon,J.F., Lejault,P., Wang,Y., McCauley,B., Honarpisheh,P., Morales Scheihing,D.A., Singh,S., Dang,W., Kim,N., Urayama,A. *et al.* (2020) Small-molecule G-quadruplex stabilizers reveal a novel pathway of autophagy regulation in neurons. *Elife*, **9**, e52283.
10. Tseng,T.Y., Wang,Z.F., Chien,C.H. and Chang,T.C. (2013) In-cell optical imaging of exogenous G-quadruplex DNA by fluorogenic ligands. *Nucleic Acids Res.*, **41**, 10605–10618.
11. Zhang,S., Sun,H., Chen,H., Li,Q., Guan,A., Wang,L., Shi,Y., Xu,S., Liu,M. and Tang,Y. (2018) Direct visualization of nucleolar G-quadruplexes in live cells by using a fluorescent light-up probe. *Biochim. Biophys. Acta Gen. Subj.*, **1862**, 1101–1106.
12. Tseng,T.Y., Chu,I.T., Lin,S.J., Li,J. and Chang,T.C. (2018) Binding of small molecules to G-quadruplex DNA in cells revealed by fluorescence lifetime imaging microscopy of o-BMVC Foci. *Molecules*, **24**, 35.
13. Sun,W., Cui,J.X., Ma,L.L., Lu,Z.L., Gong,B., He,L. and Wang,R. (2018) Imaging nucleus viscosity and G-quadruplex DNA in living cells using a nucleus-targeting two-photon fluorescent probe. *Analyst*, **143**, 5799–5804.
14. Manna,S. and Srivatsan,S.G. (2018) Fluorescence-based tools to probe G-quadruplexes in cell-free and cellular environments. *RSC Adv*, **8**, 25673–25694.
15. Yang,S.Y., Amor,S., Laguerre,A., Wong,J.M.Y. and Monchaud,D. (2017) Real-time and quantitative fluorescent live-cell imaging with quadruplex-specific red-edge probe (G4-REP). *Biochim. Biophys. Acta Gen. Subj.*, **1861**, 1312–1320.
16. Lu,Y.J., Deng,Q., Hu,D.P., Wang,Z.Y., Huang,B.H., Du,Z.Y., Fang,Y.X., Wong,W.L., Zhang,K. and Chow,C.F. (2015) A molecular fluorescent dye for specific staining and imaging of RNA in live cells: a novel ligand integration from classical thiazole orange and styryl compounds. *Chem. Commun. (Camb.)*, **51**, 15241–15244.
17. Liu,L.Y., Liu,W., Wang,K.N., Zhu,B.C., Xia,X.Y., Ji,L.N. and Mao,Z.W. (2020) Quantitative detection of G-quadruplex DNA in live cells based on photon counts and complex structure discrimination. *Angew. Chem. Int. Ed. Engl.*, **59**, 9719–9726.
18. Di Antonio,M., Ponjavic,A., Radzevicius,A., Ranasinghe,R.T., Catalano,M., Zhang,X., Shen,J., Needham,L.M., Lee,S.F., Klenerman,D. *et al.* (2020) Single-molecule visualization of DNA G-quadruplex formation in live cells. *Nat. Chem.*, **12**, 832–837.
19. Zhang,L., Liu,X., Lu,S., Liu,J., Zhong,S., Wei,Y., Bing,T., Zhang,N. and Shangguan,D. (2020) Thiazole orange Styryl derivatives as fluorescent probes for G-quadruplex DNA. *ACS Appl. Biol. Mater.*, **3**, 2643–2650.
20. Gray,L.T., Vallur,A.C., Eddy,J. and Maizels,N. (2014) G quadruplexes are genomewide targets of transcriptional helicases XPB and XPD. *Nat. Chem. Biol.*, **10**, 313–318.
21. Liu,H.H., Zheng,K.W., He,Y.D., Chen,Q., Hao,Y.H. and Tan,Z. (2016) RNA G-quadruplex formation in defined sequence in living cells detected by bimolecular fluorescence complementation. *Chem. Sci.*, **7**, 4573–4581.
22. Jimeno,S., Camarillo,R., Mejias-Navarro,F., Fernandez-Avila,M.J., Soria-Bretones,I., Prados-Carvajal,R. and Huertas,P. (2018) The helicase PIF1 facilitates resection over sequences prone to forming G4 structures. *Cell Rep.*, **24**, 3262–3273.
23. Monchaud,D. (2020) In: *Annual Reports in Medicinal Chemistry*. Academic Press.
24. Biffi,G., DiAntonio,M., Tannahill,D. and Balasubramanian,S. (2014) Visualization and selective chemical targeting of RNA G-quadruplex structures in the cytoplasm of human cells. *Nat. Chem.*, **6**, 75–80.
25. Hansel-Hertsch,R., Beraldi,D., Lensing,S.V., Marsico,G., Zyner,K., Parry,A., Di Antonio,M., Pike,J., Kimura,H., Narita,M. *et al.* (2016) G-quadruplex structures mark human regulatory chromatin. *Nat. Genet.*, **48**, 1267–1272.
26. Lam,E.Y., Beraldi,D., Tannahill,D. and Balasubramanian,S. (2013) G-quadruplex structures are stable and detectable in human genomic DNA. *Nat. Commun.*, **4**, 1796.
27. Guo,J.U. and Bartel,D.P. (2016) RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. *Science*, **353**, aaf5371.
28. Neidle,S. (2016) Quadruplex nucleic acids as novel therapeutic targets. *J. Med. Chem.*, **59**, 5987–6011.
29. Tian,T., Chen,Y.-Q., Wang,S.-R. and Zhou,X. (2018) G-Quadruplex: A regulator of gene expression and its chemical targeting. *Chem*, **4**, 1314–1344.
30. Zhao,Y., Zhang,J.Y., Zhang,Z.Y., Tong,T.J., Hao,Y.H. and Tan,Z. (2017) Real-time detection reveals responsive cotranscriptional formation of persistent intramolecular DNA and intermolecular DNA:RNA hybrid G-quadruplexes stabilized by R-Loop. *Anal. Chem.*, **89**, 6036–6042.
31. 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.
32. Stocks,M. (2005) Intrabodies as drug discovery tools and therapeutics. *Curr. Opin. Chem. Biol.*, **9**, 359–365.
33. Heddi,B., Cheong,V.V., Martadinata,H. and Phan,A.T. (2015) Insights into G-quadruplex specific recognition by the DEAH-box helicase RHAU: Solution structure of a peptide-quadruplex complex. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 9608–9613.
34. Hockemeyer,D., Wang,H.Y., Kiani,S., Lai,C.S., Gao,Q., Cassady,J.P., Cost,G.J., Zhang,L., Santiago,Y., Miller,J.C. *et al.* (2011) Genetic engineering of human pluripotent cells using TALE nucleases. *Nat. Biotechnol.*, **29**, 731–734.
35. Ran,F.A., Hsu,P.D., Wright,J., Agarwala,V., Scott,D.A. and Zhang,F. (2013) Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.*, **8**, 2281–2308.
36. Cobb,R.E., Wang,Y.J. and Zhao,H.M. (2015) High-efficiency multiplex genome editing of streptomyces species using an engineered CRISPR/Cas system. *ACS Synth. Biol.*, **4**, 723–728.
37. Zheng,K.W., Wu,R.Y., He,Y.D., Xiao,S., Zhang,J.Y., Liu,J.Q., Hao,Y.H. and Tan,Z. (2014) A competitive formation of DNA:RNA hybrid G-quadruplex is responsible to the mitochondrial transcription termination at the DNA replication priming site. *Nucleic Acids Res.*, **42**, 10832–10844.
38. Xiao,S., Zhang,J.Y., Wu,J., Wu,R.Y., Xia,Y., Zheng,K.W., Hao,Y.H., Zhou,X. and Tan,Z. (2014) Formation of DNA:RNA hybrid G-quadruplexes of two G-quartet layers in transcription: expansion of the prevalence and diversity of G-quadruplexes in genomes. *Angew. Chem. Int. Ed. Engl.*, **53**, 13110–13114.
39. Zheng,K.W., Xiao,S., Liu,J.Q., Zhang,J.Y., Hao,Y.H. and Tan,Z. (2013) Co-transcriptional formation of DNA:RNA hybrid G-quadruplex and potential function as constitutional cis element for transcription control. *Nucleic Acids Res.*, **41**, 5533–5541.
40. Wu,R.Y., Zheng,K.W., Zhang,J.Y., Hao,Y.H. and Tan,Z. (2015) Formation of DNA:RNA hybrid G-quadruplex in bacterial cells and its dominance over the intramolecular DNA G-quadruplex in mediating transcription termination. *Angew. Chem. Int. Ed. Engl.*, **54**, 2447–2451.
41. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
42. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and Genome Project Data Processing, S. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
43. Ramirez,F., Ryan,D.P., Gruning,B., Bhardwaj,V., Kilpert,F., Richter,A.S., Heyne,S., Dundar,F. and Manke,T. (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.*, **44**, W160–W165.
44. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.

45. 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.
46. Lattmann,S., Giri,B., Vaughn,J.P., Akman,S.A. and Nagamine,Y. (2010) Role of the amino terminal RHAU-specific motif in the recognition and resolution of guanine quadruplex-RNA by the DEAH-box RNA helicase RHAU. *Nucleic Acids Res.*, **38**, 6219–6233.
47. Guedin,A., Gros,J., Alberti,P. and Mergny,J.L. (2010) How long is too long? Effects of loop size on G-quadruplex stability. *Nucleic Acids Res.*, **38**, 7858–7868.
48. Li,X.M., Zheng,K.W., Zhang,J.Y., Liu,H.H., He,Y.D., Yuan,B.F., Hao,Y.H. and Tan,Z. (2015) Guanine-vacancy-bearing G-quadruplexes responsive to guanine derivatives. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 14581–14586.
49. Mukundan,V.T. and Phan,A.T. (2013) Bulges in G-quadruplexes: broadening the definition of G-quadruplex-forming sequences. *J. Am. Chem. Soc.*, **135**, 5017–5028.
50. Zhou,B., Liu,C., Geng,Y. and Zhu,G. (2015) Topology of a G-quadruplex DNA formed by C9orf72 hexanucleotide repeats associated with ALS and FTD. *Sci. Rep.*, **5**, 16673.
51. Biffi,G., Tannahill,D., McCafferty,J. and Balasubramanian,S. (2013) Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat. Chem.*, **5**, 182–186.
52. Su,Z., Zhang,Y., Gendron,T.F., Bauer,P.O., Chew,J., Yang,W.Y., Fostvedt,E., Jansen-West,K., Belzil,V.V., Desaro,P. *et al.* (2014) Discovery of a biomarker and lead small molecules to target r(GGGGCC)-associated defects in c9FTD/ALS. *Neuron*, **83**, 1043–1050.
53. Liu,J.Q., Xiao,S., Hao,Y.H. and Tan,Z. (2015) Strand-biased formation of G-quadruplexes in DNA duplexes transcribed with T7 RNA polymerase. *Angew. Chem. Int. Ed. Engl.*, **54**, 8992–8996.
54. Todd,A.K., Johnston,M. and Neidle,S. (2005) Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res.*, **33**, 2901–2907.
55. Quinlan,A.R. (2014) BEDTools: The swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinformatics*, **47**, 11 12 11–34.
56. Rozowsky,J., Euskirchen,G., Auerbach,R.K., Zhang,Z.D., Gibson,T., Bjornson,R., Carriero,N., Snyder,M. and Gerstein,M.B. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.*, **27**, 66–75.
57. Zhang,J.Y., Zheng,K.W., Xiao,S., Hao,Y.H. and Tan,Z. (2014) Mechanism and manipulation of DNA:RNA hybrid G-quadruplex formation in transcription of G-rich DNA. *J. Am. Chem. Soc.*, **136**, 1381–1390.
58. Belotserkovskii,B.P., Tornaletti,S., D'Souza,A.D. and Hanawalt,P.C. (2018) R-loop generation during transcription: Formation, processing and cellular outcomes. *DNA Repair (Amst.)*, **71**, 69–81.
59. Xia,Y., Zheng,K.W., He,Y.D., Liu,H.H., Wen,C.J., Hao,Y.H. and Tan,Z. (2018) Transmission of dynamic supercoiling in linear and multi-way branched DNAs and its regulation revealed by a fluorescent G-quadruplex torsion sensor. *Nucleic Acids Res.*, **46**, 7418–7424.
60. Yagi,R., Miyazaki,T. and Oyoshi,T. (2018) G-quadruplex binding ability of TLS/FUS depends on the beta-spiral structure of the RGG domain. *Nucleic Acids Res.*, **46**, 5894–5901.
61. Ishiguro,A., Kimura,N., Watanabe,Y., Watanabe,S. and Ishihama,A. (2016) TDP-43 binds and transports G-quadruplex-containing mRNAs into neurites for local translation. *Genes Cells*, **21**, 466–481.
62. von Hacht,A., Seifert,O., Menger,M., Schutze,T., Arora,A., Konthur,Z., Neubauer,P., Wagner,A., Weise,C. and Kurreck,J. (2014) Identification and characterization of RNA guanine-quadruplex binding proteins. *Nucleic Acids Res.*, **42**, 6630–6644.
63. Choi,H.S., Hwang,C.K., Song,K.Y., Law,P.Y., Wei,L.N. and Loh,H.H. (2009) Poly(C)-binding proteins as transcriptional regulators of gene expression. *Biochem. Biophys. Res. Commun.*, **380**, 431–436.
64. Tomonaga,T. and Levens,D. (1996) Activating transcription from single stranded DNA. *Proc. Natl. Acad. Sci. U.S.A.*, **93**, 5830–5835.
65. Kouzine,F., Gupta,A., Baranello,L., Wojtowicz,D., Ben-Aissa,K., Liu,J., Przytycka,T.M. and Levens,D. (2013) Transcription-dependent dynamic supercoiling is a short-range genomic force. *Nat. Struct. Mol. Biol.*, **20**, 396–403.
66. Kouzine,F., Liu,J., Sanford,S., Chung,H.J. and Levens,D. (2004) The dynamic response of upstream DNA to transcription-generated torsional stress. *Nat. Struct. Mol. Biol.*, **11**, 1092–1100.
67. Xiao,S., Zhang,J.Y., Zheng,K.W., Hao,Y.H. and Tan,Z. (2013) Bioinformatic analysis reveals an evolutional selection for DNA:RNA hybrid G-quadruplex structures as putative transcription regulatory elements in warm-blooded animals. *Nucleic Acids Res.*, **41**, 10379–10390.
68. Maity,A., Winnerdy,F.R., Chang,W.D., Chen,G. and Phan,A.T. (2020) Intra-locked G-quadruplex structures formed by irregular DNA G-rich motifs. *Nucleic Acids Res.*, **48**, 3315–3327.
69. Truong,T.H.A., Winnerdy,F.R. and Phan,A.T. (2019) An Unprecedented Knot-like G-Quadruplex Peripheral Motif. *Angew. Chem. Int. Ed. Engl.*, **58**, 13834–13839.
70. Piazza,A., Cui,X., Adrian,M., Samazan,F., Heddi,B., Phan,A.T. and Nicolas,A.G. (2017) Non-canonical G-quadruplexes cause the hCEB1 minisatellite instability in Saccharomyces cerevisiae. *Elife*, **6**, e26884.
71. Lightfoot,H.L., Hagen,T., Tatum,N.J. and Hall,J. (2019) The diverse structural landscape of quadruplexes. *FEBS Lett.*, **593**, 2083–2102.
72. Zheng,K.W., He,Y.D., Liu,H.H., Li,X.M., Hao,Y.H. and Tan,Z. (2017) Superhelicity constrains a localized and R-loop-dependent formation of G-quadruplexes at the upstream region of transcription. *ACS Chem. Biol.*, **12**, 2609–2618.
73. Zhang,J.Y., Xia,Y., Hao,Y.H. and Tan,Z. (2020) DNA:RNA hybrid G-quadruplex formation upstream of transcription start site. *Sci. Rep.*, **10**, 7429.
74. Brazda,V., Haronikova,L., Liao,J.C. and Fojta,M. (2014) DNA and RNA quadruplex-binding proteins. *Int. J. Mol. Sci.*, **15**, 17493–17517.
75. Liu,Y., Chen,S., Wang,S., Soares,F., Fischer,M., Meng,F., Du,Z., Lin,C., Meyer,C., DeCaprio,J.A. *et al.* (2017) Transcriptional landscape of the human cell cycle. *Proc. Natl. Acad. Sci. USA*, **114**, 3473–3478.
76. Li,G.W. and Xie,X.S. (2011) Central dogma at the single-molecule level in living cells. *Nature*, **475**, 308–315.
77. Raj,A., Peskin,C.S., Tranchina,D., Vargas,D.Y. and Tyagi,S. (2006) Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.*, **4**, e309.
78. Wang,F., Tang,M.L., Zeng,Z.X., Wu,R.Y., Xue,Y., Hao,Y.H., Pang,D.W., Zhao,Y. and Tan,Z. (2012) Telomere- and telomerase-interacting protein that unfolds telomere G-quadruplex and promotes telomere extension in mammalian cells. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 20413–20418.
79. Kouzine,F., Wojtowicz,D., Baranello,L., Yamane,A., Nelson,S., Resch,W., Kieffer-Kwon,K.R., Benham,C.J., Casellas,R., Przytycka,T.M. *et al.* (2017) Permanganate/S1 nuclease footprinting reveals non-B DNA structures with regulatory potential across a mammalian genome. *Cell Syst.*, **4**, 344–356.