Structure and possible function of a G-quadruplex in the long terminal repeat of the proviral HIV-1 genome

Beatrice De Nicola[1,2,†], Christopher J. Lech[1,†], Brahim Heddi[1,†], Sagar Regmi[1], Ilaria Frasson[2], Rosalba Perrone[2], Sara N. Richter[2,*] and Anh Tuân Phan[1,*]

[1]School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore and [2]Department of Molecular Medicine, University of Padua, Italy

ABSTRACT

The long terminal repeat (LTR) of the proviral human immunodeficiency virus (HIV)-1 genome is integral to virus transcription and host cell infection. The guanine-rich U3 region within the LTR promoter, previously shown to form G-quadruplex structures, represents an attractive target to inhibit HIV transcription and replication. In this work, we report the structure of a biologically relevant G-quadruplex within the LTR promoter region of HIV-1. The guanine-rich sequence designated *LTR-IV* forms a well-defined structure in physiological cationic solution. The nuclear magnetic resonance (NMR) structure of this sequence reveals a parallel-stranded G-quadruplex containing a single-nucleotide thymine bulge, which participates in a conserved stacking interaction with a neighboring single-nucleotide adenine loop. Transcription analysis in a HIV-1 replication competent cell indicates that the *LTR-IV* region may act as a modulator of G-quadruplex formation in the LTR promoter. Consequently, the *LTR-IV* G-quadruplex structure presented within this work could represent a valuable target for the design of HIV therapeutics.

INTRODUCTION

G-quadruplexes are nucleic acid secondary structures that may form in G-rich sequences under physiological conditions (1–3). In contrast to duplex DNA formed by Watson-Crick base-pairing, the building blocks of G-quadruplexes are stacked guanine tetrads (G-tetrads) assembled by Hoogsteen-type base-pairing. The presence of coordinating cations is important to G-quadruplex formation and stability (4–6). G-quadruplex structures are highly polymorphic, both in terms of strand stoichiometry (forming inter- and intramolecular structures) and strand orientation/topology (7,8). G-quadruplex-forming motifs have been found in telomeres, G-rich micro- and mini-satellites and near oncogene promoters (8–17). Human G-quadruplex DNA motifs have been reported to be associated with recombination prone regions (18) and to show mutational patterns that preserved the potential to form G-quadruplex DNA structures (14). Expansion of G-quadruplex-forming motifs has been associated with relevant human neurological disorders (19–25). The identification of G-quadruplex binding proteins (26,27) and their visualization in cells with antibody-based technology (28,29) have also provided convincing evidence of the existence of G-quadruplexes *in vivo*.

Recently, the presence of G-quadruplexes in viruses and their involvement in viral key steps have also been reported (30). G-quadruplexes have been implicated in pathogenic mechanisms of the Epstein-Barr virus (31,32), SARS coronavirus (33), herpes simplex virus 1 (34) and the human papillomavirus (35). We and other groups have identified functionally significant G-quadruplexes in the Nef coding region (36) and the unique long terminal repeat (LTR) promoter (37–39) of the human immunodeficiency virus (HIV), the etiologic agent of the acquired immune deficiency syndrome (AIDS), a major worldwide epidemic.

HIV establishes a persistent infection in human hosts, with the depletion of CD4+ lymphocytes, the major target cells of viral infection *in vivo*, eventually resulting in defective cellular immunity, and thus leading to full-blown AIDS. The Joint United Nations Programme on HIV/AIDS (UN-AIDS) estimates that there were 35.3 million people living with HIV at the end of 2012 compared with 26.2 million in 1999 – a 35% increase. Although important progress has been achieved in preventing new HIV infections and prolonging HIV patients' life with antiretroviral therapy, the infection cannot be eradicated and therefore AIDS-related illnesses are projected to continue as a significant worldwide cause of premature mortality in the coming decades if a decisive cure is not found (40).

HIV is composed of two copies of positive single-stranded RNA that codes for the nine virus genes. The genomic RNA is reverse transcribed into a linear double-

*To whom correspondence should be addressed. Tel: +65 6514 1915; Fax: +65 6795 7981; Email: phantuan@ntu.edu.sg
Correspondence may also be addressed to Sara N. Richter. Tel: +39 049 827 2346; Fax: +39 049 827 2355; Email: sara.richter@unipd.it
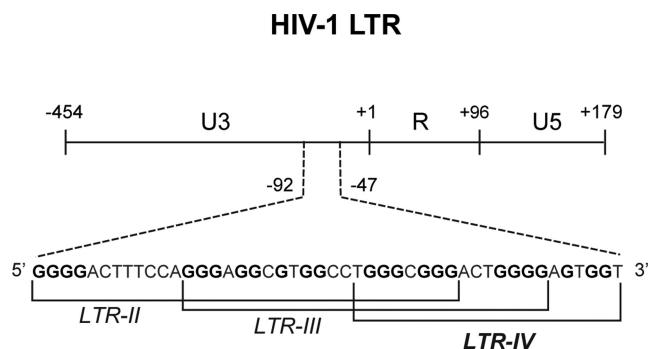†These authors contributed equally to the paper as first authors.

HIV-1 LTR

Figure 1. Sequence of the G-rich region embedded in the U3 moiety of the LTR promoter in the proviral HIV-1 genome. The *LTR-IV* sequence is the focus of this work. The *LTR-III* and *LTR-II* regions are also shown.

stranded DNA molecule that is next integrated into the human chromosome by the viral enzyme integrase. The stably integrated HIV-1 provirus, a 10 kb-long DNA molecule, serves as a template for the transcription of viral messengers and genomic RNA by the cellular Pol II polymerase (41). Viral transcription is triggered by the interaction of cellular transcription factors with the U3 region of the LTR promoter (Figure 1) (42). The LTR promoter contains multiple G-rich sequences that have been shown to adopt G-quadruplex structures. The formation of G-quadruplexes in the LTR results in decreased viral transcription (37), with an effect that is enhanced by the presence of the cellular protein nucleolin (43) and G-quadruplex ligands (44).

Interestingly, within the G-rich region of the LTR promoter one of the LTR G-quadruplexes, termed *LTR-IV*, does not readily form in physiological ionic conditions. However, *LTR-IV* can be induced by a G-quadruplex stabilizing ligand (37), making it a good G-quadruplex targets for antiviral therapy.

Several G-quadruplex ligands have been developed against cellular G-quadruplexes implicated in tumor pathogenesis; at least two molecules have also been tested and proved effective as antiviral agents against HIV (36,44). Recently, a few compounds selective for HIV-1 LTR G-quadruplexes have also been reported (45). These studies have shown the potential of developing antiviral molecules with a G-quadruplex-mediated mechanism of action. To increase the selectivity of G-quadruplex binding ligands toward a viral target over other G-quadruplexes that may form within the cell, the high-resolution structure of the target G-quadruplexes will allow the rational design and virtual screening of molecules with selective binding. Considering the value of such target elucidation, we report on the nuclear magnetic resonance (NMR) solution structure of a G-quadruplex formed by the 22-nt G-rich *LTR-IV* sequence (5′-CTG$_3$CG$_3$ACTG$_4$AGTG$_2$T-3′) from the promoter region of the HIV-1 LTR.

MATERIALS AND METHODS

DNA sample preparation

Oligonucleotides used in the *Taq* polymerase stop assay and in the construction/mutagenesis of the plasmids for the luciferase reporter assay were purchased from Sigma Aldrich (Milan, Italy). All other DNA oligonucleotides were purchased from Integrated DNA Technologies (Coralville, IA, USA) or chemically synthesized on an ABI 394 DNA/RNA synthesizer. Synthesized oligonucleotides were purified and dialyzed successively against 20 mM KCl solution and water before being lyophilized. DNA concentration was expressed in strand molarity using a nearest-neighbor approximation for the absorption coefficients of the unfolded species (46).

Gel electrophoresis

The molecular size of G-quadruplexes was visualized by non-denaturing polyacrylamide gel electrophoresis (PAGE). Samples were incubated in 10 mM potassium phosphate buffer (pH 7) before loaded on 20% polyacrylamide gels with 40% (v/v) sucrose added before loading. Gels were run with 10 mM potassium phosphate buffer (pH 7) at room temperature for 90 min at 120 volts. DNA migration was imaged using UV-shadowing. For *Taq* polymerase stop assay, the DNA extension products were separated on a 15% denaturing gel (7 M urea), run in TBE buffer for 3 h at 80 W, and finally visualized by phosphorimaging (Typhoon FLA 9000).

Circular dichroism

Circular dichroism (CD) spectra were recorded on a JASCO-815 spectropolarimeter using 1-cm path length quartz cuvettes. DNA samples (strand concentration, 5 μM) were dissolved in a buffer containing 70 mM potassium chloride and 20 mM potassium phosphate (pH 7) in a reaction volume of 500 μL. For each experiment, an average of three scans was taken, the spectrum of the buffer was subtracted, and the data were zero-corrected at 320 nm.

Thermal denaturing

The thermal denaturing of the *LTR-IV* G-quadruplex was monitored by UV absorption on a JASCO V-650 UV-Vis spectrophotometer. DNA concentration was 5 μM. Samples contained 70 mM potassium chloride and 20 mM potassium phosphate, pH 7. Absorbance was monitored at 295 nm and zero-corrected at 320 nm, over the temperature range from 20 to 90°C. Melting curves were baseline-corrected to compute the fraction of folded G-quadruplex over varying temperature. The melting temperature was determined as the temperature at which half the population was in the folded state.

NMR spectroscopy

NMR experiments were performed on 600 and 700 MHz Bruker spectrometers at 25°C. DNA samples (strand concentration, 0.2−1.5 mM) were dissolved in a buffer containing 70 mM potassium chloride and 20 mM potassium phosphate (pH 7). Resonances for guanine residues were assigned unambiguously by using site-specific low-enrichment ^{15}N labeling (47), site-specific ^{2}H labeling (48) and through-bond correlations at natural abundance (49). Spectral assignments were completed by NOESY, TOCSY, (^{13}C-^{1}H)-HMBC and (^{13}C-^{1}H)-HSQC as previously described (50). Inter-proton distances were deduced from

NOESY experiments at various mixing times. All spectral analyses were performed using the SPARKY program (51).

Structure calculation

Structures of the *LTR-IV* G-quadruplex were first calculated by performing distance geometry simulated annealing using the XPLOR-NIH program (52). The ten lowest-energy structures were then used for further refinement via molecular dynamics (MD) simulations using the AMBER 10 program (53). Prior to MD simulation, K^+ ions were introduced around the structure computed in XPLOR, including internal placement between G-tetrads of the G-quadruplex. Structures were then solvated before undergoing a series of constrained minimization and simulation steps as previously described (54). Finally, structures were refined over 1 ns of restrained MD simulations including hydrogen-bond restraints and inter-proton distance restraints. Inter-proton distances were deduced from NOESY experiments performed in H_2O (mixing time, 300 ms) and ^2H_2O (mixing times, 100 and 300 ms). The glycosidic dihedral restraints were based on intra-residue NOEs of H1′-H6/8 cross-peak intensities observed at the mixing time of 100 ms.

Data deposition

The coordinates of the *LTR-IV* G-quadruplex have been deposited in the Protein Data Bank under accession code 2N4Y.

Plasmid construction

The wild-type (*WT*) HIV-1 LTR region (corresponding to segment −381/+83 in the HIV-1 genome) was inserted into the promoterless luciferase reporter vector pGL4.10-Luc2 (Promega Italia, Italy), as previously reported (37). The *G18T* mutant pGL4.10-Luc2/LTR vector was generated using QuikChange mutagenesis kit (Stratagene/Agilent Technologies) with the following primers: PrMut *G18T* (i) 5′-CCTGGGCGGGAC TGGGGA**T**TGGCGAGCCCTCAGATCC-3′ and PrMut *G18T* (ii) 5′-GGATCTGAGGGCTCGCCAA**T**CCCCA GTCCCGCCCAGG-3′, where the mutated base is shown in boldface and underlined. Both the *WT* and *G18T* LTR sequences were confirmed by sequencing. Plasmid size was confirmed by Hind III linearization on 1% agarose gel.

Reporter assay

Luciferase activity of the *WT* and *G18T* mutant LTRs was assessed in human embryonic kidney 293T (HEK293T) cells seeded in 12-well plates (2×10^5 cells/well). Cells were transfected 24 h post-seeding with pGL4.10-LTR-*WT* or pGL4.10-LTR-*G18T* (100 ng/well) using TransIT-293 transfection reagent (Mirus Bio LLC, Madison, WI, USA), according to the manufacturer's protocol. After 30 min, cells were treated with increasing concentrations (1–4 μM) of BRACO-19 (ENDOTHERM, Saarbruecken, Germany) for 24 h. Luciferase activity was measured using the britelite plus Reporter Gene Assay System (PerkinElmer Inc., Milan, Italy) in a Victor X2 multilabel plate reader (PerkinElmer Inc., Milan, Italy). Cells

were lysed in lysis buffer (1X PBS, 1% TRITON X-100) and protein concentration was determined by bicinchoninic acid (BCA) assay (Thermo Scientific Pierce, Monza, Italy). Luciferase signals were subsequently normalized to total protein content, according to the manufacturer's protocol (http://ita.promega.com/~/pdf/resources/pubhub/cellnotes/normalizing-genetic-reporter-assays/). All experiments were performed twice and in duplicate.

HEK 293T cells sustain all viral steps with the exception of virion attachment and entry as they lack cell receptors recognized by HIV-1. They can be transfected with the HIV-1 proviral genome to produce fully infectious viral particles, indicating that their cytoplasmic/nuclear protein makeup is capable of sustaining viral transcription and replication (55).

Taq polymerase stop assay

Taq polymerase stop assay was carried out as previously described (37). Briefly, the 5′-end labeled primer (5′-GGCAAAAAGCAGCTGCTTATATGCAG-3′) was annealed to the templates (LTR-II+III+IV *WT* Taq 5′-TTTTTGGGGACTTTCCAGGGAGGCGTGGCC TGGGCGGGACTGGGGAGTGGTTTTTCTGCATA TAAGCAGCTGCTTTTTGCC-3′ or LTR-II+III+IV *G18T* Taq 5′-TTTTTGGGGACTTTCCAGGGAGGC GTGGCCTGGGCGGGACTGGGGA**T**TGGTTTTTCT GCATATAAGCAGCTGCTTTTTGCC-3′) in lithium cacodylate buffer in the presence/absence of 100 mM KCl by heating to 95°C for 5 min and gradually cooling to room temperature. Where specified, samples were incubated with BRACO-19 (62.5–250 nM). Primer extension was conducted with 2 U of AmpliTaq Gold DNA polymerase (Applied Biosystem, Carlsbad, California, USA) at 37°C for 30 min. The marker sample was obtained by treating the elongation *WT* product with the sequencing protocol of Maxam *et al.* (56). Reactions were stopped by ethanol precipitation, primer extension products were separated on a denaturing gel and visualized by phosphorimaging (Typhoon FLA 9000). The intensities of stop bands were quantified using ImageQuant TL software (GE Healthcare Europe, Milan, Italy).

Analysis of G18 conservation

Conservation of the G18 base involved in the *LTR-IV* G-quadruplex formation was evaluated by aligning 953 LTR U3 sequences of different HIV-1 subtypes including 24 subtypes A, 485 subtypes B and 119 subtypes C from the HIV Sequence Database (http://www.hiv.lanl.gov/) using Jalview (http://www.jalview.org/). The analyzed sequences belong to a broad spectrum of HIV-1 subtypes, both non-recombinant subtypes of Group M and circulating recombinant forms.

RESULTS

Formation of a monomeric parallel G-quadruplex by the *LTR-IV* sequence

The NMR spectrum of the *LTR-IV* sequence displays 12 well-resolved peaks in the imino proton region (Figure
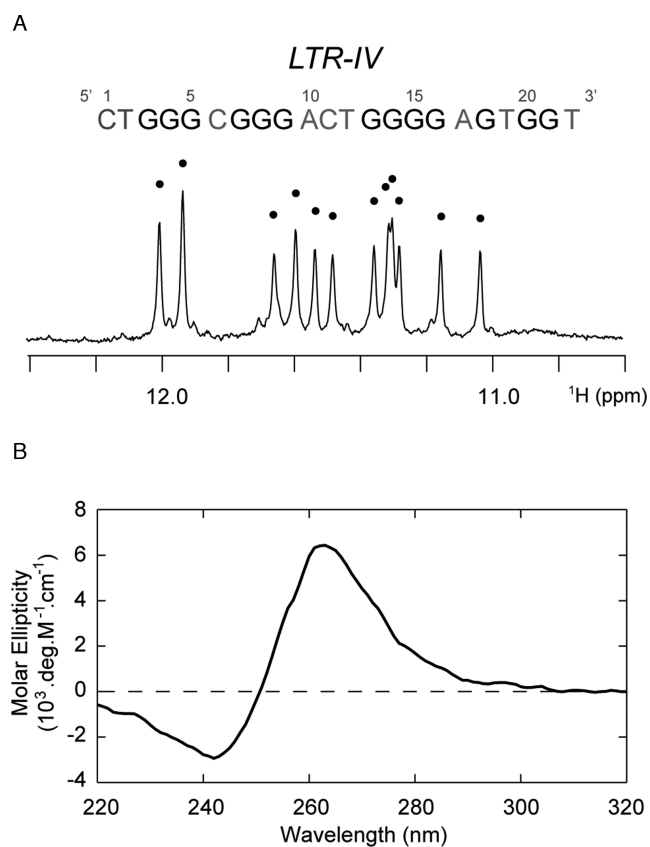
A

B

Figure 2. Biophysical characterization of the *LTR-IV* sequence. (**A**) NMR imino proton spectrum and (**B**) CD spectrum of *LTR-IV* suggest the formation of a major parallel-stranded G-quadruplex conformation.

2A), indicating the formation of a three-layer G-quadruplex structure. The CD spectrum of this sequence, showing a negative peak at 240 nm and a positive peak at 260 nm (Figure 2B), is indicative of a parallel G-quadruplex folding topology. Non-denaturing PAGE experiments show two separate species (Supplementary Figure S1): a major form that migrates at a rate consistent with a monomeric G-quadruplex reference and a small population of higher-ordered structure. Solvent exchange with ^2H_2O reveals slow exchange rates for four imino protons (Supplementary Figure S2) – corresponding to a single central G-tetrad layer. A single transition was observed in the thermal denaturing curve of *LTR-IV* (Supplementary Figure S3) showing a melting temperature of 50.5°C in ~100 mM K$^+$. Collectively, these biophysical data suggest that *LTR-IV* predominantly forms a major G-quadruplex species that adopts a monomeric parallel-stranded conformation.

LTR-IV adopts a parallel G-quadruplex containing a T-bulge

We set out to elucidate the G-quadruplex structure of the *LTR-IV* sequence. Guanine imino (H1) protons were unambiguously assigned for all guanines by carrying out ^{15}N-filtered experiments of sequences containing site-specific 4% ^{15}N enrichment (Figure 3A and Supplementary Figure S4). Unambiguous assignment of guanine H8 protons was made by site-specific ^{2}H labeling experiments (Supplementary

Figure S4). Assignment of selected aromatic and methyl protons of thymine residues was made by through-bond correlation experiments, using site-specific 4% ^{13}C,^{15}N labeled samples (Figure 3B and Supplementary Figure S5). Based on these assignments, the H8/H6-H1′ sequence connectivity could be traced allowing for assignment of cross-peaks within NOESY spectra (Figure 3D).

Examining the cyclical through-space guanine H1-H8 cross peaks in NOESY spectra allowed the identification of guanine bases that make up the same G-tetrad (Figure 3C). H1-H8 cross peaks demonstrate the formation of three guanine tetrads, G3·G7·G14·G18, G4·G8·G15·G20 and G5·G9·G16·G21, indicating a parallel-stranded G-quadruplex fold is adopted by *LTR-IV* (Figure 3E). The location of the G4·G8·G15·G20 tetrad within the center of the structure is supported by the enhanced protection of the imino protons of guanines in this tetrad during solvent exchange experiments (Supplementary Figure S2). Guanine bases are shown to adopt an *anti* conformation, consistent with the intensity of their intra-residue H8-H1′ cross-peaks observed in NOESY spectra.

The resulting folding topology indicates three continuous tracts of guanines, participating in G-tetrad formation, G3–G5, G7–G9 and G14–G16, in addition to a fragmented tract consisting of G18, G20 and G21. This fragmented tract results in a bulge at the T19 position. These G-tracts are connected by loops including a short 1-nt propeller loop formed by C6, a 4-nt propeller loop formed by the A10-C11-T12-G13 sequence and another short 1-nt propeller loop formed by A17. The flanking sequence C1-T2 is situated on the 5′-end of the molecule, while the T22 base is found on the 3′-end.

Solution structure of the *LTR-IV* G-quadruplex

Structural calculation of the *LTR-IV* G-quadruplex was carried out based on restraints derived from NMR experiments (Table 1). The ten lowest-energy structures computed were well converged (Figure 4A and Table 1).

The solution structure of *LTR-IV* shows interesting structural motifs (Figure 4B). (i) The T19 bulge readily stacks with A17 of the nearby 1-nt propeller loop. This stacking interaction is supported by multiple NOEs between the sugar and aromatic protons of these residues (Figure 3D). (ii) The 4-nt propeller loop is notably unstructured or highly dynamic, although this could potentially be due to lack of available constraints between the residues in this loop. (iii) The flanking C1 and T2 bases at the 5′-end of the molecule are located on top of the G3·G7·G14·G18 tetrad, partially capping the exposed tetrad. (iv) Similarly, T22 at the 3′-end caps the bottom G5·G9·G16·G21 tetrad. Collectively, these features give rise to a unique G-quadruplex structure containing a previously unreported loop-bulge stacking motif.

Effect of mutations on the *LTR-IV* structure

We investigated how mutations affect the structure of the *LTR-IV* sequence. Specifically, we examined the role of the T19 bulge and the ability of the G-quadruplex to form using other combinations of guanine bases. We synthesized a series of sequences containing base substitutions and/or base
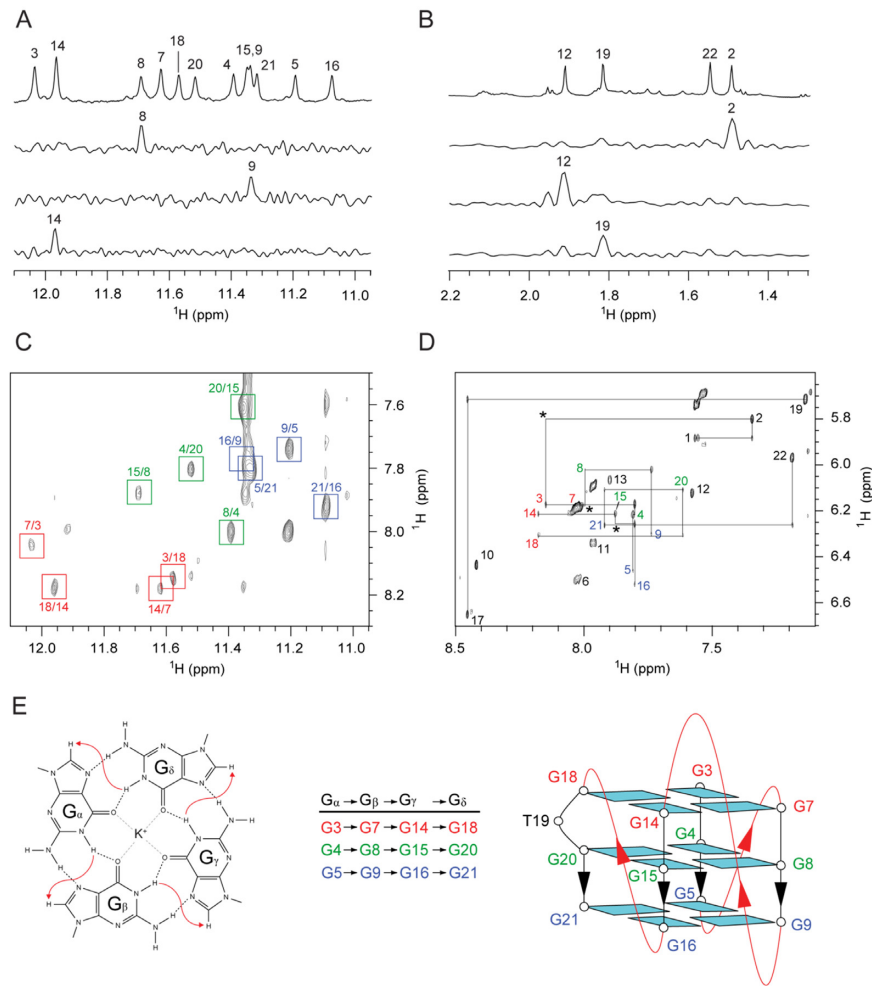
Figure 3. NMR determination of the *LTR-IV* folding topology. (**A**) Reference imino proton spectrum of *LTR-IV* with examples of ^{15}N-filtered spectra of site-specific 4% ^{15}N-enriched samples. (**B**) Reference methyl proton spectrum with examples of ^{13}C-filtered spectra of sequences containing site-specific ^{13}C-enriched thymine substitutions. (**C**) Guanine H1-H8 cyclical connectivities in NOESY spectra. Cross-peaks are colored based on G-tetrad arrangement. (**D**) Tracing the H8/6-H1' NOE sequential connectivity in NOESY spectrum (mixing time, 300 ms). Intraresidue H8/6-H1' cross-peaks are labeled with residue numbers and color coded based on G-tetrad arrangement. Missing cross-peaks are labeled with asterisks. (**E**) G-tetrad schematic highlighting the close proximity of H8-H1 protons used to determine guanine participation in G-tetrad formation. Folding topology of the *LTR-IV* sequence shows a parallel three-layered G-quadruplex conformation containing a T19 bulge.

deletions (Table 2) and examined resulting G-quadruplex formation in physiological K$^+$ conditions. The sequence *G13T* containing a G-to-T substitution at position 13 confirms that G13 does not play an important role in the G-tetrad formation. The deletion of T19 in the *ΔT19* sequence tests the formation of G-quadruplex in the absence of a T19 bulge, creating a continuous GGG tract in its place. The sequence *G13TΔT19* addresses these two changes concurrently. We also checked the ability of the *LTR-IV* sequence to utilize other guanine bases in the formation of a G-quadruplex by substituting G16 and G18 with a thymine in the *G16T* and *G18T* sequences respectively.

Mutated sequences were analyzed using NMR to ascertain their ability to form G-quadruplex and to gauge the folding similarity based on their imino proton chemical shift pattern (Figure 5). The modification of G13 to T leads to small changes in the imino proton spectra of *G13T* but maintains the general chemical shift pattern, suggesting

this guanine base is not critical to G-quadruplex formation. Similarly, the deletion of T19 in *ΔT19* results in a predominant G-quadruplex conformation with a similar imino proton chemical shift pattern to that of the unmodified *LTR-IV* sequence, although the emergence of a small minor population is visible. This suggests the T19 bulge is also not a key factor in the formation of a parallel G-quadruplex. The introduction of both of these modifications is still tolerated in *G13TΔT19*, although a slightly elevated level of a minor population is observed. Alternatively, substitution of G16 or G18 with thymine disrupted the folding topology: multiple conformations are adopted in the *G16T* sequence, while a different conformer is observed in the *G18T* sequence, displaying broad imino proton peaks. In addition to changes in NMR spectra, the mutated sequences *G16T* and *G18T* demonstrate markedly different CD spectra as compared to that of the unmodified *LTR-IV* sequence (Supplementary Figure S6) suggesting these mutations are highly disrup-
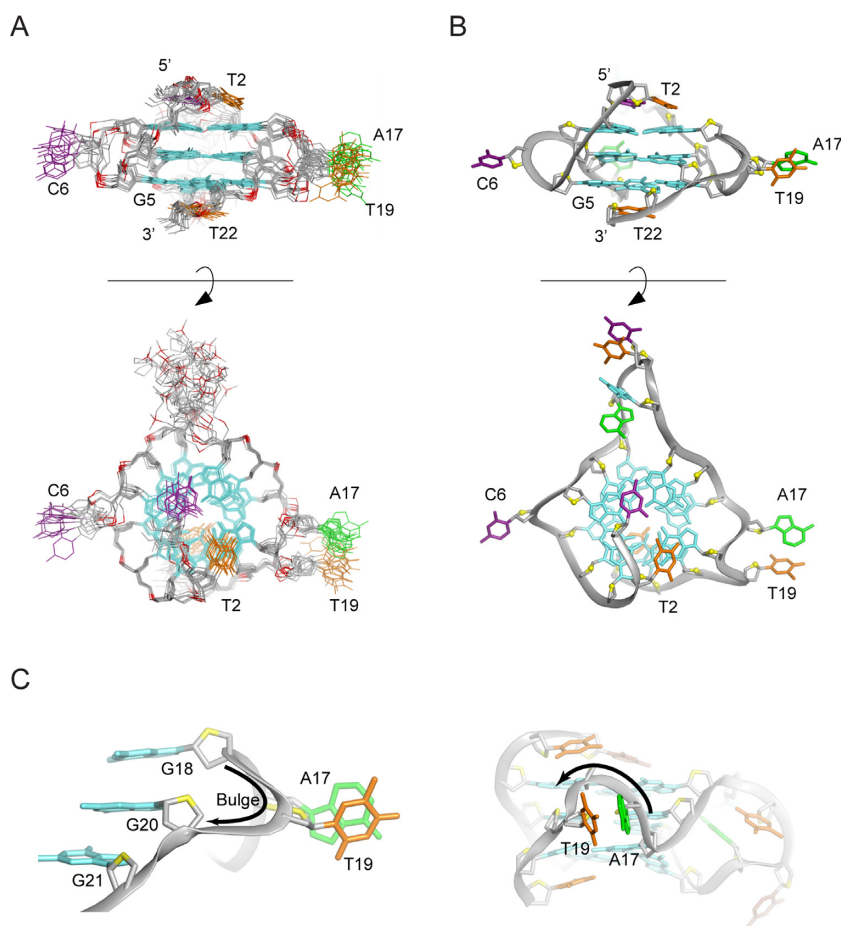
Figure 4. NMR solution structure of the *LTR-IV* G-quadruplex. (**A**) Alignment of the ten lowest-energy structures computed using XPLOR based on restraints from NMR experiments. Nucleobases of the 4-nt loop from positions 10–13 are omitted for clarity. (**B**) Ribbon view of a representative structure. (**C**) Close-up view of the T19 bulge and neighboring A17 propeller loop. Backbone are colored in grey. O4′ atoms of the sugar moiety are represented as yellow. Bases are colored as follows: guanine (cyan), adenine (green), thymine (orange) and cytosine (purple).

Table 1. Statistics of the computed structures of the *LTR-IV* G-quadruplex

NMR restraints		
Distance restraints	^2H_2O	H_2O
intra-residue	402	4
sequential (i, i + 1)	93	12
long-range (i, ≥i + 2)	18	52
Other restraints		
hydrogen bond	48	
dihedral angle	12	
Structure statistics		
NOE violations		
number (>0.2 Å)	0.100 ± 0.300	
maximum violation (Å)	0.163 ± 0.061	
R.M.S.D. of violations (Å)	0.013 ± 0.002	
Deviations from ideal covalent geometry		
bond lengths (Å)	0.003 ± 0.000	
bond angles (°)	0.714 ± 0.025	
improper dihedrals (°)	0.370 ± 0.014	
Pair-wise all heavy atom R.M.S.D. values (Å)		
all heavy atoms of G-tetrad core	0.949 ± 0.174	
all heavy atoms excluding A10, T11, C12 and G13	1.398 ± 0.192	

tive to the native conformer. Collectively, these data demonstrated the T19-bulge is not mandatory for the formation of a parallel G-quadruplex from the *LTR-IV* region of the HIV-1 LTR. In contrast, the guanines in the G14-G16 and G18-G21 stretches are required.

LTR-IV G-quadruplex formation correlates with HIV-1 LTR promoter activity

To assess the physiological role of the *LTR-IV* G-quadruplex, a G18-to-T (*G18T*) mutation was introduced in the full-length HIV-1 LTR promoter fused to a luciferase reporter gene. As the G18T mutation was shown to abolish the *LTR-IV* G-quadruplex (Figure 5), we selected this sequence as a negative structural control. Both the wild-type (*WT*) and mutant *G18T* LTR-Luc plasmids were transfected into HEK 293T cells and the luciferase activity was measured. The promoter activity of the mutant *G18T* LTR was significantly reduced to 23% of the activity of the *WT* LTR (Figure 6A). Transfected cells were next treated with BRACO-19, a G-quadruplex ligand shown to bind and stabilize *LTR-III* and *LTR-IV* G-quadruplex structures (37). In the presence of the ligand, *WT* LTR promoter activity decreased to 68% of the untreated control at the highest tested ligand concentration (37), whereas the activity of the mu-
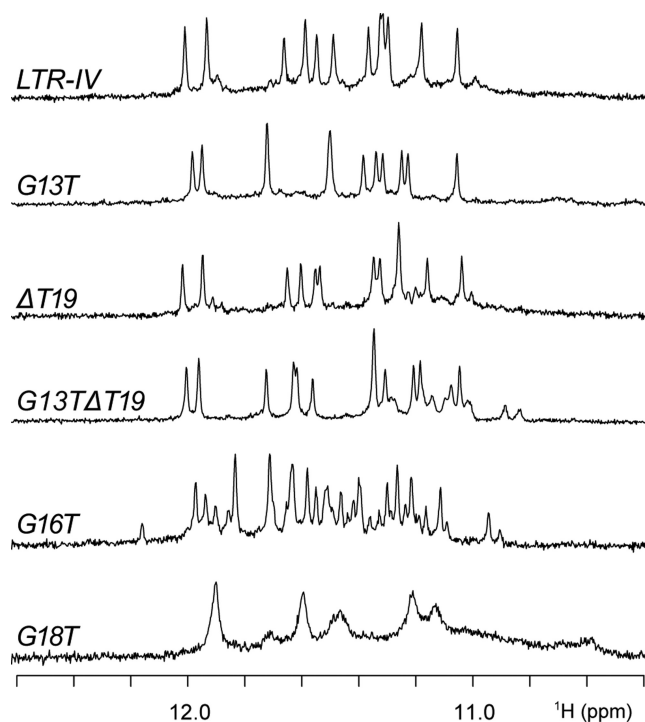
Figure 5. Imino proton NMR spectra of *LTR-IV* and sequences containing mutations and/or deletions in 70 mM potassium chloride and 20 mM potassium phosphate (pH 7).

Table 2. *LTR-IV* and mutated sequences

Name	Sequence								
	1 2	3 4 5	6	7 8 9	10 11 12	13 14 15 16	17	18 19 20 21	22
LTR-IV	CT	GGG	C	GGG	ACT	GGGG	A	GTGG	T
G13T	CT	GGG	C	GGG	ACT	TGGG	A	GTGG	T
ΔT19	CT	GGG	C	GGG	ACT	GGGG	A	G GG	T
G13TΔT19	CT	GGG	C	GGG	ACT	TGGG	A	G GG	T
G16T	CT	GGG	C	GGG	ACT	GGGT	A	GTGG	T
G18T	CT	GGG	C	GGG	ACT	GGGG	A	TTGG	T

Guanines involved in G-tetrad formation in the *LTR-IV* sequence are underlined. Sites of base mutations and deletions are boxed.

tant *G18T* LTR decreased to an even further extent to 33% of the untreated control (Figure 6B).

To assess which G-quadruplex species form in the full-length LTR promoter, a *Taq* polymerase stop assay was set up. The *WT* and mutant *G18T* templates were subjected to *Taq* polymerase processing in the absence or presence of K$^+$ and BRACO-19 (Figure 6C and D). The addition of the BRACO-19 ligand induced formation of *LTR-IV* only in the *WT* template and was found to stabilize *LTR-III* in both the *WT* and *G18T* templates. Interestingly, the *LTR-III* G-quadruplex was induced to a higher extent in the *G18T* template (Figure 6C and D).

Considering the dramatic effects of incorporating a single *G18T* base mutation on the HIV-1 promoter activity, we investigate the presence of such a mutation in about 1000 circulating viral strains available in the HIV sequence database (see 'Materials and Methods' section). We found a 99.5% degree of conservation for G18. There were five sequences observed to contain G18 mutations, four sequences contained a T mutation and one sequence an A at position G18

(Supplementary Tables S1 and 2). Collectively, these data support the crucial role of this position in virus replication and suggest that the *LTR-IV* G-quadruplex could act as a modulator of activity in the HIV-1 LTR promoter.

DISCUSSION

We demonstrate here that the *LTR-IV* G-quadruplex adopts a well-defined parallel-stranded G-quadruplex containing a T-bulge (Figure 4). The single-nucleotide T19 bulge of *LTR-IV* participates in a stacking interaction with A17 of a neighboring single-nucleotide propeller loop. While the formation and prevalence of bulged G-quadruplexes have been previously described (57), to our knowledge, the loop-bulge interaction has not been reported so far. Although base stacking interactions between loops have previously been observed (58–60), the loop-bulge stacking and its orientation in the groove of the G-quadruplex may offer specificity in the design of a ligand specific to this structure.

It has been previously shown that the *LTR-IV* region is highly conserved (37). This is particularly true for the guanine bases within this sequence, suggesting the importance of conserving the ability for *LTR-IV* to form a G-quadruplex structure. In this work, mutation studies of the *LTR-IV* sequence demonstrate that the presence of a T-bulge is not crucial to G-quadruplex formation, but can play a role in modulating the G-quadruplex stability (57). However, at the proviral level the A17 and T19 nucleotides are within the Sp1 binding site, one of the most important transcription factors in the LTR promoter (61,62). Among 953 HIV-1 strains a mutation of A17 or T19 is observed in only 5 cases (1 case A17-to-G and 4 cases T19-to-G), supporting their key role in the virus life cycle. Mutation at G18, which disrupts *LTR-IV* G-quadruplex, is also shown to be very rare with a conservation rate of 99.5%. Collectively, the high conservation of these residues suggests that the bulge feature of the *LTR-IV* G-quadruplex is maintained across strains.

The LTR promoter of the proviral HIV-1 genome is crucial both for viral transcription and latency. In fact, it alone controls transcription of all nine viral genes, and repression of transcription at this level has been proposed to be involved in the switch to viral latency (63,64). It has been previously shown that the U3 region of the LTR can form G-quadruplexes and modulate viral transcription (44).

Our results suggest that G-quadruplex formation in *LTR-III* and *LTR-IV* may act as a regulatory mechanism for transcription activity, with *LTR-IV* formation acting as a negative regulator of *LTR-III*. This interpretation is based on the following. (i) We have shown here that when the capacity for *LTR-IV* to form a G-quadruplex is lacking (by *G18T* mutation), the LTR promoter activity is highly silenced in cells (Figure 6A). This effect is likely caused by the increase in *LTR-III* G-quadruplex formation, in addition to the disruption of one of the three binding sites of the transcription factor Sp1. In contrast, we have previously shown that mutations that prevent formation of *LTR-III* lead to enhanced LTR promoter activity (37). (ii) In the presence of BRACO-19, the *G18T* mutant, lacking *LTR-IV* quadruplex-formation capacity, leads to increased *LTR-III*
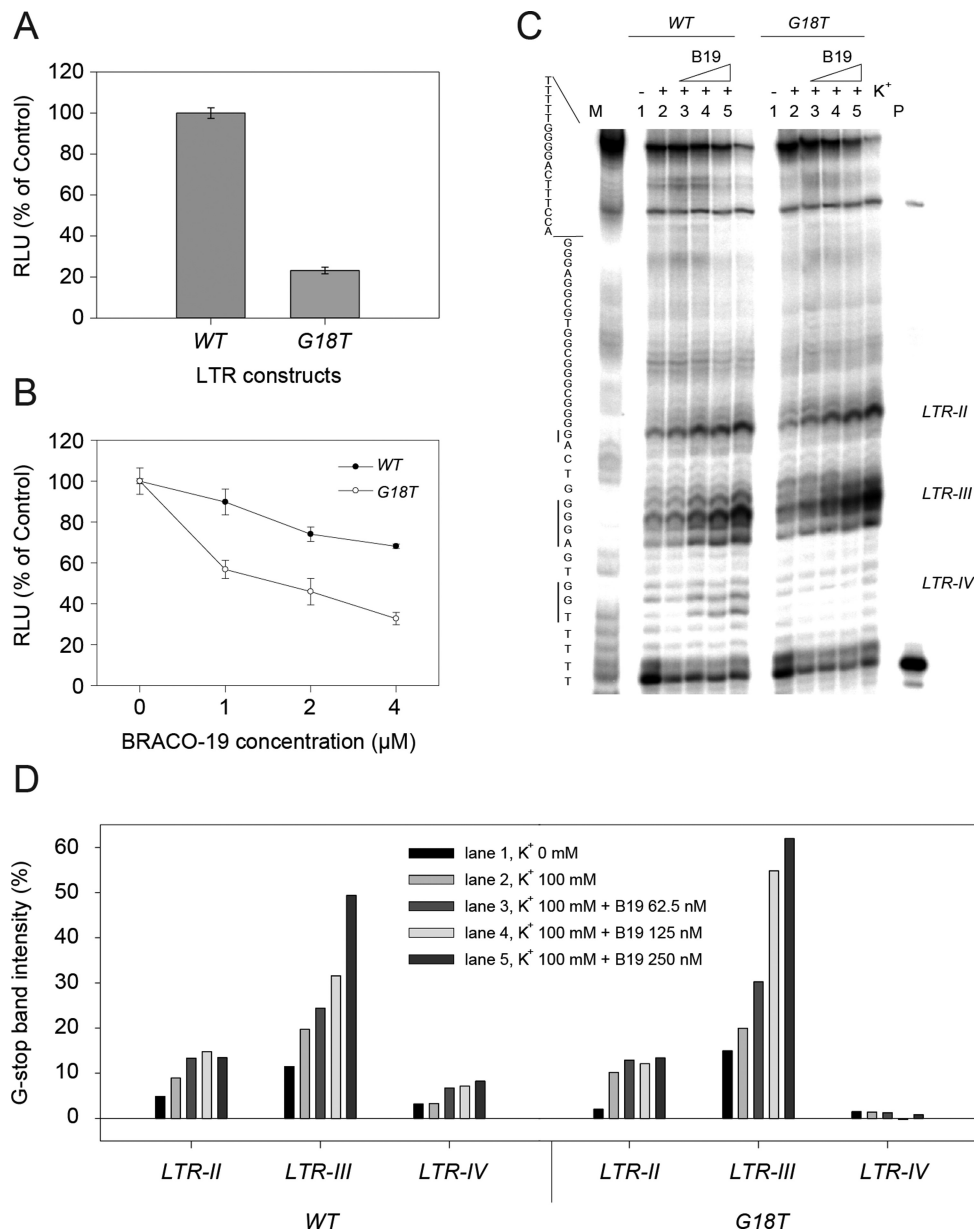
Figure 6. Biological effect of *LTR-IV* G-quadruplex. (**A**) Luciferase expression of the wild-type (*WT*) and mutant (*G18T*) *LTR-IV* normalized to total protein content in HEK 293T cells. (**B**) Normalized luciferase expression of the *WT* and mutant *G18T LTR-IV* in the presence of BRACO-19 (1.0–4.0 μM). (**C**) *Taq* polymerase stop assay performed in the absence or presence of K^+ (100 mM) and BRACO-19 (B19) (62.5–250 nM) as indicated on the LTR-II+III+IV *WT* and mutant *G18T* templates. The *WT* sequence is shown on the left. Stop regions corresponding to *LTR-II*, *LTR-III* and *LTR-IV* are shown on the right and are indicated on the *WT* sequence. M is a marker lane and P indicates the lane where only the primer was loaded as control. (**D**) Quantification of the intensity of the stop bands obtained in the *Taq* polymerase stop assay. Stop band intensity was quantified by ImageQuant TL software (GE Healthcare Europe, Milan, Italy). This assay was run in duplicate, with consistent behavior observed across experiments.

G-quadruplex formation *in vitro* and decreased promoter activity in cells (Figure 6B and D).

We demonstrate that the *LTR-IV* G-quadruplex forms less readily than the *LTR-III* G-quadruplex, with a low population of the *LTR-IV* G-quadruplex forming in the full-length promoter (Figure 6C and D). Since the *LTR-IV* and *LTR-III* G-quadruplexes are mutually exclusive, our present data suggest that an increase of the *LTR-IV* G-quadruplex folding would reduce the amount of folded *LTR-III*. Based on our data, a disruption of the *LTR-IV* G-

quadruplex formation reduced the LTR promoter activity, while a selective stabilisation of the *LTR-IV* G-quadruplex (e.g. via specific small molecules or proteins) would likely lead to increased transcriptional activity. A possible negative implication of over-active transcription could be premature cell death prior to completing virion production. Alternatively, over-active transcription could have ramifications on viral latency.

We propose the balance of *LTR-III* and *LTR-IV* G-quadruplex formation may act as a regulator of the viral

promoter. The rational and selective targeting of the *LTR-IV* structure could lead to sharp effects on HIV-1 replication. Understanding the structures of LTR G-quadruplexes is a valuable first step in the rational design of a selective small molecule to target this region of the viral genome and modulate the transcriptional behavior of HIV.

CONCLUSION

This work reports the NMR solution structure of the *LTR-IV* G-quadruplex formed in the LTR promoter of proviral HIV-1 genome. Biophysical data show that *LTR-IV* adopts a monomeric and parallel-stranded G-quadruplex in physiological solution. In the *LTR-IV* G-quadruplex, a single-nucleotide bulge interacts with a nearby single-nucleotide loop in a distinctive manner not reported previously. Conservation analysis shows these residues are highly conserved in viable HIV-1 strains. We suggest the loop-bulge interaction observed in *LTR-IV* is a promising feature that could be exploited in the design of molecules to specifically target this structure. The *LTR-IV* G-quadruplex structure presented here is valuable in understanding both the biology of the HIV-1 proviral LTR and its potential as a therapeutic target.

ACCESSION NUMBER

PDB ID: 2N4Y.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

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. Patel,D.J., Phan,A.T. and Kuryavyi,V. (2007) Human telomere, oncogenic promoter and 5′-UTR G-quadruplexes: diverse higher order DNA and RNA targets for cancer therapeutics. *Nucleic Acids Res.*, **35**, 7429–7455.
3. Rhodes,D. and Lipps,H.J. (2015) G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res.*, **43**, 8627–8637.
4. Sen,D. and Gilbert,W. (1990) A sodium-potassium switch in the formation of four-stranded G4-DNA. *Nature*, **344**, 410–414.
5. Lane,A.N., Chaires,J.B., Gray,R.D. and Trent,J.O. (2008) Stability and kinetics of G-quadruplex structures. *Nucleic Acids Res.*, **36**, 5482–5515.
6. Campbell,N.H. and Neidle,S. (2012) G-quadruplexes and metal ions. *Met. Ions Life Sci.*, **10**, 119–134.
7. da Silva,M.W., Trajkovski,M., Sannohe,Y., Hessari,N.M., Sugiyama,H. and Plavec,J. (2009) Design of a G-quadruplex topology through glycosidic bond angles. *Angew. Chem. Int. Ed.*, **48**, 9167–9170.
8. Phan,A.T. (2010) Human telomeric G-quadruplex: structures of DNA and RNA sequences. *FEBS J.*, **277**, 1107–1117.
9. Siddiqui-Jain,A., Grand,C.L., Bearss,D.J. and Hurley,L.H. (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 11593–11598.
10. 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.
11. Eddy,J. and Maizels,N. (2006) Gene function correlates with potential for G4 DNA formation in the human genome. *Nucleic Acids Res.*, **34**, 3887–3896.
12. Huppert,J.L. and Balasubramanian,S. (2007) G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.*, **35**, 406–413.
13. Eddy,J. and Maizels,N. (2008) Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes. *Nucleic Acids Res.*, **36**, 1321–1333.
14. Nakken,S., Rognes,T. and Hovig,E. (2009) The disruptive positions in human G-quadruplex motifs are less polymorphic and more conserved than their neutral counterparts. *Nucleic Acids Res.*, **37**, 5749–5756.
15. Verma,A., Yadav,V.K., Basundra,R., Kumar,A. and Chowdhury,S. (2009) Evidence of genome-wide G4 DNA-mediated gene expression in human cancer cells. *Nucleic Acids Res.*, **37**, 4194–4204.
16. Lopes,J., Piazza,A., Bermejo,R., Kriegsman,B., Colosio,A., Teulade-Fichou,M.P., Foiani,M. and Nicolas,A. (2011) G-quadruplex-induced instability during leading-strand replication. *EMBO J.*, **30**, 4033–4046.
17. 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.
18. Mani,P., Yadav,V.K., Das,S.K. and Chowdhury,S. (2009) Genome-wide analyses of recombination prone regions predict role of DNA structural motif in recombination. *PLoS One*, **4**, e4399.
19. Fry,M. and Loeb,L.A. (1994) The fragile X syndrome d(CGG)$_n$ nucleotide repeats form a stable tetrahelical structure. *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 4950–4954.
20. Fratta,P., Mizielinska,S., Nicoll,A.J., Zloh,M., Fisher,E.M.C., Parkinson,G. and Isaacs,A.M. (2012) C9orf72 hexanucleotide repeat associated with amyotrophic lateral sclerosis and frontotemporal dementia forms RNA G-quadruplexes. *Sci. Rep.*, **2**, 1016.
21. Fisette,J.F., Montagna,D.R., Mihailescu,M.R. and Wolfe,M.S. (2012) A G-rich element forms a G-quadruplex and regulates BACE1 mRNA alternative splicing. *J. Neurochem.*, **121**, 763–773.
22. Taylor,J.P. (2014) Neurodegenerative diseases: G-quadruplex poses quadruple threat. *Nature*, **507**, 175–177.
23. Haeusler,A.R., Donnelly,C.J., Periz,G., Simko,E.A., Shaw,P.G., Kim,M.S., Maragakis,N.J., Troncoso,J.C., Pandey,A., Sattler,R. *et al.* (2014) C9orf72 nucleotide repeat structures initiate molecular cascades of disease. *Nature*, **507**, 195–200.
24. Ivanov,P., O'Day,E., Emara,M.M., Wagner,G., Lieberman,J. and Anderson,P. (2014) G-quadruplex structures contribute to the neuroprotective effects of angiogenin-induced tRNA fragments. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 18201–18206.
25. Sket,P., Pohleven,J., Kovanda,A., Stalekar,M., Zupunski,V., Zalar,M., Plavec,J. and Rogelj,B. (2015) Characterization of DNA G-quadruplex species forming from C9ORF72 G$_4$C$_2$-expanded repeats associated with amyotrophic lateral sclerosis and frontotemporal lobar degeneration. *Neurobiol. Aging*, **36**, 1091–1096.
26. Wu,Y. and Brosh,R.M. Jr (2010) G-quadruplex nucleic acids and human disease. *FEBS J.*, **277**, 3470–3488.
27. Sissi,C., Gatto,B. and Palumbo,M. (2011) The evolving world of protein-G-quadruplex recognition: a medicinal chemist's perspective. *Biochimie*, **93**, 1219–1230.
28. 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.
29. Henderson,A., Wu,Y., Huang,Y.C., Chavez,E.A., Platt,J., Johnson,F.B., Brosh,R.M. Jr, Sen,D. and Lansdorp,P.M. (2014) Detection of G-quadruplex DNA in mammalian cells. *Nucleic Acids Res.*, **42**, 860–869.

30. Metifiot,M., Amrane,S., Litvak,S. and Andreola,M.L. (2014) G-quadruplexes in viruses: function and potential therapeutic applications. *Nucleic Acids Res.*, **42**, 12352–12366.

31. Norseen,J., Johnson,F.B. and Lieberman,P.M. (2009) Role for G-quadruplex RNA binding by Epstein-Barr virus nuclear antigen 1 in DNA replication and metaphase chromosome attachment. *J. Virol.*, **83**, 10336–10346.

32. Murat,P., Zhong,J., Lekieffre,L., Cowieson,N.P., Clancy,J.L., Preiss,T., Balasubramanian,S., Khanna,R. and Tellam,J. (2014) G-quadruplexes regulate Epstein-Barr virus-encoded nuclear antigen 1 mRNA translation. *Nat. Chem. Biol.*, **10**, 358–364.

33. Tan,J., Vonrhein,C., Smart,O.S., Bricogne,G., Bollati,M., Kusov,Y., Hansen,G., Mesters,J.R., Schmidt,C.L. and Hilgenfeld,R. (2009) The SARS-unique domain (SUD) of SARS coronavirus contains two macrodomains that bind G-quadruplexes. *PLoS Pathog.*, **5**, e1000428.

34. Artusi,S., Nadai,M., Perrone,R., Biasolo,M.A., Palu,G., Flamand,L., Calistri,A. and Richter,S.N. (2015) The Herpes Simplex Virus-1 genome contains multiple clusters of repeated G-quadruplex: Implications for the antiviral activity of a G-quadruplex ligand. *Antiviral Res.*, **118**, 123–131.

35. Tluckova,K., Marusic,M., Tothova,P., Bauer,L., Sket,P., Plavec,J. and Viglasky,V. (2013) Human papillomavirus G-quadruplexes. *Biochemistry*, **52**, 7207–7216.

36. Perrone,R., Nadai,M., Poe,J.A., Frasson,I., Palumbo,M., Palu,G., Smithgall,T.E. and Richter,S.N. (2013) Formation of a unique cluster of G-quadruplex structures in the HIV-1 Nef coding region: implications for antiviral activity. *PLoS One*, **8**, e73121.

37. Perrone,R., Nadai,M., Frasson,I., Poe,J.A., Butovskaya,E., Smithgall,T.E., Palumbo,M., Palu,G. and Richter,S.N. (2013) A dynamic G-quadruplex region regulates the HIV-1 long terminal repeat promoter. *J. Med. Chem.*, **56**, 6521–6530.

38. Amrane,S., Kerkour,A., Bedrat,A., Vialet,B., Andreola,M.L. and Mergny,J.L. (2014) Topology of a DNA G-quadruplex structure formed in the HIV-1 promoter: a potential target for anti-HIV drug development. *J. Am. Chem. Soc.*, **136**, 5249–5252.

39. Piekna-Przybylska,D., Sullivan,M.A., Sharma,G. and Bambara,R.A. (2014) U3 region in the HIV-1 genome adopts a G-quadruplex structure in its RNA and DNA sequence. *Biochemistry*, **53**, 2581–2593.

40. Joint United Nations Programme on HIV/AIDS (UNAIDS) (2013) http://www.unaids.org/.

41. Terwilliger,E.F., Sodroski,J.G. and Haseltine,W.A. (1990) Mechanisms of infectivity and replication of HIV-1 and implications for therapy. *Ann. Emerg. Med.*, **19**, 233–241.

42. Pereira,L.A., Bentley,K., Peeters,A., Churchill,M.J. and Deacon,N.J. (2000) A compilation of cellular transcription factor interactions with the HIV-1 LTR promoter. *Nucleic Acids Res.*, **28**, 663–668.

43. Tosoni,E., Frasson,I., Scalabrin,M., Perrone,R., Butovskaya,E., Nadai,M., Palu,G., Fabris,D. and Richter,S.N. (2015) Nucleolin stabilizes G-quadruplex structures folded by the LTR promoter and silences HIV-1 viral transcription. *Nucleic Acids Res.*, **43**, 8884–8897.

44. Perrone,R., Butovskaya,E., Daelemans,D., Palu,G., Pannecouque,C. and Richter,S.N. (2014) Anti-HIV-1 activity of the G-quadruplex ligand BRACO-19. *J. Antimicrob. Chemother.*, **69**, 3248–3258.

45. Perrone,R., Doria,F., Butovskaya,E., Frasson,I., Botti,S., Scalabrin,M., Lago,S., Grande,V., Nadai,M., Freccero,M. *et al.* (2015) Synthesis, binding and antiviral properties of potent core-extended naphthalene diimides targeting the HIV-1 long terminal repeat promoter G-quadruplexes. *J. Med. Chem.*, **58**, 9639–9652.

46. Cantor,C.R., Warshaw,M.M. and Shapiro,H. (1970) Oligonucleotide interactions. 3. Circular dichroism studies of the conformation of deoxyoligonucleotides. *Biopolymers*, **9**, 1059–1077.

47. Phan,A.T. and Patel,D.J. (2002) A site-specific low-enrichment ^{15}N,^{13}C isotope-labeling approach to unambiguous NMR spectral assignments in nucleic acids. *J. Am. Chem. Soc.*, **124**, 1160–1161.

48. Huang,X., Yu,P., LeProust,E. and Gao,X. (1997) An efficient and economic site-specific deuteration strategy for NMR studies of homologous oligonucleotide repeat sequences. *Nucleic Acids Res.*, **25**, 4758–4763.

49. Phan,A.T. (2000) Long-range imino proton-^{13}C J-couplings and the through-bond correlation of imino and non-exchangeable protons in unlabeled DNA. *J. Biomol. NMR*, **16**, 175–178.

50. Phan,A.T., Gueron,M. and Leroy,J.L. (2001) Investigation of unusual DNA motifs. *Methods Enzymol.*, **338**, 341–371.

51. Goddard,T.D. and Kneller,D.G. (2007). *SPARKY Version 3.115 ed.* University of California, San Fransico.

52. Schwieters,C.D., Kuszewski,J.J., Tjandra,N. and Clore,G.M. (2003) The Xplor-NIH NMR molecular structure determination package. *J. Magn. Reson.*, **160**, 65–73.

53. Case,D.A., Cheatham,T.E., Darden,T., Gohlke,H., Luo,R., Merz,K.M., Onufriev,A., Simmerling,C., Wang,B. and Woods,R.J. (2005) The Amber biomolecular simulation programs. *J. Comput. Chem.*, **26**, 1668–1688.

54. Cheong,V.V., Heddi,B., Lech,C.J. and Phan,A.T. (2015) Xanthine and 8-oxoguanine in G-quadruplexes: formation of a G·G·X·O tetrad. *Nucleic Acids Res.*, **43**, 10506–10514.

55. Adachi,A., Gendelman,H.E., Koenig,S., Folks,T., Willey,R., Rabson,A. and Martin,M.A. (1986) Production of acquired immunodeficiency syndrome-associated retrovirus in human and nonhuman cells transfected with an infectious molecular clone. *J. Virol.*, **59**, 284–291.

56. Maxam,A.M. and Gilbert,W. (1980) Sequencing end-labeled DNA with base-specific chemical cleavages. *Methods Enzymol.*, **65**, 499–560.

57. 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.

58. Hu,L., Lim,K.W., Bouaziz,S. and Phan,A.T. (2009) Giardia telomeric sequence d(TAGGG)$_4$ forms two intramolecular G-quadruplexes in K^+ solution: effect of loop length and sequence on the folding topology. *J. Am. Chem. Soc.*, **131**, 16824–16831.

59. Martadinata,H. and Phan,A.T. (2013) Structure of human telomeric RNA (TERRA): Stacking of two G-quadruplex blocks in K^+ solution. *Biochemistry*, **52**, 2176–2183.

60. Marusic,M., Veedu,R.N., Wengel,J. and Plavec,J. (2013) G-rich VEGF aptamer with locked and unlocked nucleic acid modifications exhibits a unique G-quadruplex fold. *Nucleic Acids Res.*, **41**, 9524–9536.

61. Jeeninga,R.E., Hoogenkamp,M., Armand-Ugon,M., de Baar,M., Verhoef,K. and Berkhout,B. (2000) Functional differences between the long terminal repeat transcriptional promoters of human immunodeficiency virus type 1 subtypes A through G. *J. Virol.*, **74**, 3740–3751.

62. Li,Y.C., Mak,G. and Franza,B.R. (1994) In-vitro study of functional involvement of Sp1, Nf-Kappa-B/Rel, and Ap1 in Phorbol 12-Myristate 13-Acetate-mediated HIV-1 long terminal repeat activation. *J. Biol. Chem.*, **269**, 30616–30619.

63. Duverger,A., Wolschendorf,F., Zhang,M., Wagner,F., Hatcher,B., Jones,J., Cron,R.Q., van der Sluis,R.M., Jeeninga,R.E., Berkhout,B. *et al.* (2013) An AP-1 binding site in the enhancer/core element of the HIV-1 promoter controls the ability of HIV-1 to establish latent infection. *J. Virol.*, **87**, 2264–2277.

64. van der Sluis,R.M., Jeeninga,R.E. and Berkhout,B. (2013) Establishment and molecular mechanisms of HIV-1 latency in T cells. *Curr. Opin. Virol.*, **3**, 700–706.