7588–7601 *Nucleic Acids Research, 2021, Vol. 49, No. 13* *Published online 2 July 2021*
https://doi.org/10.1093/nar/gkab518

GGGCTA repeats can fold into hairpins poorly unfolded by replication protein A: a possible origin of the length-dependent instability of GGGCTA variant repeats in human telomeres

Jean Chatain[1], Alain Blond[2], Anh Tuân Phan[3,4], Carole Saintomé[1,5] and Patrizia Alberti[1,*]

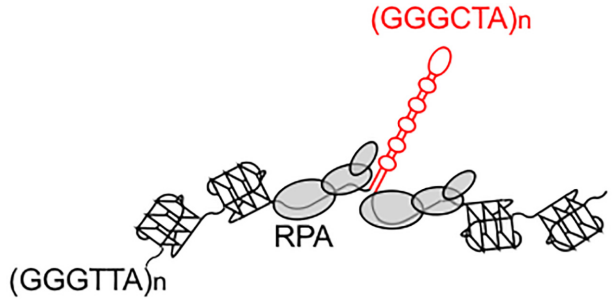
[1]Laboratoire Structure et Instabilité des Génomes (StrInG), Muséum national d'Histoire naturelle, CNRS, Inserm, Paris 75005, France, [2]Laboratoire Molécules de Communication et Adaptation des Microorganismes (MCAM), Muséum national d'Histoire naturelle, CNRS, Paris 75005, France, [3]School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, Singapore, [4]NTU Institute of Structural Biology, Nanyang Technological University, Singapore 636921, Singapore and [5]Sorbonne Université, UFR927, Paris 75005, France

Received December 07, 2020; Revised June 01, 2021; Editorial Decision June 02, 2021; Accepted June 30, 2021

ABSTRACT

Human telomeres are composed of GGGTTA repeats and interspersed with variant repeats. The GGGCTA variant motif was identified in the proximal regions of human telomeres about 10 years ago and was shown to display a length-dependent instability. In parallel, a structural study showed that four GGGCTA repeats folded into a non-canonical G-quadruplex (G4) comprising a Watson–Crick GCGC tetrad. It was proposed that this non-canonical G4 might be an additional obstacle for telomere replication. In the present study, we demonstrate that longer GGGCTA arrays fold into G4 and into hairpins. We also demonstrate that replication protein A (RPA) efficiently binds to GGGCTA repeats structured into G4 but poorly binds to GGGCTA repeats structured into hairpins. Our results (along with results obtained with a more stable variant motif) suggest that GGGCTA hairpins are at the origin of GGGCTA length-dependent instability. They also suggest, as working hypothesis, that failure of efficient binding of RPA to GGGCTA structured into hairpins might be involved in the mechanism of GGGCTA array instability. On the basis of our present and past studies about telomeric G4 and their interaction with RPA, we propose an original point of view about telomeric G4 and the evolution of telomeric motifs.

GRAPHICAL ABSTRACT



INTRODUCTION

Telomeres are the nucleoprotein complexes that constitute the end of eukaryotic linear chromosomes and protect them from erosion, degradation and inappropriate repair, solving the end-replication and the end-protection problems (1). In eukaryotes where telomeres are elongated by telomerase, the telomeric DNA strand running toward the 3′ end (the G-strand) is generally composed of repeats of a short motif (5–8 nucleotides) carrying consecutive guanines (two, three or four) (2) and ends with a 3′ single-stranded overhang (the G-overhang). The peculiar DNA and protein composition of telomeres (*shelterins*) can make them fold into a lariat structure (the t-loop) (3), first revealed, ahead of time, in the linear mitochondrial DNA of a ciliate (4) and proposed to be the primordial structure that allowed the stabilization of linear chromosomes (5,6). Because of the presence of consecutive guanines, telomeric G-strands from a variety of organisms are prone to fold into G-quadruplexes (G4) (7). G4 are four-stranded structures relying on the stacking of tetrads of guanines (G-tetrads or G-quartets) and sta-

*To whom correspondence should be addressed. Tel: +33 1 40793727; Fax: +33 1 40793705; Email: alberti@mnhn.fr

bilized by several cations (notably K^+ and, to a lesser extent, Na^+) that fit into the central cavity of the tetrahelical structure and coordinate with the carbonyl oxygens of the guanines. There is evidence that G4 do form at telomeres (in the single-stranded G-overhang as well as in the double-stranded region), where they appear to be friends or foes depending on the process under investigation (8). In human cells, several proteins are necessary for efficient replication of the telomeric G-strand, among them the helicases WRN, BLM and RTEL1 and the single-stranded DNA binding complex CST; all of them have been reported to play a role in unfolding telomeric G4 (9–15).

In humans, the telomeric motif is the hexamer GGGTTA (16). Nevertheless, human telomeres display a certain degree of sequence polymorphism. In the proximity of the subtelomeric regions, GGGTTA canonical repeats are interspersed with variant repeats, such as GGGGTT, GGGTGA, GGGTCA, … (17–20). Advances in genome sequencing technologies have allowed detecting the presence of variant motifs all along the telomeres (21–23). About ten years ago, Mendez-Bermudez *et al*. identified within the proximal region of telomeres a variant motif that was highly unstable in the male germline, the GGGCTA motif, estimated to be present in 7% of human telomeres (24). When a homogenous array of at least eleven GGGCTA repeats was present in the telomere of a father, the progeny displayed gains or losses of these variant repeats (24). GGGCTA array instability was also detected in colon samples of a single individual (24). In parallel, a nuclear magnetic resonance (NMR) study revealed that four repeats of this telomeric variant motif (the $A(GGGCTA)_3GGG$ sequence) folded, in potassium, into a robust non-canonical G4: an antiparallel chair-type G4 formed by two G-tetrads sandwiched between a Watson–Crick GCGC tetrad and a Watson–Crick GC pair in a lateral loop (25). The presence of a non-canonical GCGC tetrad and the chair-type conformation made this structure very different from the G4 structures formed by the GGGTTA canonical motif (Supplementary Figure S1). The mechanism underlying GGGCTA instability was not elucidated. It was suggested that GGGCTA instability was replication-dependent (since it was observed when GGGCTA arrays were transmitted *via* the male germline) and speculated that the non-canonical G4 structure could be an additional obstacle for G-strand replication (24).

After having investigated the structure and the stability of long telomeric sequences composed of the GGGTTA canonical motif (26), we reconsidered the GGGCTA variant motif in the light of a question: why GGGCTA instability was observed only starting from eleven repeats? We wondered what kind of structures are formed by more than four GGGCTA repeats. Do eight, twelve, sixteen, … GGGCTA repeats fold into two, three, four, … contiguous G4 as GGGTTA canonical repeats do (26) or do they fold into other structures? With the aim to obtain some insights into the origin of the length-dependent instability of the GGGCTA telomeric variant motif, we undertook a structural investigation of sequences composed of an increasing number of GGGCTA repeats and studied the interaction of human replication protein A (RPA) with these structures. RPA is considered the single-stranded DNA's first responder (27), it binds to transiently exposed single-stranded DNA and participates in the orchestration of DNA metabolic pathways. RPA is present at human telomeres during replication (28). Studies about telomere replication in the fission yeast *Schizosaccharomyces pombe* revealed a transient accumulation of RPA on the lagging G-strand (29) and provided evidence that weaking the binding of RPA to DNA led to problems in the replication of the lagging G-strand (30). In previous studies we showed that human telomeric sequences fold into contiguous G4 units (26) that are efficiently bound and unfolded by human RPA, independently of their number (31). In the present study we provide evidence that, GGGCTA arrays composed of more than four repeats fold into G4 and into hairpins and that GGGCTA hairpins impair RPA binding. Our results suggest a possible mechanism underlying the length-dependent instability of GGGCTA arrays in human telomers and raise a few considerations about telomeric G4 and the evolution of telomeric motifs.

MATERIALS AND METHODS

Oligonucleotides

Oligonucleotides were purchased from Eurogentec. GGGTTA, GGGCTA and GGGTCA repeat oligonucleotides were polyacrylamide gel electrophoresis (PAGE) purified, the other oligonucleotides were Reverse-Phase Cartridge Gold™ purified. Oligonucleotides were dissolved in bi-distilled water at a concentration of 200 μM and stocked at −20°C. Concentrations were determined using molar extinction coefficients provided by the manufacturer. Oligonucleotide sequences are listed in Supplementary Table S1.

Proteins

Recombinant human RPA1, RPA2 and RPA3 were co-expressed in the *Escherichia coli* BL21-DE3 strain transformed with the pET11a-hRPA plasmid (a kind gift from Dr Klaus Weisshart, IMB, Jena, Germany), purified following published protocols (30) and stocked in 25 mM Tris–HCl pH 7.5, 50 mM NaCl, 1 mM ethylenediaminetetraacetic acid (EDTA), 1 mM DTT, 10% glycerol; aliquots were stored at −80°C. Recombinant human POT1 and human POT1-TPP1 were expressed and co-expressed, respectively, in the baculovirus insect cell system (pFastBac HT-A vector) at the IGBMC-CERBM platform (Strasbourg). Bacmids were a kind gift from Dr Ming Lei (Shanghai Research Center, China). The full hPOT1 (residues 1–634) was expressed with 6 histidines in N-term and a truncated version of human TPP1 (87–334) with a GST-tag in N-term; tags were successively removed. Detailed purification protocols are published in Jean Chatain's Ph.D. thesis (https://hal.archives-ouvertes.fr/tel-03218825).

UV-absorption measurements: melting curves, thermal difference spectra and circular dichroism spectra

Oligonucleotides were diluted in a cacodylic acid buffer (10 mM) at pH 7.2 (adjusted with LiOH), supplemented

with LiCl or NaCl or KCl (oligonucleotide and salt concentrations are reported in figure legends).

Melting curves were obtained by recording the absorbance at fixed wavelengths (335, 295, 273, 260 and 245 nm) as a function of temperature, first upon decreasing the temperature from 95°C to 5°C at 0.2°C min^{-1} and then upon increasing the temperature from 5°C to 95°C at 0.2°C min^{-1}; between the cooling and the heating steps, the temperature was kept at 5°C for 20 min. Melting curves at 295, 273, 260 and 245 nm were corrected by subtracting the absorbance at 335 nm as a function of temperature. For each sample, cooling and heating curves were superimposable (no hysteresis), this indicated that they corresponded to equilibrium states. Melting experiments were performed at least twice, melting profiles were reproducible. We defined 'transition temperature T_t' the maximum of the first derivative of absorbance as a function of temperature.

For each sample, the thermal difference spectrum (TDS) was obtained by subtracting to the absorbance spectrum at high temperature (95°C) the absorbance spectrum at low temperature (5°), recorded after cooling the sample from high to low temperature at 0.2°C min^{-1}. TDS provide information about nucleic acid structures (32).

Circular dichroism (CD) spectra were recorded at 5°C after annealing the samples from high temperature at 0.2°C min^{-1}; each CD spectrum was obtained by averaging three scans recorded at a speed of 500 nm/min^{-1} and was corrected by subtracting the spectrum of a water filled quartz cell.

UV-absorption measurements were acquired on an Uvikon XS spectrophotometer (Secoman), coupled to a circulation thermocryostat (Julabo) for temperature control; temperature was measured with a sensor immersed into a water filled quartz cell. CD spectra were recorded on a J-810 spectropolarimeter (Jasco), equipped with a Peltier thermostated cell holder. Measurements were acquired in quartz cuvettes with an optical pathlength of 1 cm (Hellma).

Nuclear magnetic resonance

Oligonucleotides were diluted in a cacodylic acid buffer (10 mM) at pH 7.2 (adjusted with LiOH), supplemented with LiCl or NaCl or KCl (oligonucleotide and salt concentrations are reported in figure legend), heated at 90°C for a couple of minutes and slowly cooled to room temperature. D_2O (10% of total volume) was added to sample solutions before measurements. One dimensional 1H NMR spectra were acquired at 298 K on an Avance III HD 600 MHz spectrometer (Bruker), equipped with a 5 mm triple resonance TCI cryoprobe (^{1}H-^{13}C-^{15}N) including shielded Z-gradients. For each sample, DSS, dissolved in the solvents used for sample measurements, was used as external chemical shift standard. Excitation sculpting was used for peak solvent suppression (pulse program zgesgp). For each sample, 4096 scans were accumulated (with the exception of H46cta in KCl 50 mM, for which 10400 scans were accumulated).

Polyacrylamide gel electrophoresis (PAGE)

Oligonucleotides were diluted in a cacodylic acid buffer (10 mM) at pH 7.2 (adjusted with LiOH), supplemented with LiCl or NaCl or KCl 100 mM (oligonucleotide concentrations are reported in figure legends), heated at 90°C for a couple of minutes and slowly cooled to 4°C. After addition of sucrose (10%), oligonucleotide solutions were loaded (10 or 20 μl) in a 12% polyacrylamide gel (acrylamide:bisacrylamyde 19:1), prepared in a TBE buffer supplemented with LiCl or KCl or NaCl 20 mM. Electrophoresis were run in a TBE buffer supplemented with LiCl or KCl or NaCl 20 mM, in a cold room. Depending on their concentration, oligonucleotides were detected by UV-shadow at 254 nm with a G:BOX (Syngene), or stained with methylene blue (0.02%) and detected with a ChemiDoc imager (Bio-Rad), or stained with ethidium bromide (0.5 μg/ml) and detected with a Typhoon FLA9500 (GE Healthcare Life Sciences).

Electrophoresis mobility shift assays (EMSA)

Oligonucleotides were labeled with [γ^{32}P]ATP using a T4 polynucleotide kinase (NEB). Radiolabeled oligonucleotides were diluted at 2.5 nM strand concentration in a buffer containing KCl or NaCl (HEPES 62.5 mM pH 7.9, BSA 0.125 mg ml^{-1}, KCl or NaCl 100 mM and glycerol 2.5%), heated for 10 min at 85°C and slowly cooled to room temperature. Serial dilutions of proteins were incubated (20 min at 4°C) in a dilution buffer containing KCl or NaCl 100 mM (Tris–HCl 50 mM pH 7.5, KCl or NaCl 100 mM, DTT 1 mM, BSA 0.2 mg ml^{-1}, EDTA 0.1 mM and glycerol 10%). A total of 2 μl of serial dilution protein solutions were then added to 8 μl of radiolabeled oligonucleotide solutions and incubated for 20 min at 20°C. RPA electrophoresis mobility shift assays (EMSA) were carried out in a native 5% polyacrylamide gel (acrylamide:bisacrylamide 29:1), in a 0.5× TBE buffer, at room temperature. POT1 and POT1-TPP1 EMSA were carried out in a 1% agarose gel, in a 0.5× TBE buffer, at room temperature. After electrophoresis, gels were dried and exposed on a Phosphorimager screen and scanned with a Typhoon FLA9500. The intensities I of the bands were quantified using ImageQuant software. For each protein concentration, the fraction of radiolabeled oligonucleotide bound to the protein was calculated as follows:

$$I_{oligont\ bound\ to\ protein} / (I_{free oligont} + I_{oligont\ bound\ to\ protein}).$$

For each RPA binding curve, an apparent equilibrium constant K_D was calculated from the IC_{50} value, where the IC_{50} is the concentration of RPA at which half of the oligonucleotides are bound to RPA. Under our condition (2 nM oligonucleotide concentration and IC_{50} varying from 10 to 120 nM), K_D is approximately equal to the IC_{50}. K_D values from all EMSA are reported in Supplementary Table S2.

RESULTS

Arrays of the unstable GGGCTA variant motif fold into G4 and into hairpins

In order to investigate the structure of GGGCTA repeats, we carried out an UV-absorption study of $A(GGGCTA)_nGGG$ oligonucleotides, with n = 3, 7, 11, 15. The $n = 3$ sequence (here named H22cta) is the sequence previously studied by Lim *et al.* (25). In potassium, it folds

into a chair-type G4 composed of two G-tetrads and a Watson–Crick GCGC tetrad (Supplementary Figure S1) (25). The $n = 7$, 11, 15 sequences (named H46cta, H70cta and H94cta, respectively) carry eight, twelve and sixteen runs of GGG and can potentially fold into two, three and four contiguous G4 units, respectively.

We recorded thermal difference spectra (TDS) and CD spectra at the following strand concentrations: 6 μM for H22cta, 3 μM for H46cta, 2 μM for H70cta and 1.5 μM for H94cta, so to have the same concentration of potential G4 units (6 μM) for all the sequences. At these strand concentrations, folding of H46cta, H70cta and H94cta into two, three and four contiguous G4 units of similar conformation, respectively, would result in identical TDS and identical CD spectra (identical in shapes and amplitudes), as we observed for GGGTTA canonical repeats (TDS and CD spectra of GGGTTA canonical repeats folding into contiguous G4 units published in our previous studies (26,33) are shown in Supplementary Figure S2). The TDS and CD spectrum of H22cta in potassium (Figure 1A and B, red spectra) were similar to those previously reported (25). Its TDS was characteristic of a G4 structure (32) and its CD spectrum of a G4 antiparallel conformation (34), its transition temperature (T_t) was about 64°C. With increasing the number of repeats, TDS and CD spectra of H46cta, H70cta and H94cta in potassium progressively shifted from those of the single G4 formed by H22cta oligonucleotide (Figure 1A and B, blue, green and black spectra). In particular, the TDS negative peak at 295 nm decreased, while CD spectra displayed a decrease in the 290 nm band and the appearance of a 275 nm band. These changes in TDS and CD spectra (compared to H22cta) suggest folding of H46cta, H70cta and H94cta into a mixture of two structures, a H22cta-like G4 structure and a non-G4 structure (a duplex?).

To test the hypothesis of an equilibrium between a G4 and a duplex structure, we carried out measurements under conditions that allowed decreasing G4 stability without affecting duplex stability. To this purpose, we studied the behavior of GGGCTA repeats in the presence of lithium and sodium cations (Li^+ does not stabilize G4, Na^+ stabilizes G4 to a lesser extent than K^+, the three cations stabilize duplexes to a similar extent). Replacing K^+ with Li^+ highlighted the formation of non-G4 structures with a high thermal stability: H46cta, H70cta and H94cta TDS and CD spectra lose the characteristic G4 signatures (Figure 1C and D) and melting curves displayed high transition temperatures that increased with the number of repeats (T_t from 54°C for H46cta to 57°C for H94cta) (Figure 1D, inset). In Na^+, H22cta still folded into a G4 with a T_t of 55°C (Figure 1E and F, red spectra) (as already reported (25)), while H46cta, H70cta and H94cta TDS and CD spectra shifted toward non-G4 signatures similar to those displayed in Li^+ (Figure 1E and F, blue, green and black spectra). Finally, upon decreasing KCl concentration from 100 to 10 mM, the CD spectrum of H46cta progressively shifted toward the non-G4 CD spectrum displayed in LiCl (Figure 1G). Overall, the above data suggested that GGGCTA repeats folded into G4 and into non-G4 structures and that both structures coexisted in potassium solutions.

GGGCTA repeats can potentially fold into duplex structures where four Watson–Crick base-pairs (bp) alternate with a couple of G-G bp (Figure 1H). In order to ascertain the formation of this kind of local structure, we recorded the CD spectrum of the duplex formed by the autocomplementary sequence 5′CTAGGGCTAG (named dxGG) (Figure 1H). The CD spectra of GGGCTA repeats in LiCl were similar to the CD spectrum of dxGG (Figure 1J): a negative band around 255 nm, a major positive band around 275 nm and a minor positive band around 297 nm (in order to compare CD spectra, we set the strand concentration of dxGG at 12 μM so to have 6 μM of dxGG duplex structure, that is approximately the concentration of dxGG duplex units harboured in 1.5 μM H94cta, 2 μM H70cta and 3 μM H46cta). The deeper negative band around 255 nm in GGGCTA repeats compared to dxGG might arise from a higher proportion (about twice) of couples of G-G bp in GGGCTA putative hairpin structure than in dxGG duplex [*e.g.* eight couples of G-G bp per thirty Watson–Crick bp in H94cta putative hairpin versus one couple of G-G bp per eight Watson–Crick bp in dxGG (Figure 1H)]. In KCl and in NaCl, the CD spectra of GGGCTA repeats appeared to result from the contribution of a H22cta-like CD spectrum and of a dxGG-like CD spectrum (Figure 1I and K).

In order to further ascertain the presence of an equilibrium between a G4 and a duplex structure, we carried out NMR measurements (Figure 2). The spectrum of H22cta in KCl 10 mM (Figure 2A) was similar to the one published in KCl 100 mM, with peaks in the 10–12 ppm Hoogsteen imino-proton region characteristic of G-tetrads, and peaks in the 12–14 ppm Watson–Crick imino-proton region that were previously ascribed to two G-C bp in the GCGC tetrad and to a G-C bp in the central loop (25). In NaCl and LiCl, H46cta displayed major peaks in the 12–14 ppm Watson–Crick region and only minor signals in the 10–12 ppm Hoogsteen region (Figure 2D and E), supporting a strong shift toward a duplex structure, in agreement with TDS and CD spectra in NaCl and LiCl shown in Figure 1C–F. In KCl, H46cta spectra displayed major peaks in both Hoogsteen and Watson–Crick regions (Figure 2B and C). Hoogsteen peaks supported the presence of G4, while Watson–Crick peaks similar to those displayed in NaCl and in LiCl (three distinct peaks around 13.7, 13.6 and 13.5 ppm, and a major peak around 12.7 ppm) supported the presence of duplexes. Recording H46cta spectra at two KCl concentrations (50 and 10 mM) allowed strengthening the evidence of an equilibrium between a G4 and a duplex structure. Indeed, upon decreasing the concentration of KCl from 50 to 10 mM, we observed an inversion of the relative heights of Hoogsteen peaks and Watson–Crick peaks (Figure 2B and C). This modulation in the Watson–Crick and Hoogsteen components can be formalized by considering the ratio between the integral of Watson–Crick peaks and the integral of Hoogsteen peaks, here defined as WC/H ratio. Upon decreasing the concentration of KCl from 50 to 10 mM, the WC/H ratio increased from 40/60 to 60/40 (Figure 2B and C), supporting a G4-to-duplex shift in agreement with CD spectra shown in Figure 1G. Finally, minor peaks suggested additional structural information about the duplex and the G4 fractions of H46cta in KCl. In the 10–11 ppm region, H46cta spectra in KCl displayed a 10.4 ppm peak also present in dxGG spectrum (Figure 2F) and characteristic

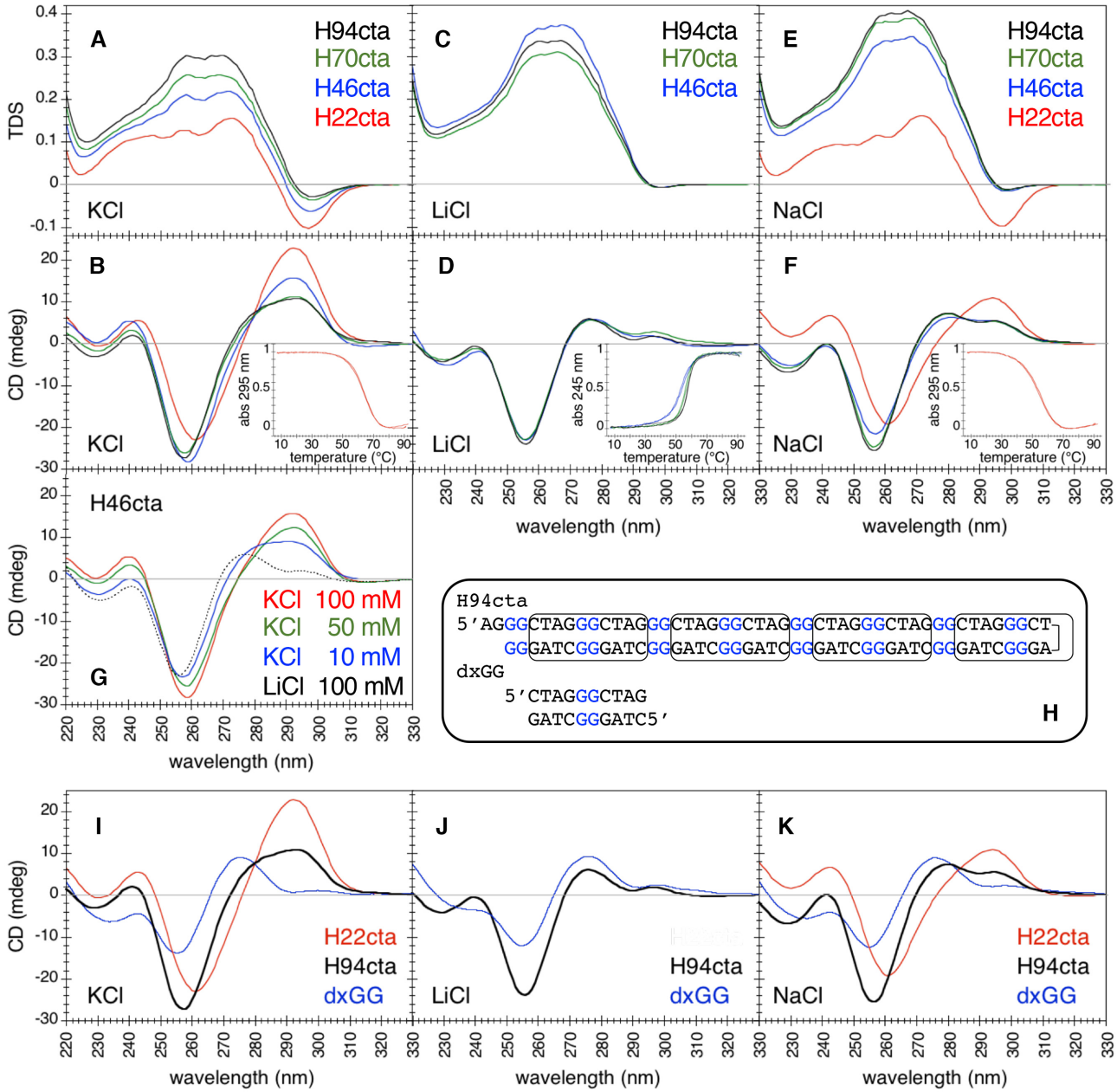


Figure 1. UV-absorption investigation of GGGCTA repeat structures. (**A**–**F**) TDS and CD spectra of H22cta 6 μM (red), H46cta 3 μM (blue), H70cta 2 μM (green) and H94cta 1.5 μM (black) in 100 mM KCl (A and B), LiCl (C and D) and NaCl (E and F). Panel D inset**:** normalized melting profiles at 245 nm of H46cta 3 μM (blue), H70cta 2 μM (green) and H94cta 1.5 μM (black) in LiCl 100 mM (T_t values range from 54°C for H46cta to 57°C for H94cta). Panel B and panel F insets**:** normalized melting profiles at 295 nm of the G4 structures formed by H22cta 6 μM in 100 mM KCl and NaCl (T_t 64°C and 55°C, respectively); in LiCl, H22cta does not fold into a stable structure. Melting profiles of H46cta, H70cta and H94cta in KCl and NaCl (not shown) are more complex and wavelength-dependent, consistently with the presence of more than one structure. (**G**) CD spectra of H46cta 3 μM in KCl 100 mM (red), KCl 50 mM (green), KCl 10 mM (blue) and in LiCl 100 mM (dotted black line). (**H**) H94cta putative hairpin structure and dxGG duplex. (**I**–**K**) CD spectra of the G4 formed by H22cta 6 μM (red), of the duplex formed by dxGG 12 μM (blue) and of H94cta 1.5 μM (black bold line) in 100 mM KCl (I), LiCl (J) and NaCl (K).

of G-G bp within Watson–Crick duplexes (35). In the 12.8–13.0 ppm region, H46cta spectrum in KCl 50 mM (Figure 2B) clearly displayed minor peaks at the same positions of the peaks arising from the GCGC tetrad in the G4 structure of H22cta (Figure 2A). The presence of these minor peaks in the spectra of H46cta in KCl suggests a dxGG-like conformation for the duplex fraction and a H22cta-like non-canonical conformation for the G4 fraction, in agreement with CD spectra.

Finally, by PAGE experiments, we assessed the molecularity of the duplex structures formed by GGGCTA repeats under conditions where the equilibrium was completely shifted toward the duplex structure, *i.e.* in LiCl. H46cta migrated in two bands with the mobility of a 22 and a 44 bp

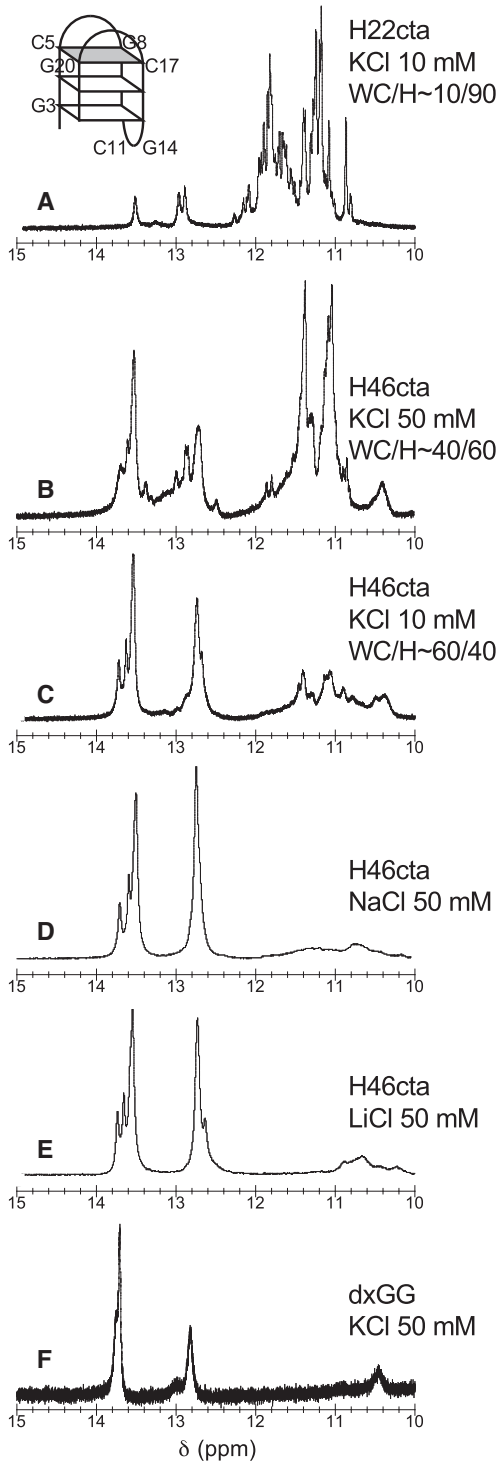


Figure 2. NMR investigation of GGGCTA repeat structures. NMR spectra of H22cta 160 μM in KCl 10 mM (**A**), of H46cta 100 μM in KCl 50 mM (**B**), KCl 10 mM (**C**), NaCl 50 mM (**D**) and LiCl 50 mM (**E**) and of dxGG 250 μM in KCl 50 mM (**F**). WC/H is the ratio between the integral of Watson–Crick peaks and the integral of Hoogsteen peaks. In (A), the folding of H22cta in KCl from (25) is schematically illustrated: the two G-tetrads are depicted in white, the non-canonical Watson–Crick GCGC tetrad is depicted in grey; the peak around 13.5 ppm had been assigned to C11-G14 Watson–Crick base-pairing in the central loop of the G4, the two peaks around 12.9 ppm had been assigned to G8-C17 and G20-C5 Watson–Crick base-pairing in the GCGC tetrad (25).

duplexes, indicating the formation of both intra- and intermolecular structures. The intramolecular structure (hairpin) was the major fraction. For longer repeats, only the hairpin structure was detected (Supplementary Figure S3). Additional PAGE experiments in KCl and in NaCl (that are discussed afterwards, Supplementary Figure S8) confirmed that GGGCTA repeats mainly fold into intramolecular structures.

In conclusion, UV-absorption, NMR and PAGE data demonstrated that H46cta, H70cta and H94cta folded into G4 structures and into duplex (mainly hairpin) structures. In LiCl and NaCl, the equilibrium was strongly shifted toward the duplex conformation. In KCl, the equilibrium between G4 and duplex markedly depended on the number of GGGCTA repeats and on K^+ concentration: increasing the number of repeats or decreasing K^+ concentration shifted the equilibrium toward the duplex conformation. Our data also provided information about the conformations of the G4 and of the duplex structures adopted by GGGCTA repeats. With respect to the conformation of the G4 fraction in potassium, H46cta, H70cta and H94cta CD spectra displayed a H22cta-like antiparallel component (Figure 1B and G). We argue that the G4 fraction in potassium may be structured into contiguous G4 units [similarly to GGGTTA canonical repeats (26)], where each G4 unit is folded into a H22cta-like conformation. With respect to the conformation of the duplex structure, CD, NMR and PAGE data supported, overall, the formation of hairpins as the one proposed in Figure 1H for H94cta (long hairpins based on a dxGG-like local structure).

Estimation of the hairpin fraction of GGGCTA oligonucleotides in solution

From CD spectra in NaCl (Figure 1F) and in KCl (Figure 1B and G), we roughly estimated the fractions of H46cta, H70cta and H94cta folded into the duplex structure under different salt conditions. Analysis of CD spectra is reported in Supplementary Figure S4A and results are shown in Figure 3A. In NaCl 100 mM, the duplex fraction attained 0.9. In KCl 100 mM, the duplex fraction increased from about 0.35 for H46cta to about 0.60 for H70cta and H94cta. When decreasing the concentration of KCl from 100 mM to 10 mM, the duplex fraction of H46cta increased from about 0.35 to about 0.75. The hairpin/G4 ratio estimated from CD spectra are consistent with the Watson–Crick/Hoogsteen ratio from NMR spectra (*e.g.* for H46cta in KCl 50 mM, the hairpin/G4 ratio from CD spectra and the WC/H ratio from NMR spectra are both equals to 0.40/0.60).

H46cta, H70cta and H94cta carry eight, twelve and sixteen GGG runs. They can potentially fold into two, three and four contiguous G4, respectively, where all GGGCTA repeats are engaged into a G4. We also studied GGGCTA oligonucleotides carrying a number of GGG runs not multiple of 4 ($A(GGGCTA)_{9,10,13}GGG$ carrying ten, eleven and fourteen GGG runs, and named H58cta, H64cta and H82cta respectively), with the aim to ascertain whether the presence of two or three repeats not engaged into a G4 might favor the hairpin form over the G4 form. Overall, the hairpin fraction in potassium solution increased with the

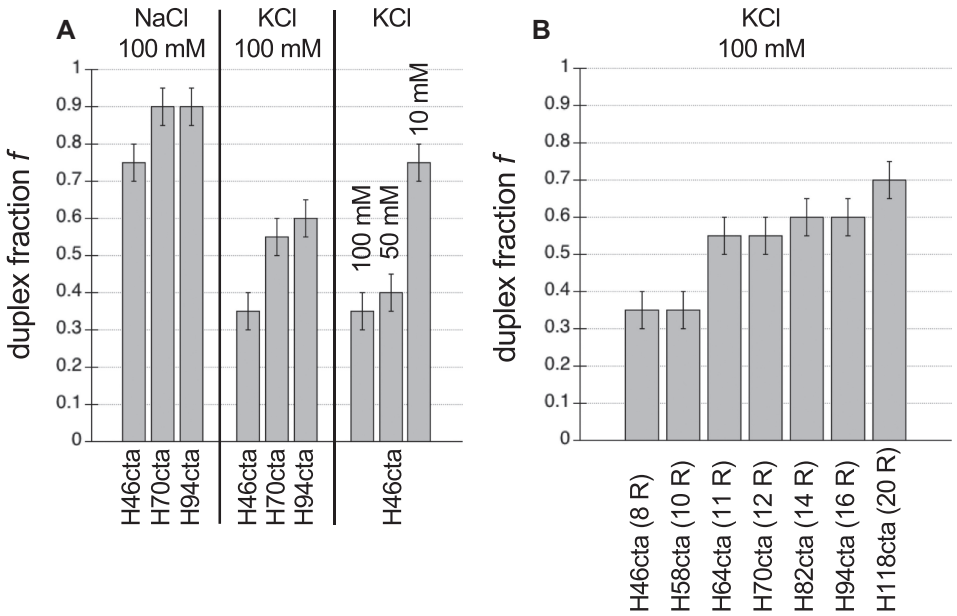


Figure 3. Estimated duplex fraction of GGGCTA repeats. (**A**) Fractions *f* of H46cta, H70cta and H94cta folded into the duplex form in NaCl and KCl 100 mM and of H46cta in KCl 100, 50 and 10 mM. (**B**) Fractions *f* of A(GGGCTA)$_n$GGG oligonucleotides folded into the duplex form in KCl 100 mM as a function of the number *R* of GGG runs ($R = n + 1$). Duplex fractions were estimated from CD spectra as detailed in Supplementary Figure S4A and B. The T_t of the duplex form (estimated in LiCl 100 mM) increased from 54°C for H46cta to 58/59°C for H118cta.

number of GGGCTA repeats (Figure 3B) (CD spectra and their analysis are reported in Supplementary Figure S4B). Notably, starting from eleven GGGCTA repeats (H64cta), the hairpin fraction became higher than 0.5, *i.e.* exceeded the G4 fraction. Finally, we ascertained that the hairpin fraction kept increasing when the number of GGGCTA repeats increased beyond 16 (Figure 3B). These data may explain two features of the length-dependent instability of GGGCTA arrays that will be discussed in the Discussion section.

RPA binds to GGGCTA repeats structured into G4 as efficiently as to GGGTTA canonical repeats, but poorly binds to GGGCTA repeats structured into hairpins

In a previous study, we showed that RPA efficiently binds to GGGTTA canonical repeats structured into contiguous G4, unfolding them, independently of the number of G4 units (31). Here we investigated by EMSA whether G4 and hairpins formed by GGGCTA variant repeats challenged RPA binding compared to contiguous G4 structures formed by GGGTTA canonical repeats (all K_D and errors are reported in Supplementary Table S2).

In potassium, RPA bound to H70cta (A(GGGCTA)$_{11}$GGG sequence) less efficiently than to its canonical counterpart H70tta (A(GGGTTA)$_{11}$GGG sequence) (Figure 4A) (we define 'binding efficiency' the percentage of DNA bound to RPA as a function of RPA concentration). From data in Figure 4A, the apparent binding constants K_D of RPA to H70cta and H70tta in potassium are (75 ± 5) nM and (36 ± 3) nM, respectively. Having demonstrated for GGGCTA repeats in potassium the existence of an equilibrium between H22cta-like G4 structures and duplexes, we wondered which of these structures impaired RPA binding to GGGCTA repeats in potassium: the non-canonical H22cta-like G4 conformation, the duplex structure or both? In KCl, we could not completely shift the equilibrium of H46cta, H70cta and H94cta toward the G4 or toward the duplex structure. Hence, in order to assess whether the non-canonical H22cta-like G4 conformation challenged RPA binding, we compared the binding of RPA to H22cta (A(GGGCTA)$_3$GGG sequence) and to its canonical counterpart H22tta (A(GGGTTA)$_3$GGG) in potassium (Figure 4B). In order to assess whether the hairpin structure challenged RPA binding, we studied the binding of RPA to H70cta in sodium, where the equilibrium of H70cta was strongly shifted toward the hairpin structure (Figure 4C). EMSA results showed that RPA bound to the non-canonical G4 formed by the GGGCTA variant motif in potassium as efficiently as to the G4 formed by the GGGTTA canonical motif (Figure 4B), but it poorly bound to the hairpin structure formed by H70cta (Figure 4C). For comparison, the apparent binding constant K_D of RPA to H70tta canonical repeats structured into G4 in potassium (Figure 4A, black curve) is (36 ± 3) nM, while the K_D of RPA to H70cta variant repeats structured into hairpins in sodium (Figure 4C) is (124 ± 16) nM). The K_D of RPA to the H70cta hairpin in sodium (Figure 4C) also provides the K_D of RPA to the H70cta hairpin fraction in potassium, since the nature of the monocation present in solution (K^+, Na^+ or Li^+) affects neither the stability of duplex structures (as shown in Supplementary Figure S5) nor the affinity of RPA for unstructured DNA (36). Overall, these EMSA results supported that the decrease in binding efficiency of RPA to H70cta compared to H70tta in potassium (Figure 4A) comes from the H70cta hairpin fraction and not from the H70cta G4 fraction. More

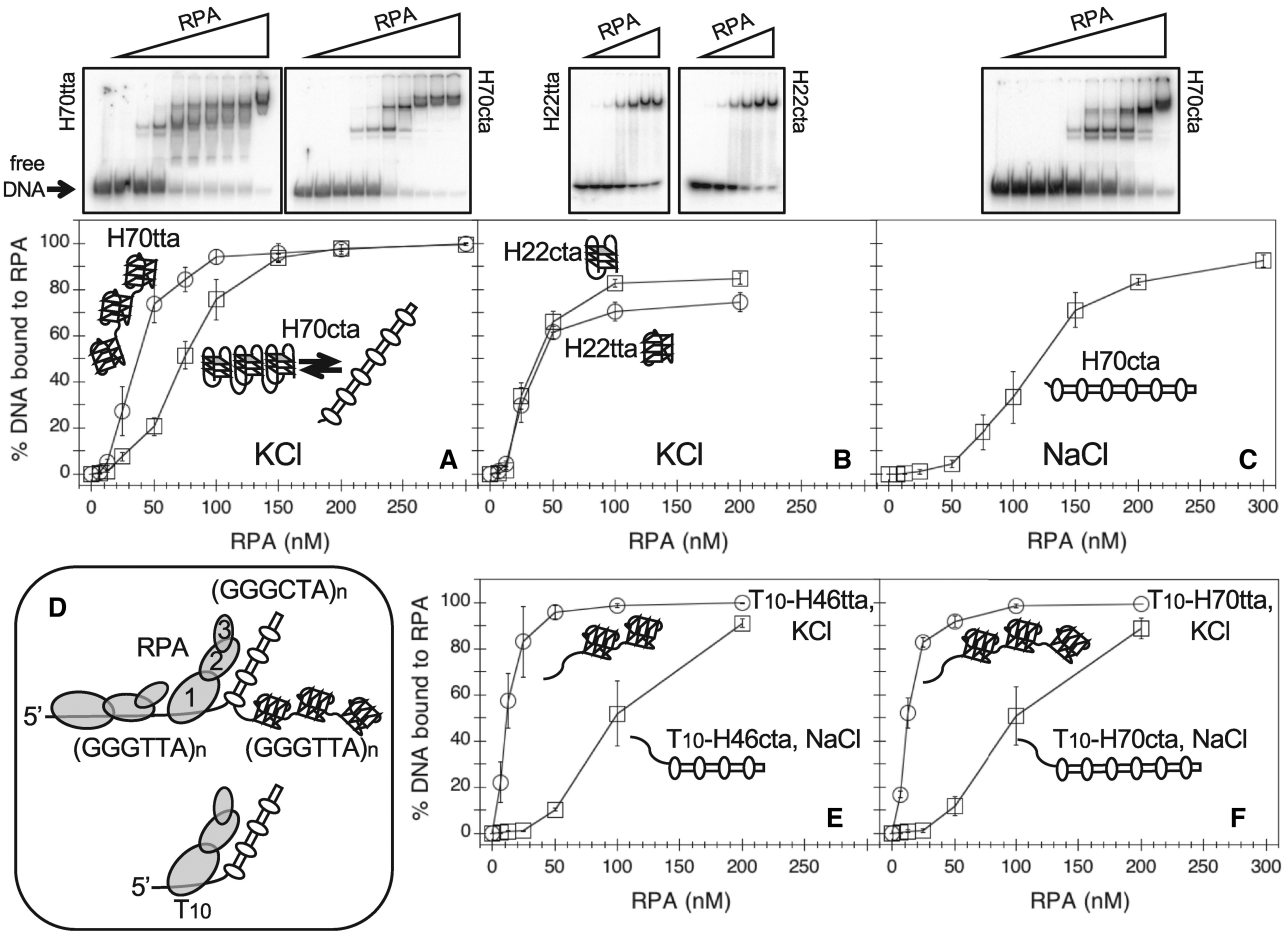


Figure 4. Binding efficiency of RPA to GGGCTA variant repeats and to GGGTTA canonical repeats. (**A**–**C**) Percentage of 2 nM radiolabeled oligonucleotide bound to RPA calculated from EMSA: (A) H70tta (circles) and H70cta (squares) in KCl 100 mM; (B) H22tta (circles) and H22cta (squares) in KCl 100 mM; (C) H70cta (squares) in NaCl 100 mM. Error bars are standard deviations from six assays. For each oligonucleotide, one representative EMSA gel is shown above the corresponding plot. RPA covers 25–30 nt in its extended binding-mode; for 70mer oligonucleotides, the shifted bands correspond to RPA:DNA complexes of increasing stoichiometry (1:1, 2:1 and 3:1). GGGTTA canonical repeat structures from our previous studies (26) and GGGCTA variant repeat structures from the present study are schematically illustrated. The Watson–Crick GCGC tetrad in H22cta G4 structure (25) is depicted in grey. (**D**) RPA can be faced with GGGCTA hairpins when already partially bound to telomeric DNA via its RPA1 subunit. In order to mimic this situation, we studied GGGCTA repeat hairpins bearing a T_{10} tail at their 5′ side, considering the 5′-to-3’ oriented and sequential binding mode of RPA to DNA. (**E,F**) Percentage of 2 nM radiolabeled oligonucleotide bound to RPA from EMSA experiments: (E) T10-H46tta in KCl 100 mM (circles) and T10-H46cta in NaCl 100 mM (squares), (F) T10-H70tta in KCl 100 mM (circles) and T10-H70cta in NaCl 100 mM (squares). Error bars are standard deviations from three assays. Representative gels are show in Supplementary Figure S10A and B.

broadly, these EMSA results demonstrated that GGGCTA repeats challenge RPA binding when they fold into hairpins but not when they fold into G4.

The heterotrimeric RPA binds DNA via four of its six OB-fold domains (A, B and C domains in RPA1 subunit and D domain in RPA2 subunit) in an oriented and sequential mode, where RPA1 binds first, followed by RPA2 toward the 3′ side of DNA (37). In a telomeric context, GGGCTA variant repeats are flanked by GGGTTA canonical repeats, hence RPA can be faced with GGGCTA hairpins when already partially bound to the telomeric G-strand via RPA1 (Figure 4D). In order to ascertain whether such an initial partial binding could increase the binding efficiency of RPA to GGGCTA repeats structured into hairpins, we carried out EMSA with repeats carrying a T_{10} tail at their 5′ side (Figure 4D). A 10-nt unstructured sequence provides an anchoring site for A and B domains in RPA1 (38,39). Its position at the 5′ side of the DNA structure can allow RPA 5′-to-3′ directional laying-out to achieve the stable 30-nt binding mode, thus favoring the unfolding of the DNA structure (37). EMSA revealed that RPA poorly bound to the hairpin structures formed by GGGCTA repeats even in the presence of an unstructured 10-nt single-stranded tail at the 5′ side of the hairpins structures (Figure 4E and F). From data in Figure 4E and F, the apparent binding constants K_D of RPA to T10-H46tta and T10-H70tta G4 in potassium are around 10 nM, while the apparent binding constants to T10-H46cta and T10-H70cta hairpins are around 100 nM.

Overall, EMSA results in Figure 4 demonstrated that RPA bound to GGGCTA repeats folded into G4 as efficiently as to GGGTTA repeats, but poorly bound to GGGCTA repeats folded into hairpins, even when it can partially bind in front of the hairpin structure (*i.e.* at the

5′ side). The 3-nt single-stranded loops in telomeric G4 may allow RPA to establish a first contact more easily than the available single-stranded bulges in GGGCTA hairpins. The higher stability of GGGCTA hairpins compared to telomeric G4 is likely an additional factor that makes RPA less efficient in binding to hairpin-structured GGGCTA repeats (Supplementary Figure S6).

Looking at the C-strand: a dozen of CCCTAG variant repeats do not fold into stable structures and are bound by RPA as efficiently as CCCTAA canonical repeats

GGGCTA variant repeats within the telomeric G-strand correspond to CCCTAG repeats within the C-strand. CCCTAG repeats might potentially fold into hairpins alternating four Watson–Crick bp and a couple of C-C bp, similarly to GGGCTA repeats where four Watson–Crick bp alternate with a couple of G-G bp (Supplementary Figure S7). We hence investigated whether CCCTAG repeats folded into hairpins and challenged RPA binding, as GGGCTA repeats did. While a dozen of GGGCTA repeats (H70cta) were sufficient to fold into stable hairpins that impaired RPA binding, a dozen of CCCTAG repeats did not fold into stable structures (at physiological pH and temperatures) and were bound by RPA as efficiently as CCCTAA canonical repeats (Supplementary Figure S7, K_D about 10 nM). We cannot exclude that a higher number of CCCTAG variant repeats might induce a local structuring of the C-strand, nevertheless this number would be beyond the 11-repeat threshold at which GGGCTA/CCCTAG repeat instability was observed (24). This asymmetry in the behavior of GGGCTA and CCCTAG repeats supports that the length-dependent instability of GGGCTA/CCCTAG repeats does not originate from structuring of CCCTAG variant repeats in the C-strand.

Arrays of the stable GGGTCA variant motif fold into structures that do not impair RPA binding

We also investigated a second variant motif, GGGTCA, that differs from the GGGCTA variant motif only for the position of the cytosine but does not display the same level of instability (24). As done for GGGCTA repeats, we investigated the structure of $A(GGGTCA)_{n=3,7,11,15}GGG$ oligonucleotides (named H22tca, H46tca, H70tca and H94tca) and their interaction with RPA. Like GGGTTA canonical repeats, also GGGTCA variant repeats folded into intramolecular structures, as shown by non-denaturing PAGE assay (Supplementary Figure S8). Overall, TDS, CD spectra and melting profiles of GGGTCA repeats in potassium provided evidence of structuring into G4 (Figure 5A and B), with melting profiles similar to the ones of canonical GGGTTA repeats (Figure 5B inset) (26). Nevertheless, CD spectra of GGGTCA in potassium markedly varied with the number of repeats (Figure 5B). We hence wondered whether a fraction of GGGTCA repeats folded into a non-G4 structure, similarly to GGGCTA repeats. Experiences in LiCl, where G4 were destabilized, revealed that H70tca and H94tca oligonucleotides folded into a structure characterized by a non-G4 TDS signature, a CD spectrum with a positive band around 254 nm and a negative band around 278 nm (we will refer to it as an 'inverted' CD spectrum) and a low thermal stability (T_t about 20°C) (Figure 5C and D). In NaCl, GGGTCA repeats displayed mixed TDS and CD signatures (Figure 5E and F): the signatures of H46tca (blue spectra) were shifted toward the antiparallel G4 signature displayed by H22tca (red spectra), while the signatures of H70tca and H94tca (green and black spectra) were shifted toward the inverted CD signature. Melting profiles further supported the presence of more than one structure in NaCl (Supplementary Figure S9). Intriguingly, a switch from a G4 CD signature in K^+ to an 'inverted' CD signature in Li^+ has also been reported for C_2G_4 repeats, for which a couple of hypothetical structures has been proposed (40). So far, we do not have enough information to propose a model for the GGGTCA structure characterized by the inverted CD signature and we cannot exclude that a minor fraction of this structure may be present in potassium solutions. Anyway, neither of the two GGGTCA repeats structures (G4 and non-G4) impaired RPA binding compared to GGGTTA canonical repeats. As shown by EMSA, RPA bound H70tca as efficiently as its canonical counterpart H70tta both in KCl (where the equilibrium of H70tca was shifted toward the G4 form) (Figure 5G) and in NaCl (where the equilibrium of H70tca was shifted toward the non-G4 form) (Figure 5H), likely because of the low stability of the non-G4 form at physiological temperatures. In conclusion, structures formed by GGGTCA arrays do not challenge RPA binding, contrary to hairpins formed by GGGCTA array.

POT1 and POT1-TPP1 do not bind to GGGCTA and GGGTCA repeats

While RPA binds to single-stranded DNA with no sequence-specificity and is present genome-wide, the protein Protection of Telomeres 1 (POT1) is specifically involved in telomere metabolism (1). *In vitro*, human POT1 has been reported to bind to its minimal tight-binding sequence TTAGGGTTAG with high specificity (41,42). In the *shelterin* complex, POT1 is bound to TPP1. Although TPP1 on its own does not bind to DNA, the complex POT1–TPP1 has a greater affinity for DNA than POT1 alone (43,44). The interaction between POT1 and TPP1 is required for localization of POT1 to telomeres (45,46). We tested the binding of POT1 and of the complex POT1–TPP1 to GGGCTA and GGGTCA variant repeats and found that neither POT1 nor POT1–TPP1 bound to sequences composed of either of these two variant motifs (Figure 6). We also verified that POT1 and POT1–TPP1 did not stably bind to longer GGGCTA and GGGTCA repeats, contrary to GGGTTA repeats (data not shown).

DISCUSSION

In previous studies, we showed that telomeric GGGTTA canonical repeats fold into contiguous G4 units (26) that are efficiently unfolded and bound by RPA, independently of their number (31). In the present study, we demonstrate that repeats of the unstable GGGCTA variant motif can fold either into G4 or into hairpins and that the hairpin form impairs RPA binding. Our results suggest hypotheses about the origin of the length-dependent instability of the telom-

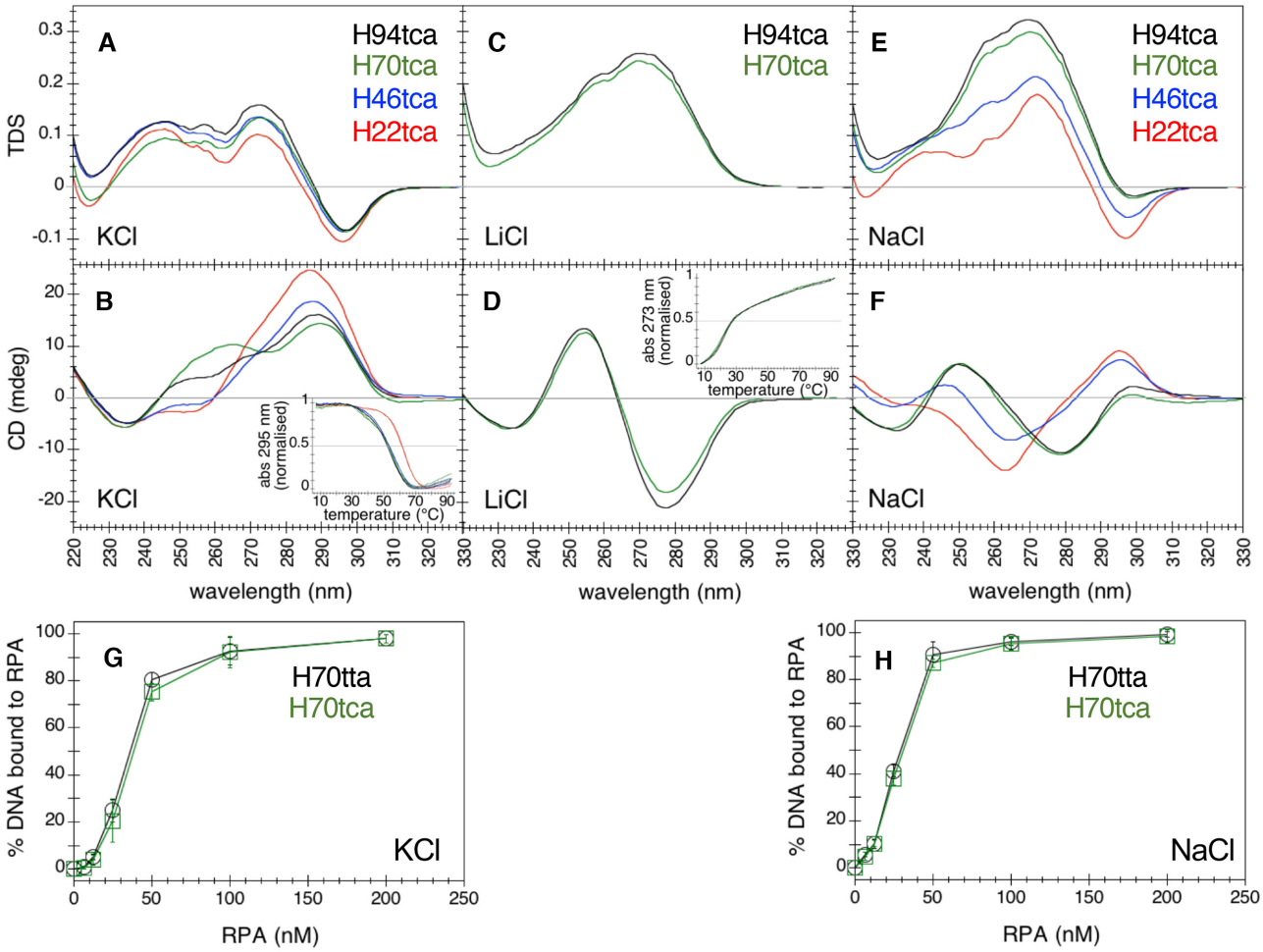


Figure 5. UV-absorption investigation of GGGTCA repeat structures. (**A**–**F**) TDS and CD spectra of H22tca 6 μM (red), H46tca 3 μM (blue), H70tca 2 μM (green) and H94tca 1.5 μM (black) in 100 mM KCl (A and B), LiCl (C and D) and NaCl (E and F). Panel B inset: normalized melting profiles of H22tca 6μM (red), H46tca 3 μM (blue), H70tca 2 μM (green) and H94tca 1.5 μM (black) in KCl 100 mM. Panel D inset: normalized melting profiles of H70tca 2 μM (green) and H94tca 1.5 μM (black) in LiCl 100 mM (H22tca and H46tca in LiCl do not fold into stable structures). Melting profiles in NaCl are shown in Supplementary Figure S9 (they are more complex and indicative of the presence of a G4 and a non G4-structure). (**G,H**) Percentage of 2 nM radiolabeled oligonucleotide bound to RPA from EMSA experiments: H70tta (black circles) and H70tca (green squares) in KCl 100 mM (G) and in NaCl 100 mM (H). Error bars are standard deviations from six assays. Representative gels are show in Supplementary Figure S10C.

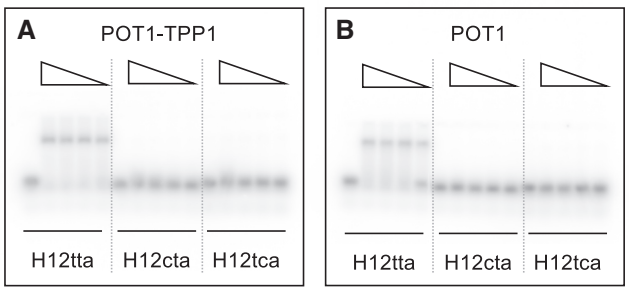


Figure 6. EMSA of canonical and variant telomeric repeats incubated with POT1 and POT1-TPP1. Representative EMSA of radiolabeled 5′GGTTAGGGTTAG (H12tta), 5′GGCTAGGGCTAG (H12cta) and 5′GGTCAGGGTCAG (H12tca) oligonucleotides (2 nM) incubated with POT1–TPP1 (**A**) or POT1 (**B**) in the presence of KCl 100 mM. For each oligonucleotide, protein concentration in lanes 1, 2, 3, 4 and 5 is 0, 200, 100, 50 and 10 nM, respectively.

eric GGGCTA variant motif and stimulate a few considerations about telomeric G4 and the evolution of telomeric motifs.

In summary, results of our structural investigation (Figure 1 and 2) demonstrate that, while the A(GGGCTA)$_3$GGG sequence folds into a non-canonical G4 [as previously demonstrated (25)], longer GGGCTA arrays can also fold into hairpins alternating $^{5'}$CTAG/$^{3'}$GATC Watson–Crick base-pairs and GG/GG base-pairs (Figure 1H). It is noteworthy that this kind of hairpin structure may be the intermediate structural state toward the folding of the A(GGGCTA)$_3$GGG sequence into the non-canonical G4 resolved in potassium solution (Supplementary Figure S1). In the presence of the physiological relevant cation K^+, G4 and hairpins coexist: a hairpin fraction was unambiguously detected starting from eight GGGCTA repeats (data not shown) and increased with the number of repeats (Figure 3). EMSA results show that RPA binds to GGGCTA variant repeats as efficiently

as GGGTTA canonical repeats when GGGCTA repeats are structured into G4, but less efficiently when GGGCTA repeats are structured into hairpins (Figure 4).

The existence of an equilibrium between G4 and hairpin structures in K^+ solutions, transposed into a chromosomal context, means that arrays of eight or more GGGCTA repeats can fold either into G4 or into hairpins (Figure 7, left side). In a simplified view (assuming that folding into G4 or into hairpins is stochastic), the probability *p* that a GGGCTA array folds into a hairpin in a chromosomal context corresponds to the fraction *f* of hairpins that it forms in solution. The instability of GGGCTA arrays appears to correlate with their propensity to fold into hairpins. Instability was observed for GGGCTA arrays of at least eleven repeats (24). Our data indicate that starting from eleven GGGCTA repeats (H64cta) the hairpin fraction *f* (or hairpin probability *p*) becomes higher than 0.5, *i.e.* the hairpin is more likely to form than the G4 (Figure 3B). Furthermore, almost all GGGCTA arrays reported in the article of Mendez-Bermudez *et al*. were composed of less than twenty repeats (out of about one hundred GGGCTA arrays, only one was longer than twenty repeats and composed of twenty-five repeats) (24). Our data indicate that the hairpin fraction *f* increases with the number of GGGCTA repeats (Figure 3B). Likely, above twenty GGGCTA repeats (H118cta, where *f* is around 0.7), the hairpin fraction further increases. These data suggest that long GGGCTA arrays, characterized by a high probability to fold into the hairpin form, might be more prone to deletion.

Since the '90s, many studies have shown that hairpin-prone Short Tandem Repeats (STR) can display replication- and transcription-dependent instability. Most of these studies focus on trinucleotide repeats (TNR), whose expansion leads to a variety of neurodegenerative diseases (many excellent reviews have been published about these topics). In human cells, replication-dependent hairpin formation by long sequences of trinucleotide repeats has been elegantly demonstrated by the use of sequence-specific hairpin-cleaving zinc finger nucleases (47). On the basis of our structural investigation, we propose that GGGCTA hairpins are the structures at the origin of the length-dependent instability of GGGCTA arrays in telomeres. GGGCTA hairpins can form, in particular, in the template or in the nascent G-strand during telomere replication or in the displaced G-strand of R-loops formed during transcription of *telomere repeat-containing RNA* (TERRA). As outlined in the Introduction section, RPA is the single-stranded DNA's first responder (27). It is present at human telomeres during replication (28). In fission yeast, it appears to transiently accumulate on the lagging G-strand (29) and to be important for its correct replication (30). RPA also colocalizes with R-loops, where it has been proposed to coat the displaced strand (48). On the basis of our EMSA results, we propose that the poor binding of RPA to GGGCTA hairpins might be involved in the mechanism of GGGCTA array instability when GGGCTA hairpins form during replication of the lagging G-strand or transcription of TERRA. Following the unwinding of the double-helix by helicases, GGGCTA hairpins might form faster than RPA binds to DNA. Alternatively, RPA might bind to GGGCTA repeats as soon as they become single-stranded and be subsequently displaced by hairpin formation (Figure 7). More broadly, we propose, as working hypothesis, that RPA might be the protein that fails first when faced with hairpins formed by Short Tandem Repeats on lagging-strands during replication or in the displaced strand of R-loops during transcription. Interestingly, studies in *E. coli* have shown the importance of SSB (the bacterial single-stranded DNA binding protein) in hairpin-prone trinucleotide repeat instability (49,50). In particular, overexpression of SSB led to a decrease in CTG/CAG trinucleotide repeat instability observed when the hairpin-prone CTG motif was on the lagging strand (50).

Human cells can repair slipped-out DNA repeats with different outcomes (correct, escaped or error-prone repair, leading to stability, contractions or expansions) (51). Several helicases have been reported to prevent hairpin-prone Short Tandem Repeat instability (for review (52)). In particular, the Werner helicase (WRN) has been shown to be involved in unwinding secondary structures formed by trinucleotide repeats, thereby promoting polδ-catalyzed DNA synthesis (53,54). Because of its interaction with both RPA and polδ (55–57), WRN is a good candidate to unwind GGGCTA hairpins on the lagging telomeric G-strand. However, the efficiency of correct repair has been shown to depend, among other factors, on the sequence of the slipped-out DNA, and more stable hairpins appears to be less efficiently correctly repaired (51,58).

In addition to RPA, at least two other single-stranded DNA binding proteins participate to telomere metabolism: the *shelterin* POT1 (1) and the CST complex [involved in the fill-in of the telomeric C-strand, in the synthesis of the telomeric lagging-strand and in rescue of stalled replication forks at both telomeric and non-telomeric sites (14,59,60)]. Might lack of binding of POT1 to variant repeats be involved in GGGCTA array instability during replication or in t-loop formation? Contrary to RPA, POT1 is depleted from replication forks at telomeres (61). The invasion of the telomeric G-overhang into a region of variant repeats would lead to a t-loop with mismatches in its double-stranded base (that might impair TRF2 binding) and with variant repeats in its single-stranded displaced portion (that, according to our results, would impair the potential binding of POT1(±TPP1)). Would such a t-loop be stable? Would it move/reform into a region of canonical repeats or would it trigger a DNA damage response? In any case, if lack of binding of POT1(±TPP1) (here demonstrated for both GGGCTA and GGGTCA repeats, Figure 6) is involved in GGGCTA instability, it cannot be the ultimate determinant of GGGCTA instability, otherwise instability would not depend on the length of GGGCTA arrays and, above all, also GGGTCA arrays would display a high level of instability, contrary to what has been observed (24). If lack of binding of POT1(±TPP1) to variant repeats is involved in GGGCTA instability, then a second single-stranded DNA binding protein must also be involved, a factor that is challenged by long GGGCTA arrays but not by GGGTCA arrays or short GGGCTA arrays. The different behavior of RPA toward the unstable GGGCTA repeats and the more stable GGGTCA repeats (Figure 5) makes it a good candidate to account for the length-dependent instability of GGGCTA arrays. Deeper investigations will be required to

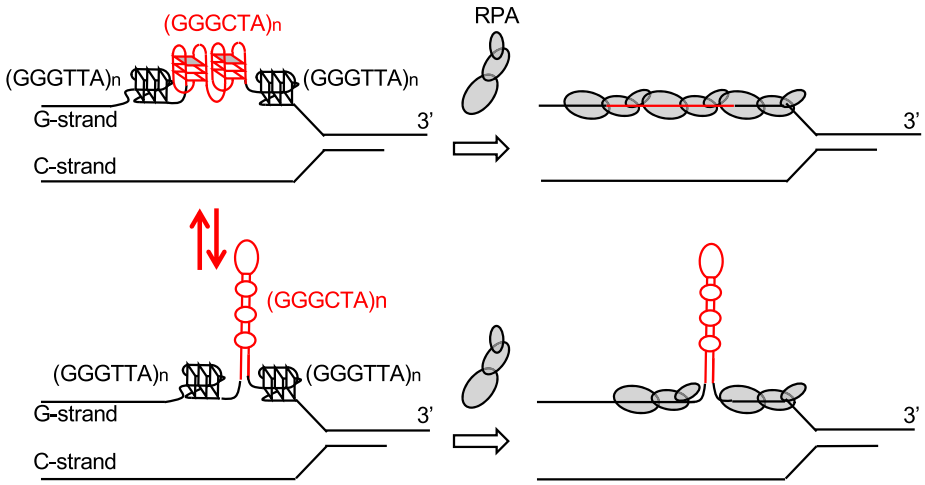


Figure 7. Hypothetical origin of the length-dependent instability of GGGCTA arrays. When tracts of the telomeric G-strand are transiently exposed as single-strands, *e.g.* during replication or TERRA transcription, GGGCTA arrays can fold into G4 or into hairpins (left side picture). The probability of folding into the hairpin structure increases with the number of GGGCTA repeats. G4 (formed by GGGTTA canonical repeats and by GGGCTA variant repeats) do not challenge RPA binding, whereas GGGCTA hairpins impairs RPA binding (right side picture). We propose that GGGCTA hairpins are the structures at the origin of GGGCTA array instability and that the poor binding of RPA to GGGCTA repeats structured into hairpins may be involved in GGGCTA array instability. Protruding hairpins can then lead to gains or losses of GGGCTA repeats via multiple possible mechanisms.

validate our working hypothesis about a possible implication of RPA in telomeric GGGCTA array instability (and more broadly in hairpin-prone Short Tandem Repeats instability) and to unravel the interplay between RPA, helicases and the CST complex in dealing with hairpins formed by Short Tandem Repeats.

Besides providing working hypotheses about the origin of GGGCTA telomeric variant repeat instability, our study stimulates a few speculative considerations about telomeric G4 and the evolution of telomeric motifs. Telomeric G4 are considered as problematic structures for the replication machinery (10). This point of view relies on evidence of troubles in telomere replication (especially of the G-strand) in cells deficient in proteins able to unfold telomeric G4 *in vitro*, in particular the helicases WRN (9), BLM (10,11) and RTEL1 (10,12,13) and the single-stranded DNA binding complexes RPA (30) and CST (14,15). Nevertheless, when cellular machineries are functional, GGGTTA repeats in lagging strand are stable through replication cycles (62). Are telomeric G4 so problematic structures? With the exception of budding yeasts (that have the most variable set of telomeric repeats), telomeric motifs mostly consist of a short sequence of the type $G_{2\text{-}4}T_{1\text{-}4}A_{1,0}$ (2), exhibit a paucity in cytosines and are prone to fold into G4 (7). The propensity of telomeric motifs to fold into G4 and the behavior of the GGGCTA variant motif (unstable and hairpin-prone), taken together, suggest that evolution of telomeric motifs might have been driven by selection of G4-prone motifs and counter-selection of hairpin-prone motifs (hence the paucity of cytosines). If, on one hand, instability might be the reason of a hypothetical counter-selection of hairpin-prone telomeric motifs, on the other hand telomeric G4 display at least two friendly features that make them easily manageable by DNA interacting proteins. *In vitro*, telomeric contiguous G4 do form under nearly physiological salt and temperature conditions (26), but they are not so stable to impair the binding of a single-stranded DNA binding protein such as RPA (31). Furthermore, *in vitro*, RPA alone allows progression of polymerase through a telomeric G4, while helicase activities are needed in the presence of more stable G4 (63,64). Consistently, a telomeric G4 forms promptly but also unfolds rapidly under mechanical forces or in the presence of its complementary strand, compared to other biological relevant G4 in promoters and in replication origins (65). Telomeric G4 present a second friendly feature. The contiguous G4 units in which telomeric sequences can fold have similar stability and do not interact with each other (26). Therefore, proteins dealing with such structures are faced with *identical* and *independent* G4 structural units. This makes the efficiency of binding/unfolding independent of the number of G4 units, *i.e.* independent of the length of the G-strand portion exposed as single-strand (31).

In conclusion, although G4-structuring Short Tandem Repeats appear to have the highest potential to induce persistent polymerase stalling (66) and although G4 appear to be challenging structures in several aspects (67), we promote a second point of view: the peculiar friendly features of telomeric G4, discussed above, make them easily manageable secondary structures, that perhaps contribute to ensuring telomere stability. The GGGCTA telomeric variant motif presented in this study (unstable (24), hairpin-prone and poorly bound by RPA in its hairpin form) acts as an enlightening counterpoint to the GGGTTA telomeric canonical motif (stable (62), G4-prone (26) and efficiently bound by RPA despite its structuring into contiguous G4 (31)). Maybe, telomeric G4 are telomeres' best friends.

DATA AVAILABILITY

The authors confirm that the data supporting the findings of this study are available within the article and its Supplementary Data. Not shown data are available from the corresponding author P.A. on request.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank all the members of the team *Nucleic Acid Structure, Telomeres and Evolution* (present and past members) for enriching discussions. We thank Anjali Sengar for preliminary NMR measurements in the early stage of the project, Pauline Lejault for her support in NMR measurements, Anne De Cian for gel staining, Phong Lan Thao Tran for testing one of our sequences in an original single molecule assay and Alexandre Deville (NMR facility of MNHN). NMR spectra reported in this manuscript were acquired by Alain Blond at the NMR facility of MNHN (Plateau technique de Résonance Magnétique Nucléaire, UMR 7245 CNRS/MNHN, Molécules de Communication et d'Adaptation des Microorganismes, Muséum National d'Histoire Naturelle, Paris).

FUNDING

Muséum National d'Histoire Naturelle (MNHN); Centre National de la Recherche Scientifique (CNRS); Institut National de la Santé et de la Recherche Médicale (INSERM); Ministère de l'Enseignement Supérieur, de la Recherche de de l'Innovation (MESRI) Ph.D. Fellowship (to J.C.). Funding for open access charge: INSERM.
Conflict of interest statement. None declared.

REFERENCES

1. de Lange,T. (2018) Shelterin-mediated telomere protection. *Annu. Rev. Genet.*, **52**, 223–247.
2. Podlevsky,J.D., Bley,C.J., Omana,R.V, Qi,X. and Chen,J.J. (2008) The telomerase database. *Nucleic Acids Res.*, **36**, D339–D343.
3. Griffith,J.D., Comeau,L., Rosenfield,S., Stansel,R.M., Bianchi,A., Moss,H. and de Lange,T. (1999) Mammalian telomeres end in a large duplex loop. *Cell*, **97**, 503–514.
4. Goldbach,R.W., Bollen-de Boer,J.E., van Bruggen,E.F. and Borst,P. (1979) Replication of the linear mitochondrial DNA of Tetrahymena pyriformis. *Biochim. Biophys. Acta*, **562**, 400–417.
5. de Lange,T. (2004) T-loops and the origin of telomeres. *Nat. Rev. Mol. Cell Biol.*, **5**, 323–329.
6. de Lange,T. (2015) A loopy view of telomere evolution. *Front. Genet.*, **6**, 321.
7. Tran,P.L., Mergny,J.L. and Alberti,P. (2011) Stability of telomeric G-quadruplexes. *Nucleic Acids Res.*, **39**, 3282–3294.
8. Bryan,T.M. (2020) G-Quadruplexes at telomeres: Friend or Foe? *Molecules*, **25**, 3686.
9. Crabbe,L., Verdun,R.E., Haggblom,C.I. and Karlseder,J. (2004) Defective telomere lagging strand synthesis in cells lacking WRN helicase activity. *Science*, **306**, 1951–1953.
10. Sfeir,A., Kosiyatrakul,S.T., Hockemeyer,D., MacRae,S.L., Karlseder,J., Schildkraut,C.L. and de Lange,T. (2009) Mammalian telomeres resemble fragile sites and require TRF1 for efficient replication. *Cell*, **138**, 90–103.
11. Zimmermann,M., Kibe,T., Kabir,S. and de Lange,T. (2014) TRF1 negotiates TTAGGG repeat-associated replication problems by recruiting the BLM helicase and the TPP1/POT1 repressor of ATR signaling. *Genes Dev.*, **28**, 2477–2491.
12. Vannier,J.B., Pavicic-Kaltenbrunner,V., Petalcorin,M.I., Ding,H. and Boulton,S.J. (2012) RTEL1 dismantles T loops and counteracts telomeric G4-DNA to maintain telomere integrity. *Cell*, **149**, 795–806.
13. Vannier,J.B., Sandhu,S., Petalcorin,M.I., Wu,X., Nabi,Z., Ding,H. and Boulton,S.J. (2013) RTEL1 is a replisome-associated helicase that promotes telomere and genome-wide replication. *Science*, **342**, 239–242.
14. Huang,C., Dai,X. and Chai,W. (2012) Human Stn1 protects telomere integrity by promoting efficient lagging-strand synthesis at telomeres and mediating C-strand fill-in. *Cell Res.*, **22**, 1681–1695.
15. Zhang,M., Wang,B., Li,T., Liu,R., Xiao,Y., Geng,X., Li,G., Liu,Q., Price,C.M., Liu,Y. *et al.* (2019) Mammalian CST averts replication failure by preventing G-quadruplex accumulation. *Nucleic Acids Res.*, **47**, 5243–5259.
16. Moyzis,R.K., Buckingham,J.M., Cram,L.S., Dani,M., Deaven,L.L., Jones,M.D., Meyne,J., Ratliff,R.L. and Wu,J.R. (1988) A highly conserved repetitive DNA sequence, (TTAGGG)n, present at the telomeres of human chromosomes. *Proc. Natl Acad. Sci. U.S.A.*, **85**, 6622–6626.
17. Allshire,R.C., Dempster,M. and Hastie,N.D. (1989) Human telomeres contain at least three types of G-rich repeat distributed non-randomly. *Nucleic Acids Res.*, **17**, 4611–4627.
18. Baird,D.M., Jeffreys,A.J. and Royle,N.J. (1995) Mechanisms underlying telomere repeat turnover, revealed by hypervariable variant repeat distribution patterns in the human Xp/Yp telomere. *EMBO J.*, **14**, 5433–5443.
19. Coleman,J., Baird,D.M. and Royle,N.J. (1999) The plasticity of human telomeres demonstrated by a hypervariable telomere repeat array that is located on some copies of 16p and 16q. *Hum. Mol. Genet.*, **8**, 1637–1646.
20. Baird,D.M., Coleman,J., Rosser,Z.H. and Royle,N.J. (2000) High levels of sequence polymorphism and linkage disequilibrium at the telomere of 12q: implications for telomere biology and human evolution. *Am. J. Hum. Genet.*, **66**, 235–250.
21. Conomos,D., Stutz,M.D., Hills,M., Neumann,A.A., Bryan,T.M., Reddel,R.R. and Pickett,H.A. (2012) Variant repeats are interspersed throughout the telomeres and recruit nuclear receptors in ALT cells. *J. Cell Biol.*, **199**, 893–906.
22. Lee,M., Hills,M., Conomos,D., Stutz,M.D., Dagg,R.A., Lau,L.M., Reddel,R.R. and Pickett,H.A. (2014) Telomere extension by telomerase and ALT generates variant repeats by mechanistically distinct processes. *Nucleic Acids Res.*, **42**, 1733–1746.
23. Lee,M., Teber,E.T., Holmes,O., Nones,K., Patch,A.M., Dagg,R.A., Lau,L.M.S., Lee,J.H., Napier,C.E., Arthur,J.W. *et al.* (2018) Telomere sequence content can be used to determine ALT activity in tumours. *Nucleic Acids Res.*, **46**, 4903–4918.
24. Mendez-Bermudez,A., Hills,M., Pickett,H.A., Phan,A.T., Mergny,J.L., Riou,J.F. and Royle,N.J. (2009) Human telomeres that contain (CTAGGG)n repeats show replication dependent instability in somatic cells and the male germline. *Nucleic Acids Res.*, **37**, 6225–6238.
25. Lim,K.W., Alberti,P., Guedin,A., Lacroix,L., Riou,J.F., Royle,N.J., Mergny,J.L. and Phan,A.T. (2009) Sequence variant (CTAGGG)n in the human telomere favors a G-quadruplex structure containing a G.C.G.C tetrad. *Nucleic Acids Res.*, **37**, 6239–6248.
26. Bugaut,A. and Alberti,P. (2015) Understanding the stability of DNA G-quadruplex units in long human telomeric strands. *Biochimie*, **113**, 125–133.
27. Chen,R. and Wold,M.S. (2014) Replication protein A: single-stranded DNA's first responder: dynamic DNA-interactions allow replication protein A to direct single-strand DNA intermediates into different pathways for synthesis or repair. *Bioessays*, **36**, 1156–1161.
28. Verdun,R.E. and Karlseder,J. (2006) The DNA damage machinery and homologous recombination pathway act consecutively to protect human telomeres. *Cell*, **127**, 709–720.
29. Moser,B.A., Subramanian,L., Chang,Y.T., Noguchi,C., Noguchi,E. and Nakamura,T.M. (2009) Differential arrival of leading and lagging strand DNA polymerases at fission yeast telomeres. *EMBO J.*, **28**, 810–820.
30. Audry,J., Maestroni,L., Delagoutte,E., Gauthier,T., Nakamura,T.M., Gachet,Y., Saintome,C., Geli,V. and Coulon,S. (2015) RPA prevents G-rich structure formation at lagging-strand telomeres to allow maintenance of chromosome ends. *EMBO J.*, **34**, 1942–1958.
31. Lancrey,A., Safa,L., Chatain,J., Delagoutte,E., Riou,J.F., Alberti,P. and Saintome,C. (2018) The binding efficiency of RPA to telomeric G-strands folded into contiguous G-quadruplexes is independent of the number of G4 units. *Biochimie*, **146**, 68–72.

32. Mergny,J.L., Li,J., Lacroix,L., Amrane,S. and Chaires,J.B. (2005) Thermal difference spectra: a specific signature for nucleic acid structures. *Nucleic Acids Res.*, **33**, e138.
33. Saintome,C., Amrane,S., Mergny,J.L. and Alberti,P. (2016) The exception that confirms the rule: a higher-order telomeric G-quadruplex structure more stable in sodium than in potassium. *Nucleic Acids Res.*, **44**, 2926–2935.
34. Masiero,S., Trotta,R., Pieraccini,S., De Tito,S., Perone,R., Randazzo,A. and Spada,G.P. (2010) A non-empirical chromophoric interpretation of CD spectra of DNA G-quadruplex structures. *Org. Biomol. Chem.*, **8**, 2683–2692.
35. Bhattacharya,P.K., Cha,J. and Barton,J.K. (2002) 1H NMR determination of base-pair lifetimes in oligonucleotides containing single base mismatches. *Nucleic Acids Res.*, **30**, 4740–4750.
36. Safa,L., Delagoutte,E., Petruseva,I., Alberti,P., Lavrik,O., Riou,J.F. and Saintome,C. (2014) Binding polarity of RPA to telomeric sequences and influence of G-quadruplex stability. *Biochimie*, **103**, 80–88.
37. Safa,L., Gueddouda,N.M., Thiebaut,F., Delagoutte,E., Petruseva,I., Lavrik,O., Mendoza,O., Bourdoncle,A., Alberti,P., Riou,J.F. *et al.* (2016) 5′ to 3′ unfolding directionality of DNA secondary structures by replication protein A: G-QUADRUPLEXES AND DUPLEXES. *J. Biol. Chem.*, **291**, 21246–21256.
38. Blackwell,L.J. and Borowiec,J.A. (1994) Human replication protein A binds single-stranded DNA in two distinct complexes. *Mol. Cell. Biol.*, **14**, 3993–4001.
39. Bochkarev,A., Pfuetzner,R.A., Edwards,A.M. and Frappier,L. (1997) Structure of the single-stranded-DNA-binding domain of replication protein A bound to DNA. *Nature*, **385**, 176–181.
40. Lat,P.K. and Sen,D. (2018) (C2G4)n repeat expansion sequences from the C9orf72 gene form an unusual DNA higher-order structure in the pH range of 5–6. *PLoS One*, **13**, e0198418.
41. Lei,M., Podell,E.R. and Cech,T.R. (2004) Structure of human POT1 bound to telomeric single-stranded DNA provides a model for chromosome end-protection. *Nat. Struct. Mol. Biol.*, **11**, 1223–1229.
42. Loayza,D., Parsons,H., Donigian,J., Hoke,K. and de Lange,T. (2004) DNA binding features of human POT1: a nonamer 5′-TAGGGTTAG-3′ minimal binding site, sequence specificity, and internal binding to multimeric sites. *J. Biol. Chem.*, **279**, 13241–13248.
43. Wang,F., Podell,E.R., Zaug,A.J., Yang,Y., Baciu,P., Cech,T.R. and Lei,M. (2007) The POT1-TPP1 telomere complex is a telomerase processivity factor. *Nature*, **445**, 506–510.
44. Xin,H., Liu,D., Wan,M., Safari,A., Kim,H., Sun,W., O'Connor,M.S. and Songyang,Z. (2007) TPP1 is a homologue of ciliate TEBP-beta and interacts with POT1 to recruit telomerase. *Nature*, **445**, 559–562.
45. Liu,D., Safari,A., O'Connor,M.S., Chan,D.W., Laegeler,A., Qin,J. and Songyang,Z. (2004) PTOP interacts with POT1 and regulates its localization to telomeres. *Nat. Cell Biol.*, **6**, 673–680.
46. Ye,J.Z., Hockemeyer,D., Krutchinsky,A.N., Loayza,D., Hooper,S.M., Chait,B.T. and de Lange,T. (2004) POT1-interacting protein PIP1: a telomere length regulator that recruits POT1 to the TIN2/TRF1 complex. *Genes Dev.*, **18**, 1649–1654.
47. Liu,G., Chen,X., Bissler,J.J., Sinden,R.R. and Leffak,M. (2010) Replication-dependent instability at (CTG) x (CAG) repeat hairpins in human cells. *Nat. Chem. Biol.*, **6**, 652–659.
48. Nguyen,H.D., Yadav,T., Giri,S., Saez,B., Graubert,T.A. and Zou,L. (2017) Functions of replication protein A as a sensor of R loops and a regulator of RNaseH1. *Mol. Cell*, **65**, 832–847.
49. Rosche,W.A., Jaworski,A., Kang,S., Kramer,S.F., Larson,J.E., Geidroc,D.P., Wells,R.D. and Sinden,R.R. (1996) Single-stranded DNA-binding protein enhances the stability of CTG triplet repeats in *Escherichia coli*. *J. Bacteriol.*, **178**, 5042–5044.
50. Andreoni,F., Darmon,E., Poon,W.C. and Leach,D.R. (2010) Overexpression of the single-stranded DNA-binding protein (SSB) stabilises CAG*CTG triplet repeats in an orientation dependent manner. *FEBS Lett.*, **584**, 153–158.
51. Panigrahi,G.B., Lau,R., Montgomery,S.E., Leonard,M.R. and Pearson,C.E. (2005) Slipped (CTG)*(CAG) repeats can be correctly repaired, escape repair or undergo error-prone repair. *Nat. Struct. Mol. Biol.*, **12**, 654–662.
52. Usdin,K., House,N.C. and Freudenreich,C.H. (2015) Repeat instability during DNA repair: insights from model systems. *Crit. Rev. Biochem. Mol. Biol.*, **50**, 142–167.
53. Kamath-Loeb,A.S., Loeb,L.A., Johansson,E., Burgers,P.M. and Fry,M. (2001) Interactions between the Werner syndrome helicase and DNA polymerase delta specifically facilitate copying of tetraplex and hairpin structures of the d(CGG)n trinucleotide repeat sequence. *J. Biol. Chem.*, **276**, 16439–16446.
54. Chan,N.L.S., Hou,C., Zhang,T., Yuan,F., Machwe,A., Huang,J., Orren,D.K., Gu,L. and Li,G.-M. (2012) The Werner Syndrome Protein promotes CAG/CTG repeat stability by resolving large (CAG) n /(CTG) n Hairpins. *J. Biol. Chem.*, **287**, 30151–30156.
55. Brosh,R.M. Jr, Orren,D.K., Nehlin,J.O., Ravn,P.H., Kenny,M.K., Machwe,A. and Bohr,V.A. (1999) Functional and physical interaction between WRN helicase and human replication protein A. *J. Biol. Chem.*, **274**, 18341–18350.
56. Kamath-Loeb,A.S., Johansson,E., Burgers,P.M. and Loeb,L.A. (2000) Functional interaction between the Werner Syndrome protein and DNA polymerase delta. *Proc. Natl Acad. Sci. U.S.A.*, **97**, 4603–4608.
57. Szekely,A.M., Chen,Y.H., Zhang,C., Oshima,J. and Weissman,S.M. (2000) Werner protein recruits DNA polymerase delta to the nucleolus. *Proc. Natl Acad. Sci. U.S.A.*, **97**, 11365–11370.
58. Hou,C., Chan,N.L., Gu,L. and Li,G.M. (2009) Incision-dependent and error-free repair of (CAG)(n)/(CTG)(n) hairpins in human cell extracts. *Nat. Struct. Mol. Biol.*, **16**, 869–875.
59. Stewart,J.A., Wang,F., Chaiken,M.F., Kasbek,C., Chastain 2nd,P.D., Wright,W.E. and Price,C.M. (2012) Human CST promotes telomere duplex replication and general replication restart after fork stalling. *EMBO J.*, **31**, 3537–3549.
60. Wang,F., Stewart,J.A., Kasbek,C., Zhao,Y., Wright,W.E. and Price,C.M. (2012) Human CST has independent functions during telomere duplex replication and C-strand fill-in. *Cell Rep.*, **2**, 1096–1103.
61. Lin,C.-Y.G., Näger,A.C., Lunardi,T., Vančevska,A., Lossaint,G. and Lingner,J. (2020) The human telomeric proteome during telomere replication. bioRxiv doi: https://doi.org/10.1101/2020.06.14.150524, 15 June 2020, preprint: not peer reviewed.
62. Damerla,R.R., Knickelbein,K.E., Kepchia,D., Jackson,A., Armitage,B.A., Eckert,K.A. and Opresko,P.L. (2010) Telomeric repeat mutagenicity in human somatic cells is modulated by repeat orientation and G-quadruplex stability. *DNA Repair (Amst.)*, **9**, 1119–1129.
63. Dahan,D., Tsirkas,I., Dovrat,D., Sparks,M.A., Singh,S.P., Galletto,R. and Aharoni,A. (2018) Pif1 is essential for efficient replisome progression through lagging strand G-quadruplex DNA secondary structures. *Nucleic Acids Res.*, **46**, 11847–11857.
64. Sparks,M.A., Singh,S.P., Burgers,P.M. and Galletto,R. (2019) Complementary roles of Pif1 helicase and single stranded DNA binding proteins in stimulating DNA replication through G-quadruplexes. *Nucleic Acids Res.*, **47**, 8595–8605.
65. Tran,P.L.T., Rieu,M., Hodeib,S., Joubert,A., Ouellet,J., Alberti,P., Bugaut,A., Allemand,J.F., Boule,J.B. and Croquette,V. (2021) Folding and persistence times of intramolecular G-quadruplexes transiently embedded in a DNA duplex. *Nucleic Acids Res.*, **49**, 5189–5201.
66. Murat,P., Guilbaud,G. and Sale,J.E. (2020) DNA polymerase stalling at structured DNA constrains the expansion of short tandem repeats. *Genome Biol.*, **21**, 209.
67. Bryan,T.M. (2019) Mechanisms of DNA Replication and Repair: Insights from the Study of G-Quadruplexes. *Molecules*, **24**, 3439.