Parity-dependent hairpin configurations of repetitive DNA sequence promote slippage associated with DNA expansion

Tze-Yun Huang[a,b,1], Chung-ke Chang[c,1], Ya-Fen Kao[a,b], Chih-Hao Chin[a], Cheng-Wei Ni[a], Hao-Yi Hsu[a], Nien-Jen Hu[d], Li-Ching Hsieh[b], Shan-Ho Chou[d], I-Ren Lee[a,2], and Ming-Hon Hou[b,2]

[a]Department of Chemistry, National Taiwan Normal University, Taipei, Taiwan 116; [b]Institute of Genomics and Bioinformatics, National Chung Hsing University, Taichung, Taiwan 402; [c]Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan 115; and [d]Institute of Biochemistry, National Chung-Hsing University, Taichung, Taiwan 402

Edited by Taekjip Ha, Johns Hopkins University, Baltimore, MD, and approved July 31, 2017 (received for review May 26, 2017)

Repetitive DNA sequences are ubiquitous in life, and changes in the number of repeats often have various physiological and pathological implications. DNA repeats are capable of interchanging between different noncanonical and canonical conformations in a dynamic fashion, causing configurational slippage that often leads to repeat expansion associated with neurological diseases. In this report, we used single-molecule spectroscopy together with biophysical analyses to demonstrate the parity-dependent hairpin structural polymorphism of TGGAA repeat DNA. We found that the DNA adopted two configurations depending on the repeat number parity (even or odd). Transitions between these two configurations were also observed for longer repeats. In addition, the ability to modulate this transition was found to be enhanced by divalent ions. Based on the atomic structure, we propose a local seeding model where the kinked GGA motifs in the stem region of TGGAA repeat DNA act as hot spots to facilitate the transition between the two configurations, which may give rise to disease-associated repeat expansion.

DNA tandem repeats | DNA slippage | single-molecule spectroscopy | X-ray crystallography

DNA replication is a crucial process in all living organisms. Mishaps in the replication process generally lead to deleterious consequences but also drive biological evolution (1). Changes in the number of tandem copies of a specific DNA sequence within the genome are associated with devastating neuropathies and various types of cancer (2, 3). On the other hand, these changes also help shape normal genomic features such as microsatellite polymorphism, which are often used as markers for population biology studies (4).

The unit sizes of repetitive DNA sequences involved in repeat number changes range from a single base (e.g., microsatellites) to dodecanucleotides (12 bases, e.g., in progressive myoclonic epilepsy type 1) (5, 6). DNA slippage is believed to be a primary mechanism driving the change in repeat number of various unit sizes. Repetitive DNA sequences often form alternative structures such as bulges and hairpin loops in addition to canonical DNA conformations (7, 8). A repeat unit may slip between being part of a hairpin loop, a bulge, or a duplex in a dynamic fashion, which may alter the course of normal cellular DNA chemistry and ultimately lead to repeat expansion associated with neurological diseases (9). $(TGGAA)_n$ repeats, for example, may form noncanonical structures such as a hairpin arm (10, 11) or an antiparallel duplex (12). Expansion of this pentanucleotide sequence has been associated with spinocerebellar ataxia 31 (SCA31), an adult-onset autosomal-dominant neurodegenerative disorder (13).

In this article, we probed the conformational heterogeneity and stability of hairpins composed of repetitive TGGAA sequences using single-molecule fluorescence resonance energy transfer [single-molecule FRET (smFRET)] spectroscopy and X-ray crystallography as primary tools. Remarkably, we were able to detect two distinct hairpin configurations, with each being dominant under different repeat number parity (even or odd). The ability to convert between the two configurations is dependent on the number of repeats and can be modulated by the presence of divalent ions. Only sequences with large even number of repeats are able to interconvert between the two forms. Based on our structural studies, we propose a local seeding model where the central kinked GGA motifs in the $(TGGAA)_n$ DNA repeat act as hot spots to facilitate the transition between the two parity-dependent configurations. Our findings suggest a mechanism by which a binary dynamic property of DNA repeats may affect repeat expansion and may be applicable to other repetitive DNA systems.

Results

$d(TGGAA)_3$ and $d(TGGAA)_4$ Adopt Distinct Structural Configurations. A scheme of the configurations probed by our smFRET assay is shown in Fig. 1*A*. We found that $d(TGGAA)_3$ folds into a hairpin structure with the two ends of the single-stranded oligonucleotide being brought into close proximity, corroborated by the high E_{FRET} value (~0.8) compared with the value of 0.3 for the single-stranded dT_{15} control, which does not form secondary structures (Fig. 1*B*). The end-to-end alignment was further confirmed by the very similar

Significance

We found that TGGAA DNA repeats, which are involved in the neurological disease spinocerebellar ataxia 31, are capable of assuming two different hairpin structures depending on repeat number parity. We determined the interconversion kinetics by single-molecule spectroscopy and probed the interconversion mechanism through elucidation of the TGGAA repeat stem structure. Our results suggest that the two hairpin structures interconvert through motion slippage, and the process can be explained by the overall stem stability and local destabilization of the kinked GGA motif. Divalent cations and stem length affected the equilibrium and kinetics of slippage. Our findings suggest a mechanism by which a binary dynamic property of DNA repeats may affect repeat expansion and may be applicable to other repetitive DNA systems.

Author contributions: I-R.L. and M.-H.H. designed research; T.-Y.H., Y.-F.K., C.-H.C., C.-W.N., and H.-Y.H. performed research; T.-Y.H., C.-k.C., Y.-F.K., C.-H.C., C.-W.N., H.-Y.H., N.-J.H., L.-C.H., S.-H.C., I-R.L., and M.-H.H. analyzed data; and C.-k.C., I-R.L., and M.-H.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: Crystallography, atomic coordinates, and structure factors have been deposited in Protein Data Bank (accession no. 5GUN).

[1]T.-Y.H. and C.-k.C. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: irenlee@ntnu.edu.tw or mhho@dragon.nchu.edu.tw.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1708691114/-/DCSupplemental.

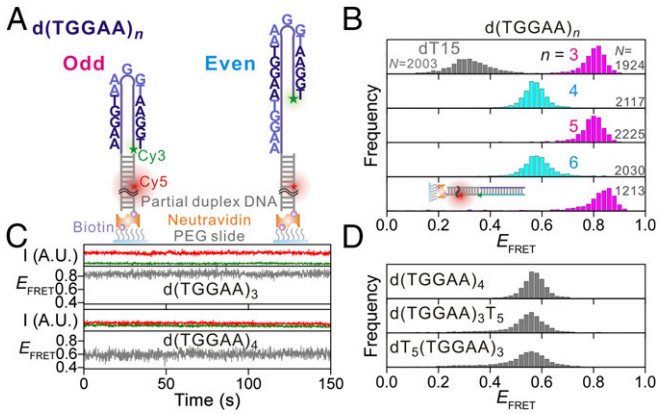

Fig. 1. Structural characterization of d(TGGAA)$_n$ using single-molecule FRET. (*A*) Illustrations of the single-molecule assay used in this experiment. The green and red stars represent the donor (Cy3) and acceptor (Cy5) labeling sites, respectively. (*B*) E_{FRET} histograms of d(TGGAA)$_{3-6}$ (colored) and the assay used as a caliper of the end-to-end alignment (cartoon in *Bottom*). The E_{FRET} histogram of secondary structure-free dT$_{15}$ is overlaid as a gray histogram for comparison. (*C*) Representative E_{FRET} time traces for d(TGGAA)$_3$ and d(TGGAA)$_4$. The green and red curves are intensities of the Cy3 and Cy5 fluorescence signals, respectively. The gray curves represent the E_{FRET} (unitless). The vast majority of the traces, including those not shown here, are virtually stable at a single E_{FRET} state. I, intensity; A.U., arbitrary unit. (*D*) E_{FRET} histograms of d(TGGAA)$_4$ and mutation assays for simulating the hairpin with overhang structures. *N* denotes the number of molecules used for building the histogram.

E_{FRET} observed when a complementary strand labeled with Cy3 at the 3′ end was annealed to our construct (Fig. 1*B*, *Bottom*). At room temperature, the time-dependent E_{FRET} traces of d(TGGAA)$_3$ remained static over a 2-min observation window, suggesting that the end-to-end hairpin conformation was very stable and without distinguishable conformational isomers (Fig. 1*C*).

However, we observed a substantial drop in the E_{FRET} value (~0.8 to ~0.6; Fig. 1*B*) for the even-numbered repeat d(TGGAA)$_4$ compared with that of odd-numbered d(TGGAA)$_3$, indicating the presence of an offset between the two termini of d(TGGAA)$_4$. The E_{FRET} value was still significantly higher than that of dT$_{15}$ (Fig. 1*B*), suggesting that the stem region where base pairing occurs was still present in d(TGGAA)$_4$. Similar to d(TGGAA)$_3$, the time-dependent E_{FRET} traces remained constant for d(TGGAA)$_4$ (Fig. 1*C*), suggesting that this offset configuration was also stable at room temperature. To control for the possibility of G-quadruplex formation, we conducted the same experiments in buffer solutions containing 150 mM potassium (which favors G-quadruplex formation), sodium, or lithium (which inhibits G-quadruplex formation) cations and observed similar behavior (Fig. S1), suggesting that G-quadruplex formation is not likely to happen under our experimental conditions. We also generated two mutants, dT$_5$(TGGAA)$_3$ and d(TGGAA)$_3$T$_5$, where the 5′- and 3′-terminal dTGGAA were changed to dTTTTT, respectively, and used these mutants as calipers to measure the offset in d(TGGAA)$_4$. The good agreement between the E_{FRET} values of d(TGGAA)$_4$ and those of the two mutants suggests that the terminal dTGGAA adopts the random coil state of dTTTTT, indicating that the offset is caused by the presence of a single dTGGAA overhang (Fig. 1*D*).

TGGAA Repeat Number Parity Determines the Preferred Configuration. The observation of two distinct E_{FRET} values in the d(TGGAA)$_3$ and d(TGGAA)$_4$ repeats led us to undertake a systematic study of oligonucleotides containing different numbers of TGGAA repeats. Remarkably, we found that the E_{FRET} oscillated between ~0.8 and ~0.6 depending on the repeat number parity (Fig. 1*B*). Lower E_{FRET} values corresponding to the overhang configuration were exclusively observed for oligonucleotides with even repeat numbers.

We have compared the binding affinity of dTGGAA to another dTGGAA versus to the fully complementary sequence dTTCCA by surface plasmon resonance (SPR) (Fig. 2 *A* and *B*). The higher association and lower dissociation rates of the fully complementary duplex formation compared with those of dTGGAA repeat duplex formation, suggesting that the dTGGAA repeat duplex is less stable than the fully complementary duplex. This allowed the determination of the relative stabilities of the hairpin structures in relation to the number of repeats with a kinetic competition assay using complementary oligonucleotides as competitors (Fig. 2*C*). We recorded the time-dependent E_{FRET} histograms (Fig. 2*D*) of oligonucleotides containing different number of repeats and converted these to population fractions (Fig. 2*E*), which were then used to extract the kinetic rates (Fig. 2*F*). Hairpins that consisted of even number of repeats were more prone to melt and form a duplex in our assay (Fig. 2*F*, *Inset*), which suggests that the energy barrier between the canonical duplex and the overhang configuration is lower than that between the duplex and the end-to-end configuration.

dTGGAA Repeats Form a Kinked Antiparallel Duplex. To provide insight into the hairpin formation of TGGAA repeat at the stem region, we solved the crystal structure of dG(TGGAA)$_2$C at a resolution of 2.58 Å by multiple-wavelength anomalous diffraction (MAD) using a brominated oligonucleotide (Table S1). The rmsd between the crystal and NMR sructures is 1.98 Å, indicating that they are overall similar. The oligonucleotide self-assembles into an antiparallel duplex corresponding to the stem region of the hairpin conformations of dTGGAA repeats. Each duplex contains two zipper cores formed by the two d(TGGAA)$_2$

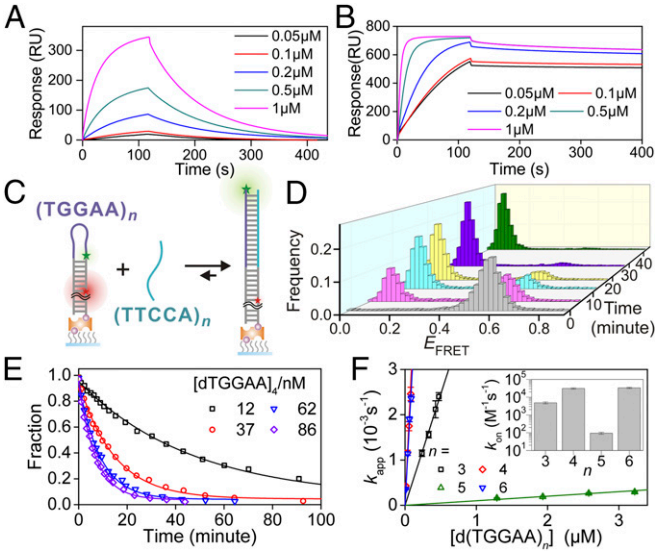

Fig. 2. Kinetic competition experiment using complementary strands. (*A*) SPR sensorgram of DNA–DNA interaction between immobilized dG(TGGAA)$_2$G and the ligand dC(TGGAA)$_2$C at various ligand concentrations. (*B*) Same as *A* but with dC(TTCCA)$_2$C as the ligand instead. (*C*) Illustration of the competition assay. Excess amount of d(TTCCA)$_n$ oligonucleotides was added to the d(TGGAA)$_n$ single molecule assay. Opening of the looped structure and formation of duplex DNA in the product state increases the distance between the two labeling sites (marked as red and green stars), resulting in low E_{FRET}. (*D*) Time-dependent E_{FRET} histograms of d(TGGAA)$_4$ + d(TTCCA)$_4$. (*E*) Representative time evolutions of the fraction of folded d(TGGAA)$_4$ at various concentrations. The fraction was obtained using the population fraction with E_{FRET} greater than 0.27. Single-exponential fits are overlaid as solid lines. (*F*) Concentration dependence of the apparent kinetic rate constants obtained for different number of repeats. Linear fits are overlaid as solid lines. The forward rate constants, obtained from the slopes of the linear fits, are shown in the inset. Error bars represent SDs of three individual experiments.

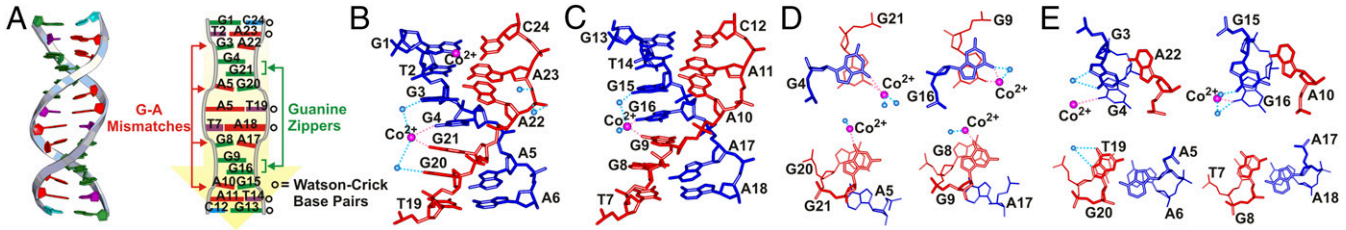

Fig. 3. Structure of the dG(TGGAA)$_2$C duplex. (*A*) Ribbon representation of the duplex structure (*Left*) and a schematic of important interactions between the two strands (*Right*). The two strands are colored in gray (G1 to C12, from top to bottom) and yellow (G13 to C24, from bottom to top). Guanines are in green, thymines are in purple, cytosines are in cyan, and adenines are in red. Close-up view of the molecular structure of the dG(TGGAA)$_2$C at the (*B*) G1 to A6 and (*C*) G13 to A18 terminal base pair steps. Water molecules and Co(II) ions are shown as light blue and pink balls, respectively. The dashed line represents direct hydrogen bonds and coordinated bonds. (*D*) Structural detail of the Co(II) (or water)–DNA interactions. (*E*) Structural detail of the interactions involving hydrogen bonds of mismatched G:A pairs with and without water stabilization.

motifs from each strand. The central region of each zipper core comprises a double-stranded intercalated motif in which the two strands of the duplex intersect and are held together by hydrogen bonds from G:A homopurine base pairs and two intercalated G bases (Fig. 3*A*). An unpaired guanosine base in each dTGGAA repeat from one strand intercalates with another unpaired guanosine base from the opposite strand to form a guanine zipper, which in turn stacks with the guanosine bases of the flanking sheared G:A base pairs on both sides to form a stable continuous G4 stack (Fig. 3 *A*–*C*). The exocyclic NH_2 at C2 of the unpaired guanine and the cross-strand backbone phosphate oxygen atoms form unique hydrogen bonds that may stabilize the cross-strand stacking between the two unpaired guanines in the $(GGA)_2$ motif. Some water molecules not observed in the NMR structure were also found to form hydrogen bonds with the guanine O6 and N1 atoms, which may further stabilize the overall structure (Fig. 3 *D* and *E*) (14). These interactions at the stem may all assist in stabilizing the hairpin structure. For duplexes with short repeats, this stabilization would be more modest than that of a fully cDNA duplex, although longer repeats would provide stronger stabilization (Fig. S2 *A* and *B*). The δ torsion angles of most residues are in the *trans* (*t*) conformation, with the exception of the unpaired guanosine residues, which are closer to the g^- domain. The two strands in the intercalated stem of the DNA duplex (Fig. 3*A*) twist in a clockwise direction, with the major and minor grooves retaining a right-handed B-helical structure with a narrow minor groove in the GGA region (Fig. 3 *B* and *C*). The majority of sugar puckers preserve the C2′-*endo* or closely related C3′-*exo* structures, except for the two contiguous unpaired guanines, which adopt various conformations including O1′-*endo*, C1′-*exo*, and C4′-*exo*. In addition, the β torsion angles of the adenine residues in the sheared G:A pairs adopt the g^- conformation, which differs from the *trans* conformation usually observed for A-DNA and B-DNA.

The roll angles between the sheared G3:A22 and G15:A10 base pairs are negative in the crystal structure (Fig. 4*A*). This results in a sharp kinking of the DNA helix toward the major groove which was not reported in the NMR structure. The sharp kink may act as a hot spot to destabilize the duplex and enable formation of alternative DNA structures (Fig. 4*B*). In addition, the helical twist angles between the A:T and G:A base pair steps have an average of 50.5° and result in a locally overwound DNA conformation (Fig. S3*A*). The average twist angle between the A:G base pair steps is back to 35°, which is similar to that of the B-form DNA. The two sheared G:A pairs flanking each zipper core are nonplanar with asymmetric propeller twist angles and form N2–H…N3, and N2–H…N3 hydrogen bonds (Fig. S3*B*). These two sheared G:A pairs exhibit high buckle angles (~30°) and stretch (~4 Å) and shear (~6 Å) distances (Fig. S3 *C*–*E*). The average stacking gap between the A:G pairs separated by two intercalated and unpaired G bases is 11 Å (Fig. S3*F*). The stacking gap between the A:T pair and the G:A pair is 3 Å.

Conformational Slippage Occurs in Longer TGGAA DNA Repeats. The kinked structure of the stem regions of dTGGAA repeat hairpins suggest a means by which the hairpin may be able to slip between different conformations. Indeed, transient spikes that change the E_{FRET} from ~0.6 to ~0.8 were observed for the smFRET traces of d(TGGAA)$_6$ and d(TGGAA)$_8$ (Fig. 5*A*). In contrast, d(TGGAA)$_5$ and d(TGGAA)$_7$ did not exhibit such transitions at all (Fig. S4). A closer look at the FRET histogram of d(TGGAA)$_6$ and d(TGGAA)$_8$ revealed a small shoulder at E_{FRET} ~0.8, which matches the end-to-end hairpin configuration (Fig. 5*B*). The forward (overhang to end-to-end) and backward (vice versa) rate constants of the slippage transitions of d(TGGAA)$_8$ were ~20% ($0.04 \pm 0.02\ s^{-1}$ versus $0.05 \pm 0.01\ s^{-1}$) and ~35% ($0.9 \pm 0.1\ s^{-1}$ versus $1.4 \pm 0.2\ s^{-1}$) slower than those of d(TGGAA)$_6$, respectively, resulting in a slight increase in the fraction of the end-to-end configuration in the longer repeat. This trend continued for d(TGGAA)$_{10}$ (Fig. S5). Because increasing the number of repeats also allowed the formation of longer and hence more stable stem regions, we suggest that stabilization of the stem region may play a key role in allowing the formation of the end-to-end configuration.

We also performed the experiments in buffer containing different Mg^{2+} concentrations to examine the effect of ions on the two configurations in d(TGGAA)$_6$ and d(TGGAA)$_8$. A significant increase in the fraction of the end-to-end configuration was observed with increasing Mg^{2+} concentration (Fig. 5*C*), although the overhang configuration was still favored (Fig. 5*D*). A slight shift toward high E_{FRET} was also observed and is likely due to an apparent charge screening effect (15). In the crystal structure of the dG(TGGAA)$_2$C duplex, a number of Co^{2+} and water molecules were coordinated to the DNA duplex (Fig. 3 *B*–*E* and Table S2) and help stabilize the structure of the stem region. One Co^{2+} ion was coordinated to the O6 atom of G13 (Fig. 3*C*),

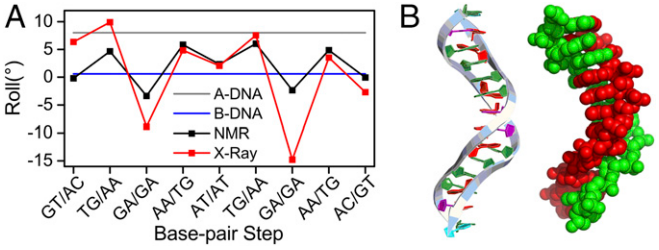

Fig. 4. Kinked structure of the dG(TGGAA)$_2$C duplex. (*A*) Roll angles between base pair steps of the duplex. Red and black lines represent values for the crystal and NMR structures[Protein Data Bank (PDB) ID: 103D], respectively. Typical values for A-DNA (gray line) and B-DNA (blue line) are also shown. (*B*) Side view of the dG(TGGAA)$_2$C crystal structure in ribbon (*Left*) and space-filled (*Right*) representations.

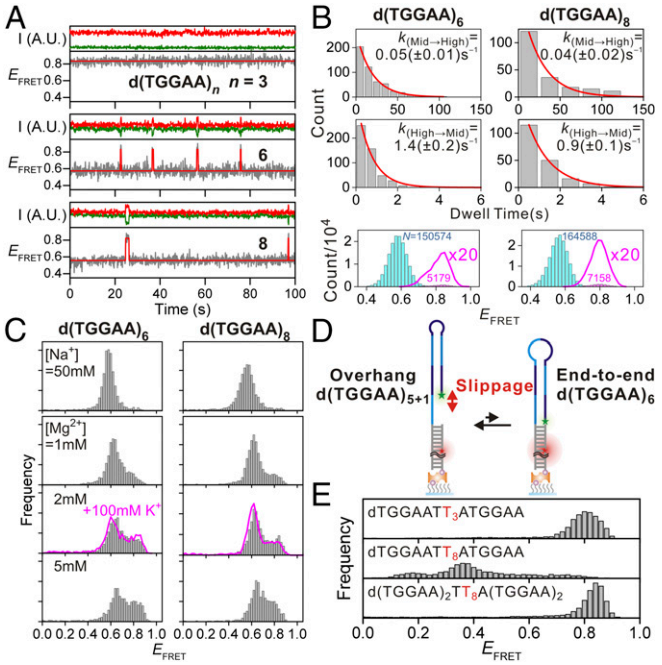

Fig. 5. Structural dynamics of d(TGGAA)$_6$. (*A*) Representative E_{FRET} time traces (lower panel in each set of panels) in 50 mM Na^+ and the corresponding donor (green) and acceptor (red) signal intensities (upper panel in each set of panels) for d(TGGAA)$_{3,6,8}$. Except for the virtually static trace of d(TGGAA)$_3$, multiple sudden increases in E_{FRET} can be clearly observed. The time traces were fit to a hidden Markov algorithm assuming a two-state model and are overlaid in red. I, intensity; A.U., arbitrary unit. More than 85% (n = 6) and 55% (n = 8) of valid traces longer than 30 s showed similar transitions, whereas the rest remained steady. (*B*) Dwell time histograms of the low (E_{FRET} = 0.6; *Left*) and the high (E_{FRET} = 0.8; *Right*) E_{FRET} states. Forward and backward kinetic rate constants of the slippage process were obtained by fitting the histograms to a single-exponential decay model (red lines). E_{FRET} population histograms of d(TGGAA)$_{6,8}$ are shown in *Bottom*. Magnified views (20×) of the high E_{FRET} population are shown as red lines. The numbers (N) indicate frame counts of each state, which are color coded. Data were obtained by analyzing 208 and 257 molecules for n = 6 and n = 8, respectively. (*C*) E_{FRET} histogram of d(TGGAA)$_{6,8}$ under various salt conditions. The fractions of high E_{FRET} (>0.8) increase as Mg^{2+} concentrations increase. The overlaid red line was obtained in the presence of [K^+] = 100 mM and [Mg^{2+}] = 2 mM to mimic physiological conditions. No obvious differences were observed compared with the data obtained in the presence of 2mM Mg^{2+} alone. (*D*) Schematic of d(TGGAA)$_{6,8}$ hairpin slippage. The transient slippage from the overhang to the end-to-end configuration may originate from the formation of an octanucleotide loop which shortens the distance between the dye pair (marked as green and red stars) and leads to a transient increase in E_{FRET}. (*E*) Effect of loop formation on hairpin stability. (*Top*) Trinucleotide mutation (GGA→T$_3$) in the loop region of d(TGGAA)$_3$ does not alter hairpin formation. (*Middle*) Octa-nucleotide mutation (GGAATGGA→T$_8$) in the loop region of d(TGGAA)$_4$ prevents hairpin formation as evidenced by the broad and downshifted E_{FRET} distribution. (*Bottom*) Octanucleotide mutation (GGAATGGA→T$_8$) in the loop region of d(TGGAA)$_6$ retains stable hairpin formation.

whereas two additional Co^{2+} ions were *bis*-coordinated to the O6 oxygen atoms of two consecutive unpaired guanines with an incomplete hydration shell (Fig. 3 *B*–*D*). These interactions were not observed in the previous NMR structure. Our results indicate divalent ions are an integral part of the stem and may increase the population of the end-to-end configuration by enhancing stem region stability.

Long TGGAA DNA with Even Number of Repeats Form Transient Octaloop Structures During Slippage. Unlike in the longer d(TGGAA)$_n$ with even number of repeats, the end-to-end configuration was not observed in the shorter d(TGGAA)$_4$, indicating that the end-to-end configuration that is potentially formed by an antiparallel dTGGAA/dTGGAA stem, a T:A base pair, and a dGGAATGGA octanucleotide loop is inherently unstable. To examine this hypothesis, we conducted a series of experiments using d(TGGAA)$_n$ oligonucleotides containing poly-dT mutations in the loop region (Fig. 5 *D* and *E*). We first tested the effect of the size of the loop on the stability of the end-to-end configuration using dTGGAAT$\underline{T_3}$ATGGAA and dTGGAAT$\underline{T_8}$ATGGAA, which have the same length as d(TGGAA)$_3$ and d(TGGAA)$_4$, respectively. Both have the potential to form end-to-end hairpin configurations held together by an identical antiparallel dTGGAA/dTGGAA stem and a T:A base pair but with different loop sizes: 3 nt for dTGGAAT$\underline{T_3}$ATGGAA and 8 nt for dTGGAAT$\underline{T_8}$ATGGAA. Our results show that dTGGAAT$\underline{T_3}$ATGGAA did form a stable hairpin configuration, evident by the sharp E_{FRET} distribution at ∼0.8, whereas dTGGAAT$\underline{T_8}$ATGGAA gave a broad and downshifted E_{FRET} distribution indicative of conformational heterogeneity containing folded and unfolded species. This finding suggests that the octanucleotide loop is less stable than the trinucleotide loop and cannot form a stable state when the stem region contains a single dTGGAA/dTGGAA pair. However, an extended version of the octanucleotide loop oligonucleotide, d(TGGAA)$_2$TT$\underline{T_8}$A(TGGAA)$_2$, which has the same length as d(TGGAA)$_6$ and contains a total of four repeat units in the duplex stem region, was able to reform a stable end-to-end hairpin configuration, suggesting that the extended stem composed of antiparallel d(TGGAA)$_2$/d(TGGAA)$_2$ stabilizes the octaloop hairpin and allows a stationary state to be formed. Long stem regions may be required for stabilization of the octaloop in the end-to-end configurations of the even-numbered repeat sequences, which may explain why slippage motions are only observed for d(TGGAA)$_6$ and longer even-numbered repeat sequences.

Discussion

Many repetitive DNA sequences including trinucleotide, tetranucleotide, and pentanucleotide repeat adopt a variety of noncanonical structures such as hairpin loops and quadruplexes in either the single-stranded or double-stranded state in the cell, which may predispose these repeats to expand (5). For example, single-stranded (CNG)$_n$ repeats are able to form hairpin DNA structures that consist of both Watson–Crick base pairs and mismatched base pairs (16, 17). Individual strands of (CCTG)$_n$/(CAGG)$_n$ repeats have also been shown to fold into hairpin-like structures with zipper-like composition (18). Transient intrastrand hairpins containing noncanonical structures have also been proposed to promote DNA slippage and are causative factors for DNA expansion (19). We have chosen the pentanucleotide TGGAA repeat, which carries guanine mismatches and zipper-like interactions, as an ideal model system to study the structural and conformational dynamics of repetitive DNA sequences. Owing to its variety of structural interactions, the information garnered from TGGAA DNA repeats may be potentially applied to other DNA repeats containing different sequences. Because dTGGAA repeat expansion is associated with the neurodegenerative disease SCA31 (13), understanding its structural and chemical bases of configuration slippage may also be important in a physiological context.

It has been shown by NMR that dTGGAA tends to form a hairpin with a single G residue in the loop closed by a sheared G:A mismatch (11, 20). Nucleic acids that contain 5′-GGA/AGG-5′ or 5′-GAAA/AAAG-5′ motifs can form antiparallel duplexes stabilized by unpaired purine bases that extend their stacking interactions until reaching the sheared G:A base pair. The continuous stacking interactions in the purine sequences are the major forces responsible for the stabilization of these DNA conformations. By combining this information with our high-resolution structure of the dTGGAA repeat duplex, the structural basis of the molecular behavior observed in the smFRET experiments becomes clear. For sequences containing an odd number of dTGGAA repeats, the

end-to-end configuration is optimal because it forms the highest number of duplex interactions while maintaining the 1-nt loop structure proposed by Zhu et al. (11). In contrast, sequences containing an even number of dTGGAA repeats can either maintain the 1-nt loop structure and leave a dangling overhang or maintain the end-to-end configuration and form octaloop structures (Fig. 5*D*). For longer even-numbered repeats, the destabilizing effect of the octaloop would be compensated by the stabilizing effect of the long flanking duplexes. The loss of entropic energy from the overhang region may be compensated by the increase in enthalpic energy of the end-to-end configuration duplex interactions. One would thus expect that longer repeats would facilitate the transition to the end-to-end configuration because of the higher number of duplex interactions available. Indeed, we did observe an increase in the fraction of the high E_{FRET} population with increasing even-numbered repeat lengths (Fig. S5). The effect of divalent cations on the population fraction of the end-to-end configuration in $d(TGGAA)_6$ and $d(TGGAA)_8$ further highlights the importance of stem stability in the slippage process. Presence of divalent cations enhances the overall stability of the hairpin by forming a network of interactions with the dTGGAA stem duplex, including an unusual *bis* coordination to the O6 of two consecutive unpaired guanosines (Fig. 3*D*). Similar binding modes between metal ions and bases have been observed for Pt and cytosine cross-linked interaction in cisplatin–DNA complex structure (21). This implies that it could be possible to control the slippage process of dTGGAA repeats by metal ions in a manner similar to that proposed for controlling ribozyme activity (22).

Given that both the end-to-end and overhang configurations are accessible by the DNA, the question of how they interconvert remains. One potential model for the conversion is through the complete opening and closing of the hairpin loop. This is unlikely because the stability of the hairpin formed by dTGGAA repeats is very high (Fig. S2*B*), and the fully open conformation would be inaccessible in a closed system at room temperature. A more viable conversion model is one where the antiparallel dTGGAA repeat duplex sites in the stem region of the hairpin serve as hot spots for duplex melting due to the lower stability of the $(GGA)_2$ motifs and sharp bent conformation of the TGGAA repeats. It has been reported that DNA bending may promote DNA melting and further enable formation of alternative DNA structures (23). This would allow the dissociation of the different strands in the stem region to be initiated at multiple sites by thermal fluctuation and thus lower the energy requirement for duplex melting. We envision that multiple rounds of local unwinding at the hot spot regions followed by rearrangement of the repeat registers, e.g., rearrangement of the duplex between repeats 1 and 5 to a duplex between repeats 2 and 5 with repeat 1 bulging out, to eventually lead to conversion to the alternative conformation. The bulges may propagate during the next few rounds of local unwinding toward the final location of the loop. This local unwinding would defy observation by bulk experiments such as spectrophotometry because most repeats in the sample would remain in the duplex conformation, thus resulting in almost no change to the observable spectrum at a given temperature. Of course, the larger the number of repeats forming the stem duplex, the higher the energy cost to slip the two strands of the stem against each other, and the longer one would have to wait for a slippage event to occur. The retarded forward and backward kinetics of configuration slippage in $d(TGGAA)_8$ compared with $d(TGGAA)_6$ (Fig. 5*A*) may reflect this fact.

Finally, we propose a consecutive expansion model for dTGGAA tandem repeats involving DNA slippage, illustrated in Fig. 6. Odd-numbered (n) repeats of dTGGAA form end-to-end aligned hairpin structures that have the potential to induce repeat expansion during DNA replication, recombination, or repair (24). The expanded even-numbered $n + 1$ repeat product forms an overhang-containing hairpin, which can be converted back to a normal DNA duplex in the presence of its complementary strand (Fig. 2). This

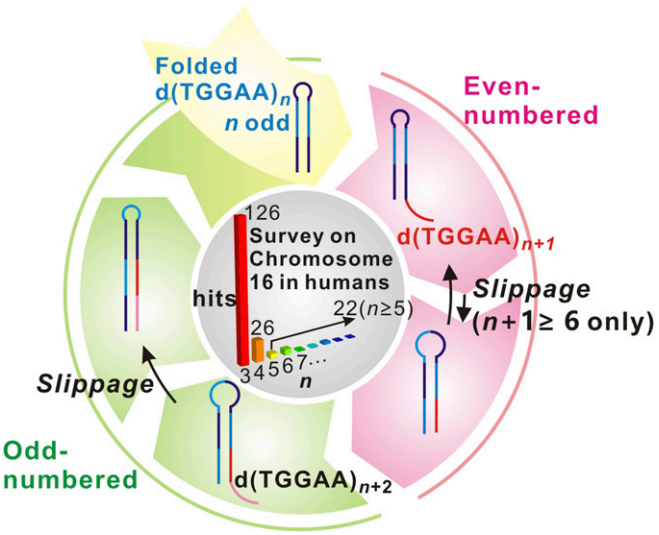

Fig. 6. Proposed model for consecutive expansion of DNA repeats. From top and clockwise, initially odd-numbered $d(TGGAA)_n$ DNA folds into a hairpin structure, which has a high repeat expansion propensity. Expansion of one pentanucleotide unit dTGGAA results in a hairpin with an overhang product $d(TGGAA)_{n+1}$. This product can undergo slippage motion and convert to an end-to-end aligned hairpin with an octanucleotide loop only if the repeat number $n + 1$ is greater than or equal to 6. Further expansion of the end-to-end aligned $d(TGGAA)_{n+1}$ results in an odd-numbered $d(TGGAA)_{n+2}$ product which has forms a hairpin structure again. The cycle can be perpetuated to yield long repeat expansions. A bioinformatics survey of the prevalence of each repeat number (n) in human chromosome 16 is shown in the center. The number of occurrences is shown at the top of the bars.

would temporarily stall the expansion process. However, for $n + 1 > 4$, configurational slippage would allow the $n + 1$ product to transiently form end-to-end hairpins capable of initiating a second round of repeat expansion. The odd-numbered $n + 2$ repeat product would revert back to the energy-favorable hairpin configuration and perpetuate the cycle. The only exception is $d(TGGAA)_4$, or $n + 1 = 4$, which cannot adopt an end-to-end configuration and may completely halt the expansion. Our bioinformatics analysis of the human chromosome 16, as shown in the center circle of Fig. 6, not only shows the expected high abundance of $n = 3$ repeats but also shows a sharp inflection in abundance between repeat numbers $n = 4$ and $n = 5$. We surmise that sequences with $n = 3$ and $n = 4$ are considered safe since they do not undergo slippage and have a large enough margin for error even if other mechanisms result in abnormal expansion. In contrast, while $n = 5$ also prevents slippage, it provides a narrow error margin because other expansion mechanisms may result in $n = 6$ repeats which would increase the probability of slippage-induced expansion. This finding implies that nature has devised a way to use the conformational dynamics properties of $d(TGGAA)_4$ to act as a checkpoint against DNA slippage-induced repeat expansion.

In conclusion, we have demonstrated that the repetitive DNA sequence dTGGAA is capable of assuming two interconvertible parity-dependent hairpin configurations. We propose a consecutive expansion model as a possible molecular mechanism for TGGAA repeat expansion in diseases such as SCA31. In addition to the biological implications, the divalent cation dependency of the transition suggests a way to control the phenomenon, which may have potential applications in the field of DNA-based sensor development and nanotechnology (25).

Materials and Methods

General materials and methods used in this work are described in *Supporting Information*. Oligonucleotide sequences are listed in Table S3.

Single-Molecule FRET Experiments. Detailed descriptions of the experimental setup are available in *Supporting Information*. The single-molecule FRET apparatus was built following the guidelines in Roy et al. (26). An oligonucleotide labeled with Cy5 dye was tethered on the fluidic chamber surface. The oligonucleotide contained a handle region which was used to anneal a second oligonucleotide containing the complementary handle sequence, followed by the sequence of interest and a Cy3 dye label at the 5′ end. The end-to-end distance of the structure of a DNA sequence of interest can be measured as a function of FRET efficiency, E_{FRET}, given in the following equation (27) and shown in Fig. 1*A*:

$$E_{FRET} = \frac{I_A}{I_D + I_A} = \frac{1}{1 + \left(\frac{R}{R_0}\right)^6},$$

where I_D and I_A represent the fluorescence intensities of Cy3 and Cy5, respectively. R represents the distance between the two fluorophores, and R_0 is the Förster distance of the donor–acceptor pair.

Competition assays using complementary strand hybridization (Fig. 2*C*) were carried out with static smFRET experiments. Assuming a pseudo–first-order reaction,

$$\mathrm{d(TGGAA)}_n + \mathrm{d(TTCCA)}_n \underset{k_{off}}{\overset{k_{on}}{\rightleftarrows}} ds\mathrm{DNA},$$

the decrease in the hairpin fraction as a function of time (Fig. 2 *D* and *E*) can be fit to a single-exponential decay with an apparent decay rate k_{app} described by the following formula:

$$k_{app} = k_{on}\left[\mathrm{d(TTCCA)}_n\right] - k_{off}[ds\mathrm{DNA}].$$

The association rate constant k_{on} can then be extracted from the slope of the linear fit to the plot of k_{app} against complementary oligonucleotide concentrations (Fig. 2*F*).

Single-molecule time traces extracted from dynamic smFRET experiments were analyzed using HaMMy software package, which employs a hidden Markov algorithm for state identification and dwell time analysis (26). The traces were manually screened to remove the interference of unwanted effects such as early photobleaching or dye molecule instability before the analysis.

Crystallography. Crystals of $dG(TGGAA)_2C$ were obtained from a solution of 1.5 mM single-stranded DNA, 4 mM $CoCl_2$, 50 mM sodium cacodylate buffer (pH 7.0), 1 mM spermine, and 3% 2-methyl-2,4-pentanediol (MPD) at 4 °C using the sitting-drop vapor diffusion method. The 5′ guanine and 3′ cytosine were included to push the equilibrium toward the self-assembly of the antiparallel duplex through the complementary base pairs on the two termini, which favors crystallization. Cylinder-shaped crystals of $d[G(TGGAA)_2C]$ appeared after 2 wk. Diffraction data were collected at 100 K using an Advanced Detector Systems Corp. (ADSC) Q315r detector at beamline 13B1 of the National Synchrotron Radiation Research Center (Taiwan). The software package HKL2000 was used to index, integrate, and scale the X-ray diffraction data (28). Multiple-wavelength anomalous diffraction (MAD) data were collected from three wavelengths using a brominated oligonucleotide. The phase was solved with SHELX C/D/E program in the Collaborative Crystallography Project Number 4 Graphical User Interface (CCP4i). The resulting well-defined MAD electron density maps at 2.58-Å resolution were used to build the initial models using MIFit (github.com/mifit/) and structure refinement carried out in Refmac5 (29) using the DNA force field parameters reported by Parkinson et al. (30). Each asymmetric unit contained three similar duplexes with twofold symmetry (Fig. S6). Two different types of contacts are present between the three DNA duplexes in each asymmetric unit: end-to-end and side-to-side interactions, mediated by π–π stacking and hydrogen contacts, respectively (Fig. S6*A*). Torsion angles were calculated using Curves v5.3 software (31, 32) and w3DNA (33) web server. Crystallographic data are summarized in Table S1.

ACKNOWLEDGMENTS. We thank Mr. Roshan Satange for the structural refinements. We also thank the National Synchrotron Radiation Research Center (Taiwan) staff for the data collection. This work was supported by Grants 106-2628-M-005-001-MY3 (to M.-H.H.) and 105-2113-M-003-009-MY2 (to I-R.L.) from the Ministry of Science and Technology, Taiwan.

1. Carvalho CMB, Lupski JR (2016) Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet* 17:224–238.
2. Gacy AM, Goellner G, Juranić N, Macura S, McMurray CT (1995) Trinucleotide repeats that expand in human disease form hairpin structures in vitro. *Cell* 81:533–540.
3. Kim T-M, Park PJ (2014) A genome-wide view of microsatellite instability: Old stories of cancer mutations revisited with new sequencing technologies. *Cancer Res* 74:6377–6382.
4. Leclercq S, Rivals E, Jarne P (2010) DNA slippage occurs at microsatellite loci without minimal threshold length in humans: A comparative genomic approach. *Genome Biol Evol* 2:325–335.
5. Ou C-Y, et al. (1999) Analysis of microsatellite instability in cervical cancer. *Int J Gynecol Cancer* 9:67–71.
6. Mirkin SM (2007) Expandable DNA repeats and human disease. *Nature* 447:932–940.
7. Völker J, Gindikin V, Klump HH, Plum GE, Breslauer KJ (2012) Energy landscapes of dynamic ensembles of rolling triplet repeat bulge loops: Implications for DNA expansion associated with disease states. *J Am Chem Soc* 134:6033–6044.
8. Chen Y-W, Jhan C-R, Neidle S, Hou M-H (2014) Structural basis for the identification of an i-motif tetraplex core with a parallel-duplex junction as a structural motif in CCG triplet repeats. *Angew Chem Int Ed Engl* 53:10682–10686.
9. López Castel A, Cleary JD, Pearson CE (2010) Repeat instability as the basis for human diseases and as a potential target for therapy. *Nat Rev Mol Cell Biol* 11:165–170.
10. Grady DL, et al. (1992) Highly conserved repetitive DNA sequences are present at human centromeres. *Proc Natl Acad Sci USA* 89:1695–1699.
11. Zhu L, Chou SH, Reid BR (1996) A single G-to-C change causes human centromere TGGAA repeats to fold back into hairpins. *Proc Natl Acad Sci USA* 93:12159–12164.
12. Chou SH, Zhu L, Reid BR (1994) The unusual structure of the human centromere (GGA)2 motif. Unpaired guanosine residues stacked between sheared G.A pairs. *J Mol Biol* 244: 259–268.
13. Sato N, et al. (2009) Spinocerebellar ataxia type 31 is associated with "inserted" penta-nucleotide repeats containing (TGGAA)n. *Am J Hum Genet* 85:544–557.
14. Tseng W-H, et al. (2017) Induced-fit recognition of CCG trinucleotide repeats by a nickel-chromomycin complex resulting in large-scale DNA deformation. *Angew Chem Int Ed Engl* 56:8761–8765.
15. Chen H, et al. (2012) Ionic strength-dependent persistence lengths of single-stranded RNA and DNA. *Proc Natl Acad Sci USA* 109:799–804.
16. Lo Y-S, Tseng W-H, Chuang C-Y, Hou M-H (2013) The structural basis of actinomycin D-binding induces nucleotide flipping out, a sharp bend and a left-handed twist in CGG triplet repeats. *Nucleic Acids Res* 41:4284–4294.
17. Hou M-H, Robinson H, Gao Y-G, Wang AHJ (2002) Crystal structure of actinomycin D bound to the CTG triplet repeat sequences linked to neurological diseases. *Nucleic Acids Res* 30:4910–4917.
18. Edwards SF, Sirito M, Krahe R, Sinden RR (2009) A Z-DNA sequence reduces slipped-strand structure formation in the myotonic dystrophy type 2 (CCTG) x (CAGG) repeat. *Proc Natl Acad Sci USA* 106:3270–3275.
19. Qiu Y, Niu H, Vukovic L, Sung P, Myong S (2015) Molecular mechanism of resolving trinucleotide repeat hairpin by helicases. *Structure* 23:1018–1027.
20. Chou SH, Zhu L, Reid BR (1996) On the relative ability of centromeric GNA triplets to form hairpins versus self-paired duplexes. *J Mol Biol* 259:445–457.
21. Pizarro AM, Sadler PJ (2009) Unusual DNA binding modes for metal anticancer complexes. *Biochimie* 91:1198–1211.
22. Schnabl J, Sigel RKO (2010) Controlling ribozyme activity by metal ions. *Curr Opin Chem Biol* 14:269–275.
23. Bikard D, Loot C, Baharoglu Z, Mazel D (2010) Folded DNA in action: Hairpin formation and biological functions in prokaryotes. *Microbiol Mol Biol Rev* 74:570–588.
24. Figueroa AA, Cattie D, Delaney S (2011) Structure of even/odd trinucleotide repeat sequences modulates persistence of non-B conformations and conversion to duplex. *Biochemistry* 50:4441–4450.
25. Mao C, Sun W, Shen Z, Seeman NC (1999) A nanomechanical device based on the B-Z transition of DNA. *Nature* 397:144–146.
26. Roy R, Hohng S, Ha T (2008) A practical guide to single-molecule FRET. *Nat Methods* 5: 507–516.
27. Sabanayagam CR, Eid JS, Meller A (2005) Using fluorescence resonance energy transfer to measure distances along individual DNA molecules: Corrections due to nonideal transfer. *J Chem Phys* 122:061103.
28. Minor W, Otwinowski Z (1997) HKL2000 (Denzo-SMN) software package. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol* 276: 307–326.
29. Murshudov GN, Vagin AA, Dodson EJ (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr* 53: 240–255.
30. Parkinson G, Vojtechovsky J, Clowney L, Brünger AT, Berman HM (1996) New parameters for the refinement of nucleic acid-containing structures. *Acta Crystallogr D Biol Crystallogr* 52:57–64.
31. Lavery R, Moakher M, Maddocks JH, Petkeviciute D, Zakrzewska K (2009) Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res* 37:5917–5929.
32. Lavery R, Sklenar H (1988) The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. *J Biomol Struct Dyn* 6:63–91.
33. Zheng G, Lu X-J, Olson WK (2009) Web 3DNA—A web server for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures. *Nucleic Acids Res* 37:W240–W246.