doi: 10.1093/nar/gkaa036

Dynamics of strand slippage in DNA hairpins formed by CAG repeats: roles of sequence parity and trinucleotide interrupts

Pengning Xu†, Feng Pan†, Christopher Roland, Celeste Sagui and Keith Weninger*

Department of Physics, North Carolina State University, Raleigh, NC 27695-8202, USA

Received June 19, 2019; Revised December 11, 2019; Editorial Decision January 06, 2020; Accepted January 15, 2020

ABSTRACT

DNA trinucleotide repeats (TRs) can exhibit dynamic expansions by integer numbers of trinucleotides that lead to neurodegenerative disorders. Strand slipped hairpins during DNA replication, repair and/or recombination may contribute to TR expansion. Here, we combine single-molecule FRET experiments and molecular dynamics studies to elucidate slipping dynamics and conformations of $(CAG)_n$ TR hairpins. We directly resolve slipping by predominantly two CAG units. The slipping kinetics depends on the even/odd repeat parity. The populated states suggest greater stability for 5′-AGCA-3′ tetraloops, compared with alternative 5′-CAG-3′ triloops. To accommodate the tetraloop, even(odd)-numbered repeats have an even(odd) number of hanging bases in the hairpin stem. In particular, a paired-end tetraloop (no hanging TR) is stable in $(CAG)_{n=even}$, but such situation cannot occur in $(CAG)_{n=odd}$, where the hairpin is "frustrated" and slips back and forth between states with one TR hanging at the 5′ or 3′ end. Trinucleotide interrupts in the repeating CAG pattern associated with altered disease phenotypes select for specific conformers with favorable loop sequences. Molecular dynamics provide atomic-level insight into the loop configurations. Reducing strand slipping in TR hairpins by sequence interruptions at the loop suggests disease-associated variations impact expansion mechanisms at the level of slipped hairpins.

INTRODUCTION

Trinucleotide repeats (TRs) are a subset of microsatellite repeats in the human genome where a triplet of nucleotides is repeated multiple times (1). TR tracts may expand or contract in multiples of the trinucleotide unit (2,3). If an expansion within susceptible genes crosses a certain threshold, it gives rise to Trinucleotide Repeat Expansion Disorders (2,4–6). TR expansion is associated with more than 40 neurodegenerative and neuromuscular disorders (3,7–9), some of which display 'anticipation', where the age of disease onset can decrease coincidentally with intergenerational expansion of a TR (4,10–12).

Among all the possible TRs, CAG repeats are associated with the largest category of neurodegenerative diseases. CAG repeats in the exons of diverse genes cause a variety late-onset, progressive neurodegenerative disorders including Huntington's disease (HD), dentatorubral-pallidoluysian atrophy (DRPLA), spinal and bulbar muscular atrophy (SBMA, popularly known as Kennedy's disease) and several spinocerebellar ataxias (SCAs 1, 2, 3, 6, 7 and 17) of which SCA2 is related to amyotrophic lateral sclerosis (ALS) and parkinsonism (8,12,13). These pathologies are collectively referred to as polyglutamine (polyQ) diseases (14), since the CAG expansion in these genes leads to proteins with polyQ expansions, which ultimately form aggregates before eventual neuronal death (15–17).

Expansion of TR regions is believed to be related to the repetitive structure of TRs that could cause slippage during DNA replication, repair and/or recombination (7,8,12,18–22). Although mechanisms of TR expansion may be different for small scale (expansion of one trinucleotide) or large scale (expanding large runs of TR at once) mutations (23), there is broad consensus that secondary structures of TRs, including hairpins, contribute to both phenomena (24). Numerous experimental analyses dating back to 1995 (25,26) and extending to the present have confirmed that non-B DNA secondary structures in the expanded repeats contribute to TR expansion (3). In particular, CAG repeats are associated with hairpins in *in vitro* experiments with both DNA (27–34) and RNA (35), and although directly visualizing a TR hairpin *in vivo* is still not possible, indirect experimental data suggests CAG DNA hairpins form *in vivo* (36,37).

*To whom correspondence should be addressed.Tel: 919 513 3696; Email: krwening@ncsu.edu

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

A key behavior of TR hairpins believed to be associated with TR expansion is their tendency for the two strands to slip along each other by integer units of the trinucleotide, a phenomenon known as strand-slipping. Strand-slipping in CAG DNA TR hairpins leading to overhanging single strand DNA regions has been inferred by indirect methods including single-strand nuclease susceptibility, polymerase-based extension, and chemical probing of exposed bases (28,30,38). The dynamics of the strand slipping behavior in TR hairpins has been estimated by observing temporal evolution of ensemble distributions, kinetics of conversion of hairpins to duplex, thermodynamic cycling and single molecule FRET (28–31,39).

Interruptions of the TR region by mutation of one repeat unit to a different codon can play important roles in TR disorders (40). Genetic studies find that interrupts increase the stability of alleles for several disease-related TRs (40–46). In particular, interruptions of a $(CAG)_n$ repeat tract by mutation to CAA has strong phenotypic effects in several diseases (47). CAG TRs expand less and are associated with different disease age-of-onset phenotypes when interrupted by CAA in SCA17 (46) and SCA2 (13,48–50). In the Huntington gene, the (CAG) repeat region normally terminates with a CAA codon interrupting the (CAG) tract in the penultimate position. Rare mutations involving this interrupting codon in HD recently were linked to instability of the TR region and changes in the HD age-of-onset: earlier in individuals with the loss of CAA but later in those with duplicated CAACAG at the end of the TR region (51,52). Because CAG and CAA both code for glutamine, the impact of this change on disease phenotype, despite the unchanged polyQ protein produced, taken together with evidence that mRNA levels do not impact age-of-onset (52) confirms that at least some TR disorder phenomena occur at the level of the DNA.

In vitro studies showed that CAT interrupts in CAG TR and AGG interrupts in (CGG) TRs both reduce slipped strand configurations of hairpins (53). Interestingly, CAA interrupts also increase the fidelity of polymerase chain reaction (PCR) amplification of CAG repeated sequences (54) and reduced TR expansion in a yeast genetic assay (55). These biochemical characterizations support the hypothesis that interrupts enhance genomic stability of TRs by suppressing strand slippage in TR hairpins.

Despite the extensive evidence that CAG-containing TR DNA forms interconverting, slipped-strand hairpins, at present, there is neither a known molecular structure for a CAG-repeat hairpin nor a complete description of its strand slipping kinetics. Experimental structural studies of CAG repeats are limited to RNA using X-ray crystallography (56–58), and NMR (58,59). Given that the expansions that characterize TRs originate at the DNA level, an understanding of the dynamics of hairpin slipping along with a structural characterization of these repeats at the atomic level in DNA is particularly important.

In our previous work (60), we performed free energy and molecular dynamics (MD) studies to determine the preferred conformations of the A–A non-canonical pairs in $(CAG)_n$ and $(GAC)_n$ TRs ($n = 1$ to 4) and the consequent changes in the overall structure of both RNA and DNA duplexes. We found that the global free energy minimum corresponds to A–A pairs stacked inside the core of the helix with anti-anti conformations.

In this work, we combine experimental and computational investigations to directly probe the conformational ensemble and dynamic slipping of CAG TR hairpins. We use single-molecule fluorescence resonance energy transfer (smFRET) to directly observe dynamic slipping in CAG TR hairpins by integer numbers of CAG units, predominantly two CAG units. The observed hairpin dynamics, which agrees with previously reported parity-dependent behaviors (29–31), taken together with our molecular characterization of the conformers points toward the slipping dynamics being governed by a balance between free energies in the stem and loop of the TR hairpin that leads to more frequent slipping in hairpins with an odd number of repeat units. The details of the loop configuration and the impact on slipping are confirmed by studies using TR sequence interrupts that vary the specific loop sequence and result in strongly preferred conformers. We complement the smFRET studies with classical MD studies, which completely characterize the atomic configurations, and the structural origin of the competing energetic trends between loop and stem that drive the slipping of the odd-numbered hairpins. The simulations also provide insight into the triggering instability that initiates the slipping process. Our characterization of these spontaneous TR slips defines fundamental behaviors of TR hairpins that replication machinery or other repair or recombination-related proteins must manage.

MATERIALS AND METHODS

For smFRET experiments, we designed a two-strand system involving an anchor strand and a hairpin strand positioning donor (Cy3) and acceptor (Atto647N) fluorophores at consistent positions for all hairpins (see Figure 1(A), Supplementary Figures S1, S4 and S8) (61,62). All DNAs were purchased from IDT (Coralville, IA, USA), and the sequences are listed in Supplementary Table S1. Biotinylated DNAs were immobilized at widely spaced locations on biotinylated bovine serum albumin (biotin-BSA)/streptavidin coated quartz slides inside flow chambers. Single molecule FRET signals were recorded with a homebuilt, prism-based TIRF microscope as described in (63) using a 60× 1.2 N.A. objective, a Dualview image splitter with a 645dcxr dichroic mirror, HQ585/70m (donor) and HQ700/75m (acceptor) filters (Chroma) and an emCCD camera (Cascade 512B, Photometrics) recording at 15 ms per frame using 3 × 3 binning on chip before readout. Images were background subtracted, and then donor intensity (Id) and acceptor intensity (Ia) were extracted from the movie at each spot identified to have an active acceptor in the first 150 ms of the movie. FRET efficiency was defined and calculated as Ia/(Ia + Id) and not corrected for gamma factor (64). Histograms contain all the timepoints of FRET efficiency from multiple picked molecules. All the histograms are fit with multiple Gaussian functions to identify the peak locations. Estimates of the uncertainty in the reported peak locations are described in the Supplementary methods and Supplementary Table 2. We used automatic software (65) or manual identification to identify edges of transitions in FRET

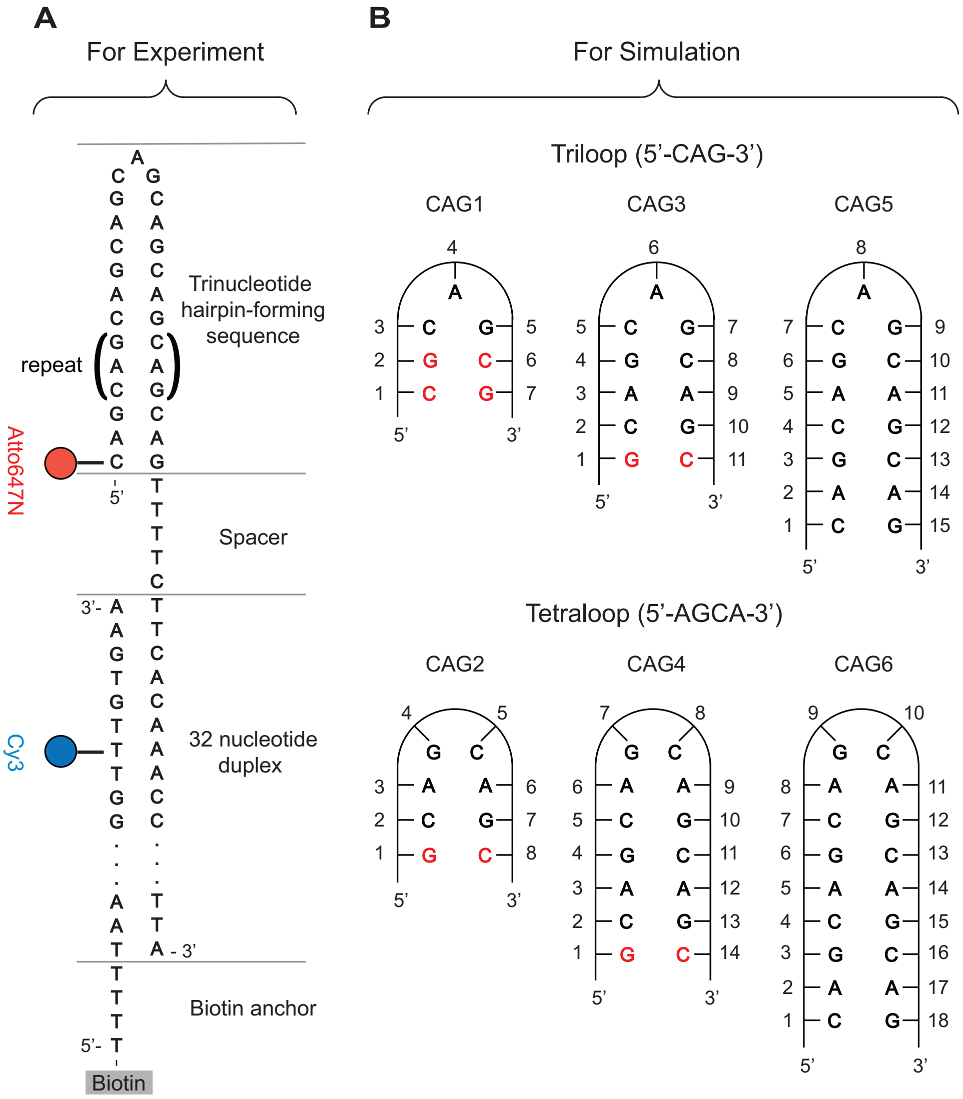

Figure 1. (**A**) Schematic DNA design for the smFRET analysis of CAG repeat hairpins, with donor in blue and acceptor in red. The CAG sequence in parenthesis is repeated for various hairpins. The hairpin loop of interest is immobilized to a surface by a partial complimentary DNA anchor strand. The TTTTC spacer helps reduce the interaction between the hairpin and junction duplex. (**B**) CAG hairpins sequences considered in the MD simulations. CG Watson-Crick pairs were added to the ends of short repeats to mimic long canonical stems.

intensity traces (Supplementary Figure S2(A)). Histograms of dwell times in each state (Supplementary Figure S2B–F) were fit to single exponential decay function to extract fitting parameters. All measurements were conducted at 21°C in buffer containing 20 mM Tris–HCl pH 8.0, 10 mM NaCl, and an oxygen scavenging system of 1% glucose, glucose oxidase (100 units/ml), catalase (1000 units/ml), 0.05 mg/ml cycooctatetraene and 1% β-mercaptoethanol unless otherwise noted.

The simulations were carried out using the PMEMD module of the AMBER v.16 (66) software package with force fields ff99 BSC1 (67) for DNA. For the waters, the TIP3P model (68) was used, as well as the standard AMBER force field parameters for the ions (69). The main CAG hairpins simulated are shown in Figure 1B. To model the long-range Coulomb interactions, the Particle-Mesh Ewald (PME) method (70) with a 9 Å cutoff and an Ewald coefficient of 0.30768 was used. The van der Waals interactions were calculated by means of a 9 Å atom-based nonbonded list, with a continuous correction applied to the long-range part. MD production runs were generated using the leapfrog algorithm with a 2 fs timestep utilizing Langevin dynamics with a collision frequency of 1 ps^{-1} at 300 K. The pressure of the system was maintained at 1 bar using the Berendsen barostat, with isotropic position scaling and relaxation time of 1 ps. The SHAKE algorithm was applied to all bonds with hydrogen atoms. Regular long MD simulations up to 2 and even 3 μs were run for all sequences, using different initial conformational values for the glycosyl torsion angles χ as will be discussed.

Additional methodological detail in supplementary methods.

RESULTS

smFRET analysis of DNA containing CAG TRs

For smFRET studies, we used a two strand DNA system that allowed the specific hairpin to be changed while maintaining a consistent placement of the donor and acceptor dyes (Figure 1A). When the hairpin closes, the acceptor and donor are close and we expect high FRET. When the hairpin is open, we expect low FRET (near zero). We verified these expectations with a hairpin formed from a stem of 6 matched Watson-Crick base pairs and a loop of 31 adenine (A) bases. This A31 loop hairpin has 2 states with FRET efficiency 0.01 and 0.67, corresponding to open and closed respectively, see Figure 2A.

We measured smFRET signals from DNA containing 14 CAG repeats, which we designate $(CAG)_{14}$ (Figure 2B). When folded into a symmetric hairpin, MD simulations indicate a conformation of a tetraloop (AGCA) with the remaining bases paired in the stem containing G-C Watson-Crick base pairs and A–A mismatches, as shown in models shown in Figure 2D. smFRET time traces from $(CAG)_{14}$ have transitions among three different FRET efficiencies of 0.65, 0.31 and 0.01 (Figure 2B,D). The 0.01 state was rarely visited, whereas there were many of transitions between the 0.31 and 0.65 states (Figure 2H). The 0.65 state was most stable with the hairpin populating that state for the majority of the time.

To assist identifying the configuration of the DNA generating these FRET efficiencies, we compare these results to the A31 loop hairpin (Figure 2A). The similarity in the highest and lowest smFRET efficiency states in the $(CAG)_{14}$ and the A31 hairpin experiments leads us to assign the lowest (0.01) state to the completely open state and the highest state to the folded hairpin state in both. The slight difference in the of the FRET efficiency of the closed states (0.67 for A31 and 0.65 for $(CAG)_{14}$), which is highly reproducible (beyond our uncertainty of ±0.01 for those FRET states. Error estimates and details on FRET efficiency measurements are described in Supplementary Tables S2 and S3.), is possibly due to transient unpairing of the CG terminal basepair of the $(CAG)_{14}$ as it is positioned beside an AA mismatch. There was no analogue in the A31 experiment for the 0.31 state observed in the $(CAG)_{14}$ experiments, which we will pursue further below.

We next measured DNA containing 15 CAG repeats, referred to as $(CAG)_{15}$. Surprisingly, smFRET measurements of $(CAG)_{15}$ revealed transitions among four states with FRET efficiencies 0.73, 0.46, 0.25 and 0.01 (Figure 2C, D). In agreement with the observation in the $(CAG)_{14}$ study, the 0.01 state was the rarest. Unlike the $(CAG)_{14}$ study, the 0.73 and 0.46 states were more equally populated, suggesting they have similar stabilities. We attribute the rare 0.01 state to the fully open state based on the similarity with $(CAG)_{14}$ and the A31 open state. The other states for $(CAG)_{15}$ (0.25, 0.46, 0.73) interleave with the states from $(CAG)_{14}$ (0.31 and 0.65). The highest state of $(CAG)_{15}$ (0.73) as higher than that of A31 and the $(CAG)_{14}$. This difference indicates that for the 0.73 state in $(CAG)_{15}$, the acceptor end is closer to the donor in the anchor compared to the 0.65 state of $(CAG)_{14}$, which could happen from the $(CAG)_{15}$ hairpin slipping by one CAG unit toward the donor on the anchor strand (Figure 2D), which we denote as '–1 slip' in Figure 2(D). Similarly, the 0.46 state could be the $(CAG)_{15}$ hairpin slipping one CAG unit away from the donor, denoted as '+1 slip' in Figure 2D. Slipping forward or backward by one CAG unit would allow $(CAG)_{15}$ to form an AGCA tetraloop turn with the stem assembling into the CAG/GAC aligned pairings, whereas aligning the CAG/GAC pair at the end of the $(CAG)_{15}$ hairpin (denoted '0 slip' in Figure 2D) would require the loop to contain one CAG unit, a CAG triloop (Figure 2D). Previous studies indirectly observed odd repeat TR hairpins tolerating single trinucleotide stem overhangs to accommodate tetraloop structures as reported by nuclease susceptibility, polymerase extension, and chemical susceptibility of specific guanine bases (28,30,38). Additionally, our MD simulations (discussed below) indicate CAG triloops are substantially less stable than AGCA tetraloops. Continuing with this hypothesis, the 0.31 state of $(CAG)_{14}$ could be slipping two CAG units and the 0.25 state of $(CAG)_{15}$ slipping three CAG units ('+2' slip and '+3' slip respectively).

We designed DNAs to test whether the intermediate states are consistent with hairpin slipping by multiples of trinucleotides. Adding unpaired T bases between the 3′ end of the hairpin and the anchor displaces the folded hairpin and associated acceptor fluorophore on the 5′ end by a known number of unpaired trinucleotides away from the donor. We inserted three extra Ts in CAG repeats containing 12, 13, 14 and 15 triplet units (called $(CAG)_{12}T_3$, $(CAG)_{13}T_3$, $(CAG)_{14}T_3$ and $(CAG)_{15}T_3$ respectively (Figure 2E, F and Supplementary Figure S1). Our models of $(CAG)_{14}$ and $(CAG)_{15}$ slipping (Figure 2D) suggest that the lower state of $(CAG)_{15}$ ($E = 0.46$ due to a slip of one CAG unit – 3 bases) should be similar to the highest state in $(CAG)_{14}T_3$, which has $E = 0.45$. Similarly, our model for the low state in $(CAG)_{14}$ being due to a slip of two CAG units (six bases) should compare well with the low state of $(CAG)_{15}T_3$ (assuming the low state in $(CAG)_{15}$ is due to one CAG unit slip), which they do, $E = 0.29$ versus 0.31. These results strongly suggest that the lower FRET state in $(CAG)_{15}$ is due to slipping by one CAG unit and the lower FRET state in $(CAG)_{14}$ is due to slipping two CAG units.

These observations highlight systematic differences in behavior of odd and even numbers of CAG repeats. Even-numbered repeats (e.g. $(CAG)_{14}$) accommodate an AGCA tetraloop with a fully strand-paired stem (without hanging trinucleotides) (Figure 2D). An alternate, less preferred configuration for even TR hairpins has the stem slipped by two CAG units, which also forms an AGCA tetraloop. In odd-numbered repeats (e.g. $(CAG)_{15}$), a paired-end stem requires formation of a CAG triloop, which we find spontaneously slips forward or backward by one trinucleotide to form an AGCA tetraloop with a hanging trinucleotide in the stem. We measured the dwell times in each state between transitions among these states in many single molecules for both $(CAG)_{14}$ and $(CAG)_{15}$ hairpins (Supplementary Figures S2 and S3). Fitting histograms of these dwell times with single exponential functions allows the characteristic lifetimes to be estimated. These quantitative dwell time measurements (Supplementary Table S4) confirm that the

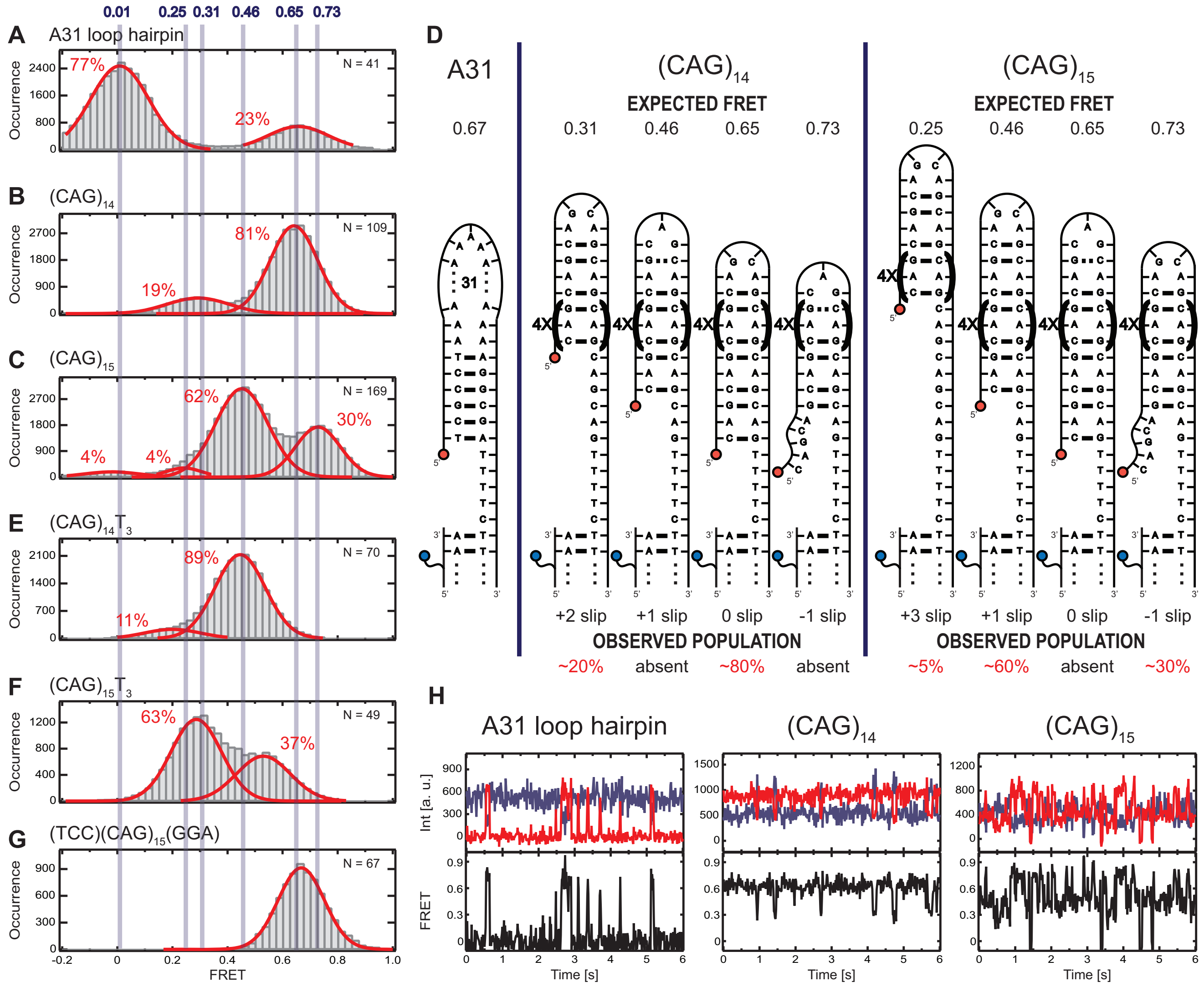

Figure 2. smFRET analysis and interpretation of A31, $(CAG)_{14}$ and $(CAG)_{15}$ hairpin loops, as well as associated hairpin variants. Histograms contain all the timepoints of FRET efficiency from multiple molecules. All the histograms are fit with multiple Gaussian functions (red line) to identify the peak locations. Here we show results for (**A**) A31; (**B**) $(CAG)_{14}$; (**C**) $(CAG)_{15}$; (**E**) $(CAG)_{14}T_3$; (**F**) $(CAG)_{15}T_3$ and (**G**) $(TCC)(CAG)_{15}(GGA)$. In (**D**), we present schematic diagrams to show molecular configurations, expected FRET values (upper edge) and the corresponding observed populations (lower edge) of the individual hairpins. The slip designation is indicated below each schematic. (**H**) Representative smFRET time traces (upper panels, donor signal in blue, acceptor signal in red) and calculated FRET values (lower panels, black) for A_{31}, $(CAG)_{14}$ and $(CAG)_{15}$ hairpin loops.

$(CAG)_{15}$ hairpin is more dynamic with more frequent slipping transitions than the $(CAG)_{14}$ hairpin (Figure 2H), suggesting the potential triloop for odd-parity hairpins frustrates their stability. We have verified this general pattern of states and kinetics is observed in other numbers of even and odd CAG repeats by measuring hairpins with 7–13 repeats (see Supplementary Figures S5 and S6). The dwell times become shorter as the hairpins become shorter. It is possible that the FRET states observed in the shortest hairpins (e.g. $(CAG)_7$) arise from rapid averaging of two underlying configurations interchanging at a rate just below our 15 ms resolution (71). Transitions between the fully open hairpin and any closed hairpin state (any degree of slipping) were rare compared to transitions between closed states of the hairpin with different degrees of slipping for both even and odd CAG repeat numbers suggesting the barriers to open are larger than to slip. The characteristic dwell time for $(CAG)_{14}$ were longer at 15°C compared to our standard 21°C measurements (Supplementary Table S4) while the populations of the states were not strongly affected (Supplementary Figure S7), suggesting the barriers between the states are on order of thermal energy around room temperature. In addition, higher cation concentration stabilized the symmetric folding state (indicated by an increase in the highest FRET state population) and slightly increase the FRET efficiency of the state (0.73 in 10 mM NaCl, 0.76 in 100 mM NaCl, 0.78 in 1000 mM NaCl, and 0.81 in 10 mM $MgCl_2$) for $(CAG)_{15}$ (Supplementary Figures S3 and S7). The difference in stability of the slipped CAG states in even and odd numbers of repeats suggests that a delicate balance between energies in the combination of the stem and either triloops or tetraloops that form in CAG repeat sequences make critical contributions to the overall stability of the slipped states of the CAG repeat hairpins.

CAA interrupts select preferred slipped strand conformers of $(CAG)_n$ TRs

Because CAA interrupts within $(CAG)_n$ tracts impact disease phenotypes in HD and in SCA2, we measured $(CAG)_{15}$ hairpins that were interrupted by CAA at the seventh and eighth position (noted as $(CAG)_6(CAA)(CAG)_8$ and $(CAG)_7(CAA)(CAG)_7$). The interrupt is located near the natural loop position in these constructs (Figure 3A). Placing the CAA interruption at position 7 (Figure 3D) dramatically stabilized the +1 slip with a new, minor population forming at 0 slip (FRET 0.65). The CAA interruption at position 8 (Figure 3(C)) stabilized the –1 slip state and also generated a small population of 0 slip not observed in $(CAG)_{15}$. For the seventh and eighth position CAA interrupts, slipping transitions were strongly suppressed compared to uninterrupted $(CAG)_{15}$ (Figure 3E).

The stabilized states with the CAA interrupt at the seventh and eighth positions both are consistent with a tetraloop sequence of AACA closed by two G–C Watson–Crick bonds, differing from the AGCA loop that is present in the two states that interconvert for the $(CAG)_{15}$ hairpin. The minor 0 slip state in the CAA interrupted hairpins may be a larger, 7-base loop assuming the possible G–C basepair within that loop is not stabilized. Importantly, the slipping by two CAG units (between +1 and –1) that is characteristic of $(CAG)_{15}$ is eliminated by CAA near the loop. In total, these studies indicate that the presence of a point mutation generating a single CAA interrupt dramatically stabilizes the strand-slipping dynamics in the CAG TR hairpins when the sequence change occurs in the loop.

smFRET experiments of loop sequence variants

We tested additional interrupts in $(CAG)_{15}$ to investigate the relative influence between the loop and the closest base pairs in the stem on overall hairpin configuration bias. Changing the middle CAG in $(CAG)_{15}$ to AAA ($(CAG)_7(AAA)(CAG)_7$) results in a triloop configuration that has AAA or a tetraloop with AGAA and the loss of a G-C at the base of the loop (Supplementary Figure S8C). For this construct, we measured predominantly 0.65 with only short, rare excursions to 0.37 and nearly no open hairpin (0.01), indicating that the fully matched CAG/GAC paired stem (0 slip) with the AAA triloop was highly stabilized relative to the slipped stem with a tetraloop (middle model, Supplementary Figure S8C).

The triloop configuration of $(CAG)_{15}$ could be stabilized by adding three Watson–Crick base pairs to the end of the hairpin (Figure 2(G) and Supplementary Supplementary Figure S1F). For $TCC(CAG)_{15}GGA$, we observed only one state at $E = 0.67$, consistent with the ends of the hairpin being stable locked by the TCC/GGA basepairs. The $(CAG)_{15}$ middle section could either form a matched CAG/GAC stem with a CAG triloop, or yet another more complex configuration containing a bubble of unpaired bases. These sequence variants confirm that changes as small as 2 bases in the loop or 3 bases at the end of the stem dramatically change the configuration of $(CAG)_{15}$.

In another interrupt variant of $(CAG)_{15}$, $(CAG)_6(CGG)(CAG)_8$ (Supplementary Figure S8D), we observed preference for a different tetraloop. In this sequence, tetraloops can be either GGCA (+1 slip) or AGCA (−1 slip). smFRET measurements for this sequence found the 0.46 state to be nearly 10 times more populated indicating the GGCA loop is substantially favored over the AGCA. Notably, the potential CAG triloop that would accompany matched stem alignment was absent. Demonstrating the subtle energies of these configurations, the AGCA tetraloop can be stabilized relative to the TGCA tetraloop (left model 0.46 FRET, Supplementary Figure S8(E)) by changing a single base (G to T) as demonstrated in the variant $(CAG)_6(CTG)(CAG)_8$ (Supplementary Figure S8(E)). In this variant, the 0.73 state is dramatically stabilized compared to that state in $(CAG)_6(CGG)(CAG)_8$.

In contrast, an ACCA tetraloop with one CAG unit slip in the stem can be stabilized in the sequence $(CAG)_6(CAC)(CAG)_8$, which replaces only one G base of the loop in the original $(CAG)_{15}$ by a C (Supplementary Figure S8A). We observed $(CAG)_6(CAC)(CAG)_8$ primarily exists in the 0.46 FRET state, confirming that the ACCA tetraloop was preferred. A minor population was observed at 0.65, which could be a matched stem with larger loop ACCAGCA (expected $E = 0.64$) and seems unlikely to be the forward slipped AGCA tetraloop closed by one fewer basepairs in the adjacent stem (expected $E = 0.73$).

Shifting the G to C substitution to the other side of the loop has unexpected consequences. For $(CAG)_8(CAC)(CAG)_6$ we measure three states (in addition to the rare $E = 0$ open state): 0.25 (20% population), 0.46 (35% population) and 0.65 (40% population) (Supplementary Figure S8B). The 0.46 state is consistent with the hairpin slipped back by one CAG unit (+1 slip), forming an AGCA tetraloop. The G to C substitution sacrifices one W-C base pair far down the stem. Similarly, the 0.25 is similar to the hairpins slipped by two CAG units (+2 slip). The 0.65 state is substantially different than the $E = 0.73$ FRET level expected on forward slip by one CAG unit (−1 slip). It is closer to the values seen when the CAG/GAC pair symmetrically at the end of the stem with no slipping. In this case, the G to C substitution eliminates one C–G basepair in the lock between the stem and loop. This one base pair lock may not be sufficient and it could open to form a nine base pair loop.

Taken together, these result highlights the subtle balance between loop and stem free energies in determining the stability of the triloop and tetraloop configurations of CAG repeat sequences. Measurements of the energy differences for some hairpin states derived from DNA melting are reported in Supplementary Table S5 and Supplementary Figure S9.

Atomistic MD simulations

Our smFRET experiments indicate different hairpin slipping in odd and even number of repeats in CAG sequences but cannot unequivocally determine the number of bases that form the hairpin loop. Based on indirect evidence (27,72–74), it has been suggested that these loops can be tight, i.e. formed by a few bases. We used MD simulations to provide insight into the full atomic level conformations

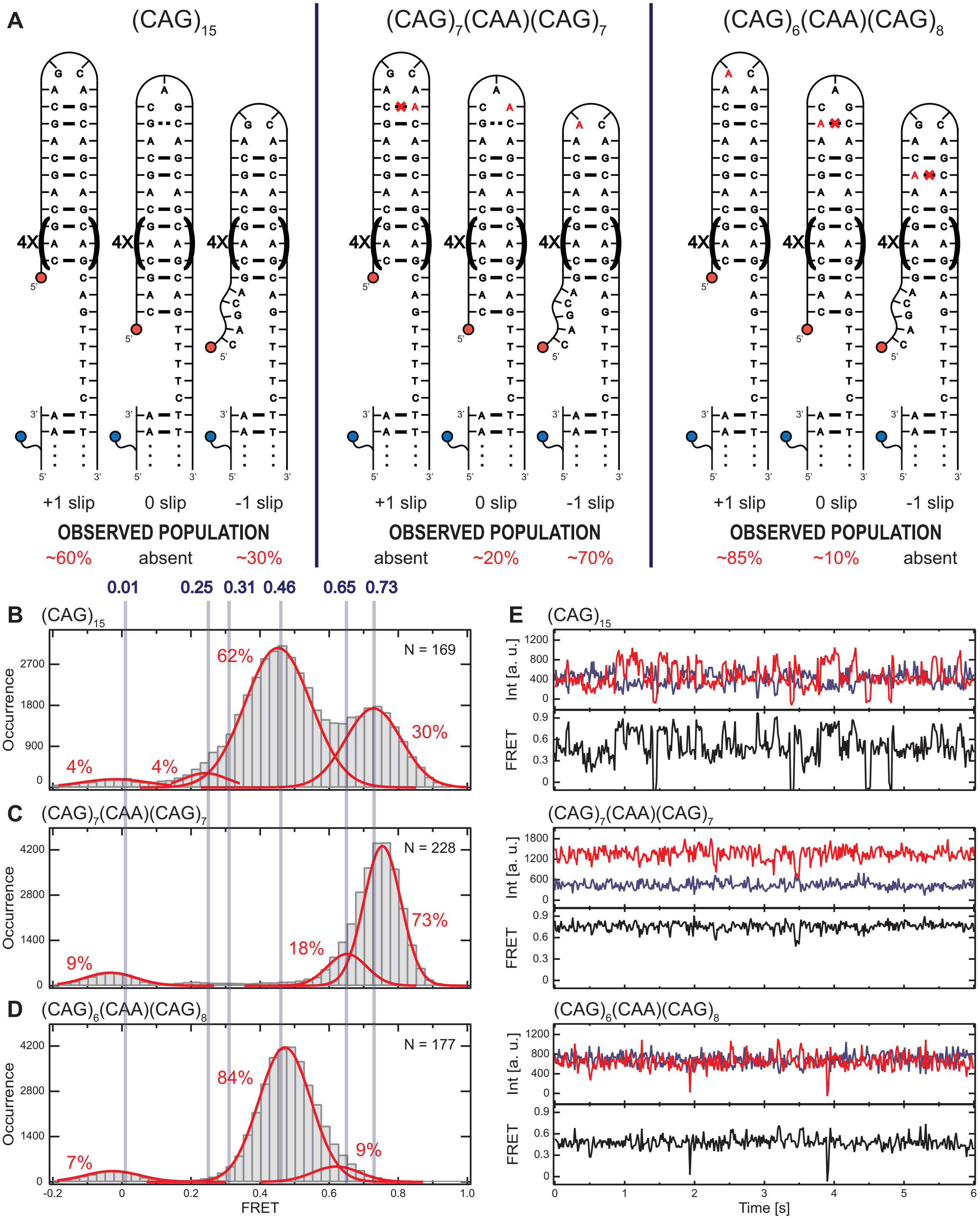

Figure 3. CAA Interrupts near the loop can stabilize $(CAG)_{15}$ strand-slipping transitions. (**A**) Schematics of possible folding configurations and smFRET results for $(CAG)_{15}$ and several CAA interrupted $(CAG)_{15}$ TR hairpins. Sequences are indicated at the top of the diagram. Slip designation and observed populations are indicated along the lower edge. Results of $(CAG)_{15}$ experiment from Figure 2 are listed for comparison. (**B–D**) Histograms of smFRET measurements for the hairpins diagramed in (A). (**E**) Representative time traces (upper panels, donor signal in blue, acceptor signal in red) and calculated FRET values (lower panels, black) for the $(CAG)_{15}$ and the CAA interrupted $(CAG)_{15}$ TR hairpins.

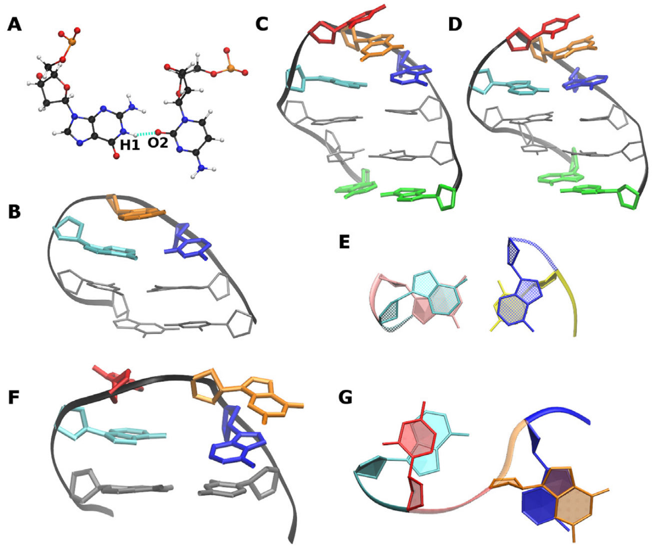

Figure 4. (**A**) The sheared C·G pair with a C(O2)–G(H1) H-bond in the stable clamped triloop of CAG1 (Figure 1). (**B**) A side view of the CAG1 triloop with C3(blue), A4(orange) and G5(cyan). (**C** and **D**) Two long-lived loop conformations for CAG4 (Figure 1): (C) S(aa)-L(as)-G(a) (stable) and (D) S(as)-L(ss)-G(s) (kinetic trap). The cartoon only shows the residues from A3 to A12 with A6(blue), G7(orange), C8(red), A9(cyan) and A3–A12(green). (**E**) The strong stacking between the A–A mismatch in the tetraloop of CAG4 and the C–G Watson–Crick basepair immediately below: C5(yellow)/A6(blue) and A9(cyan)/G10(pink) in the S(aa)-L(as)-G(a) conformation (bases in the 5′-AGCA-3′ tetraloop in conformation *anti–anti–anti–syn*, with *anti–anti* conformation for the first mismatch closer to the loop, see text). (**F**) The two-stack loop in L(aa) (A–A in 5′-AGCA-3′ in anti-anti) for CAG2. Bases in colors: A3(blue), G4(orange), C5(red) and A6(cyan). (**G**) View for (F) of the stacking of A3(blue)/G4(orange) and C5(red)/A6(cyan).

of CAG hairpins and relative stabilities of specific loops configurations (Figure 1B). Simulations with 27 different initial structures for triloops for a total time of 55.5 μs and 38 initial structures for tetraloops for a total time of 67.5 μs were performed. In addition, we ran CAA, AAA, CAC, CGG mutated sequences starting from both triloops and tetraloops (Supplementary Table S11), and $(CAG)_9$, $(CAG)_{11}$ and $(CAG)_{15}$ hairpins both at 0 and 100 mM NaCl in order to identify the spontaneous transition from the less stable triloop towards the tetraloop.

Triloop simulations

Simulations of the CAG1 hairpin (5′-CG-CAG-CG-3′, Figure 1(B)) considered four possibilities regarding the χ angles in the CG base pair closing the CAG triloop: C3(anti)-G5(anti), C3(anti)-G5(syn), C3(syn)-G5(anti) and C3(syn)-G5(syn). Combined with two conformations for the middle A base: A4(anti) and A4(syn) (Supplementary Figure S10), gives a total of eight different triloop conformations. All conformations for the CG pair in CAG involving a syn angle are unstable, and only C(anti)-G(anti) is stable. A4(*syn*) converts into A4(*anti*) after 10ns, which results in a stronger stacking of A4/G5, as shown in Supplementary Figure S10. The hairpin with every base in CAG in anti-conformation is stable after 1 μs simulation. This triloop shows a sheared C·G pair (Figure 4A), which has been reported on a GAC triloop (75). The sheared C·G pair has a typical C(O2)–G(N1-H1) hydrogen bond. A side view of the CAG triloop is shown in Figure 4B. The two G–C Watson–Crick base pairs in the stem help stabilize this triloop.

To study the stability of hairpins with longer stems, we simulated CAG3 (5′-G-$(CAG)_3$-C-3′) and CAG5 (5′-$(CAG)_5$-3) (Figure 1(B)). Both were found to be unstable although there is increased stability as the stem length is increased. In particular, CAG5 begins unraveling at approximately 5 μs. See supplemental discussion for more details (Supplementary Table S6, Supplementary Figures S11–S13) as well as a discussion of an observed anti-syn versus anti-anti transitions in the A–A stem mismatches (60) as ion concentration is increased.

Tetraloop simulations

We simulated even CAG repeat numbers to examine 5′-AGCA-3′ tetraloops. For studies of CAG2 [5′-G-$(CAG)_2$-C-3′], numbering in the tetraloop as A3–G4–C5–A6, Figure 1B), we set initial conformation of the C5 base as anti because it can flip fast. There are four possibilities regarding the χ angles in the A–A base pair closing the loop: A3(*anti*)–A6(*anti*), A3(*anti*)–A6(*syn*), A3(*syn*)–A6(*anti*) and A3(*syn*)–A6(*syn*). We name them L(aa), L(as), L(sa), L(ss) (L stands for loop) (Supplementary Figure S14). The G4 base can have two conformations: G(*anti*) and G(*syn*), which results in eight different initial conformations.

Time courses for CAG2 demonstrate that loop A3–G4–C5–A6 populates possible stable conformations during the 2 μs simulation (Supplementary Table S7, Supplementary Figure S15). These include L(as)–G(a); L(sa)–G(s); L(ss)–G(s); and L(aa)–G(a) and L(sa)–G(a) that interconvert into each other. Additionally, a different type of loop structure with A/G and C/A stacked on different sides was found in L(aa)–G(a) and L(ss)–G(a) (see Figure 4F for L(aa)). We name it a two-stack loop since it has two stacks of overlapping bases. Stacking of A3/G4 and C5/A6 contribute to the relative stability of the structure (Figure 4G). Two-stack loops have not been observed experimentally in DNA but have been reported in some RNA tetraloops (PDB ID: 1K4A (76), 1AFX (77), 1K6G (78)). None of the tetraloops studied exhibit a full blown instability, as some of the triloops did. Supplementary Movie S1 in the supplemental discussion shows an example of a transition from L(aa)–G(s) to L(aa)–G(a) at ~890 ns. This loop then transitions to a two-stack loop ~1370 ns and it finally transitions back to the regular tetraloop ~1500 ns.

We next simulated CAG4 [5′-G-$(CAG)_4$-C-3′], where an extra stem A–A mismatch may impact hairpin stability (Figure 1B). We considered three initial conformations for the A–A mismatch (A3(*anti*)–A12(*anti*), A3(*anti*)–A12(*syn*) and A3(*syn*)–A12(*anti*), named S(aa), S(as) and S(sa) respectively, where S stands for stem (Supplementary Figures S16– S18, Supplementary Table S8). We found that the S(aa)–L(as)–G(a) conformation has high stability in the 2 μs simulation and other initial states transition to this configuration frequently (Figure 4C). Another conformation of S(as)–L(ss)–G(s) in kinetic trap was also observed (Figure 4D). L(aa)–G(a) conformations (with different A–

A arrangements) coexist with two-stack loops, and are also long-lived. The stability of S(aa)–L(as)–G(a) is enhanced by C5/A6 and A9/G10 stacking (Figure 4E). Other transitions to two-stack loops are discussed in the supplemental. The analysis of hydrogen bond and base-stacking are presented in Supplementary Tables S9–S10, while Supplementary Figure S19 gives a principle component analysis.

Since transitions between different hairpin conformations are relatively rare in the simulation time scales, we cannot carry out a strict cluster analysis but we can present an approximation to it, as shown in Figure 5. This figure is a schematic summary of the 72 times trajectories of the RMSD of the AGCA tetraloop in the CAG4 hairpin and of the χ angles of the neighboring A3–A12 mismatch, as well as of the associated transitions between different conformations (Supplementary Figure S16–S18). Circles represent the eight loop conformations and the three two-stack loops found in our simulations, averaged over the conformation of the A–A mismatch nearest to the loop. The area of each circle represents the percentage of the total simulation time that the hairpin spends on that conformation, while the width of the arrows is proportional to the frequency of transitions. We define a ‘well-connected’ circle as one that has at least two arrows linked to it. A growing circle is one where the net incoming arrows (incoming arrows minus outgoing arrows), weighed by their width, is positive. A shrinking circle has net weighted incoming arrows negative. Thus, the most favored conformations correspond to large, growing circles. The large circles corresponding to L(ss)–G(a) and L(ss)–G(s) are not well connected because they correspond to kinetic traps (see more details and principal component analysis in Supplementary Figure S19 in supplemental discussion). The most favored conformations correspond to large, growing circles. There are two net growing circles, the red L(as)–G(a); and the two versions of L(aa)–G(a): the blue circle (one stack) in coexistence with its two-stack alternate conformation (shaded circle). These are the most favored hairpin conformations for the tetraloops.

Interrupts in CAG hairpins to test loop stability

In simulations of hairpins containing variants of the CAG repeat sequences (Supplementary Table S11), we observe dynamic loop re-arrangements consistent with experimental results. For both triloops and tetraloops with CAA interrupts, we simulated the structures with the interrupt at different positions as the experiment did. First, we notice that the triloops (with the CAA interrupt in the middle or to either side of the triloop) tend to show weaker stability, in agreement with the smFRET results. The initial triloops convert into a heptaloop [$(CAG)_6$–CAA–$(CAG)_8$], or manifest a tendency to slip [$(CAG)_7$–CAA–$(CAG)_7$, which shifts one base towards the 5′ direction, consistent with the –1 slip behavior observed in experiments]. With respect to the tetraloops, our simulations show that a CAA interrupt in the 5′ side of an AGCA tetraloop is not stable, in agreement with the zero population for –1 slip in the $(CAG)_6$–CAA–$(CAG)_8$ smFRET result (Figure 3). The simulation where the bases C and A of the interrupt CAA form part of the tetraloop A**GCA-A** is not stable either, in agreement with the zero population for the +1 slip in the $(CAG)_7$–CAA–$(CAG)_7$ smFRET result (Figure 3). Finally, the **AA**CA tetraloop shows good stability, in agreement with the –1 slip in in the smFRET $(CAG)_7$–CAA–$(CAG)_7$ hairpin (70% population) and the +1 slip in the smFRET $(CAG)_6$–CAA–$(CAG)_8$ hairpin (85% population).

With respect to the AAA mutation, the AAA-triloop remains stable during the simulation, in agreement with the smFRET AAA triloop hairpin. On the other hand, simulations show that the AAA-mutated tetraloop shifted to the usual GAA triloop. For CAC and CGG mutations, the triloop is not stable, also in agreement with the experiments. For the CAC and CGG mutations in tetraloop form, the simulations are stable. This supports the assumption in FRET experiments that the triloops with CAC and CGG mutations shifted to the lower free-energy tetraloops detected by FRET.

Structurally, these results can be explained as follows. When the synonymous point mutation G→A changes the 5′-A**G**CA-3′ tetraloop into the 5′-A**A**CA-3′ tetraloop, the stacking previously described for the AGCA tetraloop is preserved, and the same stability considerations apply. However, when this mutation takes place anywhere else in or close to the loop in the hairpin, it either adds mismatches or destabilizes the stacking, therefore resulting in less stable hairpins.

The AAA triloop has a stronger purine-purine stacking (A/A) compared with the pyrimidine-purine (C/A) stacking in the CAG triloop, which enhances the AAA triloop stability. The AAA-mutated tetraloop morphs into a larger CAGAAA loop, which increases disorder and reduces the number of Watson-Crick base pairs. It becomes unstable and transitions to a stable GAA triloop. Similar explanations apply to the CAC and CGG triloop mutations. When the CAC and CGG mutations occur in tetraloops, the tetraloop stability is conserved due to the presence of purine-purine mismatches closing the loop: A–A in the ACCA tetraloop formed by the CAC mutation, and G–A in the GGCA tetraloop formed by the CGG mutation. In addition, they have a G–C Watson–Crick base pair immediately after, helping lock the tetraloop.

Extended $(CAG)_n$ triloop

Both experiments and simulations indicate that CAG triloops are less stable than AGCA tetraloops. Thus, we carried out MD simulations on an initial $(CAG)_n$ triloop (n = 9, 11, 15, at 0 and 100 mM NaCl) to gauge the steps in the transition towards a tetraloop. Depending on the case, base shifting produces GCA triloops and GCAG tetraloops (Supplementary Table S12). For instance, the $(CAG)_{15}$ hairpin under 100 mM NaCl excess salt undergoes large fluctuations after 1 μs, and at a certain point a base pairing shift results in a GCAG tetraloop. As shown in Supplementary Figure S20, the original A20–A26 mismatch immediately preceding the triloop breaks due to the flipping of the A26 base into the minor groove. This causes the C25 base to shift down along the 3′ direction to form an A20–C25 mismatch, leaving G21, C22, A23 and G24, in a temporary GCAG tetraloop conformation. We suggest that this is the first step in the mechanism of the shift transition from an unstable

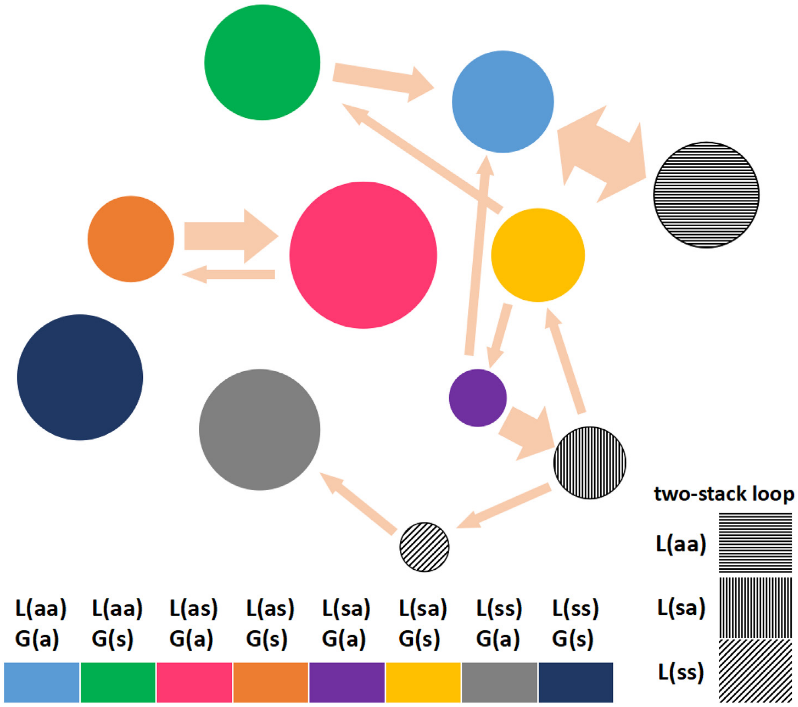

Figure 5. Results for the cluster-like analysis of different conformations for the AGCA tetraloop in the CAG4 hairpin. Different conformations are represented as different colors. Three different kinds of two-stack loops are shown in different shadows. The areas of the circles are representative of the percentage of time that the system spends in that conformation. Likewise the arrows indicate the observed transitions during the simulation with the width of the arrow representing the estimated frequency of transitions. Favorable conformations are indicated by large, growing circles. The simulation results show that the L(as)–G(a) (red; AGCA in *anti–anti–anti–syn* conformation) and the two versions of L(aa)–G(a) (blue, gray with horizontal line shading; AGCA all in *anti* conformation) are the most favorable conformations, as discussed in the text.

CAG triloop to a tetraloop. As experiments observed transitions between slipped states both involving tetraloops, it is suspected this transient, unstable triloop might be in intermediate state during the transition that is too short-lived to be resolved in experiments. A similar shifting behavior was observed in $(CAG)_9$ with zero salt, but a GCAG tetraloop with two-stack structure formed after transition from an intermittent unstable CAGC loop structure (see Supplementary Movie S2).

DISCUSSION

In this work, we have used smFRET measurements and MD simulations to characterize the structure and dynamics of DNA hairpins formed from CAG TRs. Including different lengths and/or mutations due to interrupting trinucleotides, the smFRET experiments have included 23 different sequences, and the MD simulations have included 13 different sequences with 71 different initial conditions. The smFRET results for the $(CAG)_{14}$ and $(CAG)_{15}$ oligomers show that the $(CAG)_{14}$ oligomer displays transitions between 0.65, 0.31 and 0.01 FRET states, while transitions for $(CAG)_{15}$ correspond to 0.73, 0.46, 0.25 and 0.01 FRET states (Figure 2). The 0.01 state corresponds to the open hairpin, a state that is rarely observed. These FRET values for even and odd hairpins interleave and represent sequential amounts of slipping of one strand of the hairpin stem with respect to the other by integer units of the CAG TR. Using the hairpin with paired ends as reference (zero slip), the FRET states and the associated slips are: 0.73 (–1 slip, 3′ → 5′, such that one TR overhangs at the 5′ end), 0.65 (0 slip), 0.46 (+1 slip, 5′ → 3′, one TR overhanging at the 3′ end), 0.31 (+2 slip) and 0.25 (+3 slip) (Figure 2D). Using smFRET to observe real-time trinucleotide stem slipping dynamics confirms previous results using chemical probes of guanine accessibility, polymerase based extension, or single strand nuclease activity that support such configurations (28–30,38). However, there are subtle differences, such as the use of only even-numbered sequences $(CAG)_n$ in those studies. For instance, the sequences employed in gel electrophoresis experiments in (29) are of the form $(CAG)_{16}(CTG)_4$ (where triloops are enforced through Watson-Crick pairing with the self-priming CTG repeats). In addition, smFRET directly reveals spontaneous, dynamic slipping between configurations with variable degrees of slip and underlines the completely different dynamic patterns between even- and odd-numbered sequences.

Folding CAG TRs into hairpins without any overhang in the stem requires triloops for odd repeat numbers and tetraloops for even numbers of repeats (Figure 2). Therefore, the offset in slipping number we observe between odd and even TR suggests slipping is driven to achieve the tetraloop configurations, as was previously suggested (29,30). The ability to stabilize triloops by changing only a few bases either in the loop or the stem indicates that a competition between loop and stem free energy is delicately balanced. Indeed, a recent smFRET study (79) of CAG repeat hairpins that intentionally removed the CAG pattern in the loop region and replaced it with polyA or polyT linkers (15 bases) did not observe slipped hairpin. These CAG stems with polynucleotide linkers were useful for studying stability of the stem region, but did not reveal spontaneous slipping of the stem, confirming the significance of the requirement of the CAG sequence in the loop for driving the phenomena.

MD simulations reveal details of the basic interactions that stabilize various configurations of CAG hairpins. CAG loops are marginally stable in a tight 5′-CAG-3′ triloop arrangement, as long as there is some clamping in the end of the stem, while tight 5′-AGCA-3′ tetraloops are stable even without clamping. Clamping can naturally occur in the cell when the ends of the loop interact with other molecules. Overall, all simulations indicate that tetraloops are more stable than triloops, in agreement with the experimental findings. In a 5′-CAG-3′ triloop, the three nucleotides are in *anti* conformation, the C base flips out and the weak sheared C·G pair is held by a single hydrogen bond (Figure 4A, B). The 5′-CAG-3′ triloop is then 'locked' by a weak AG/CA step (where the A bases are mismatched), whose melting temperature has been estimated as ∼13°C (29). In contrast, 5′-AGCA-3′ tetraloops (Figure 4C, D) are stabilized relative to triloops by favorable stacking energy within the loop, less bending deformation of the backbone; and locking by a GC/GC step, the strongest of all possible steps, whose melting energy has been estimated as ∼136°C (29). Considering the values of the glycosidic angle χ, the 5′-AGCA-3′ tetraloop shows two preferred conformations, where the four nucleotides are either in *anti–anti–anti–syn*, L(as)–G(a), or *anti–anti–anti–anti* L(aa)–G(a) conformations (Figure 5) The L(as)–G(a) conformation actually consists of two dynamically coexisting conformations with either single A/G or single G/C stacking (Supplementary Figure S19). The L(aa)–G(a) conformation also consists of two main dynamically coexisting conformations with (i) a single A/G or single G/C stacking similar to Supplementary Figure S19 that is not long-lived and dissolves into no stacking; or (ii) a long-lived double A/G and C/A stacking (Figure 4F, G).

In addition to the stability difference between the fundamental tri- and tetraloop configurations (including the locking basepair step at the base of the loop), one has to consider the free energy of the stem. Even-numbered sequences can accommodate a tetraloop while forming a paired-end stem without overhangs, which minimizes the free energy of the full hairpin. The stability of this state is reflected in the highly populated FRET 0.65 state of $(CAG)_{14}$. Odd-numbered sequences, on the other hand, can only accommodate an AGCA tetraloop if one strand is displaced with respect to the other by at least one TR (becoming *di facto* an even-numbered hairpin with a hanging base). However, the hanging base in the hairpin stem costs extra energy, although apparently not enough to deter the preference for tetraloops. Taking the paired-end state as reference, odd-numbered hairpins slip by odd (mainly 1 but also 3) CAG TR units. The resulting dynamics are therefore rather different: even-numbered hairpins reside most of the time in the paired-end state with occasional two CAG TR slips, while odd-numbered hairpins are essentially 'frustrated', never populating by a measurable amount of time the paired-end triloop, and slipping back and forth in one direction or the other. Of course, since the zero-slip reference for odd-numbered hairpins is not populated (at least in a measurable way), the odd-numbered hairpin is also slipping by two units as it goes from one state to the other. Molecular details of the mechanisms leading to these changes are further discussed in the supplemental discussion, where results for other hairpin lengths are also presented. We have carried out MD simulations to try to elucidate how the triloop favored by pairing the two strands of the stem would transition to a tetraloop. Although the simulations were not long enough to elucidate the mechanism of the transition, they indicate that the transition from a triloop in a $(CAG)_{odd}$ sequence is triggered by the disruption of the A-A mismatch closest to the loop, where the A base on the 3′ strand switches towards the minor groove allowing the temporary formation of a GACG tetraloop.

Finally, we analyzed the effect of trinucleotide mutations in the hairpin stability. In principle, arbitrary trinucleotide mutations can alter hairpin structure and dynamics in a myriad of ways, and the effect of these on hairpin structure has been briefly discussed above and further presented in the supplemental discussion. Genetic studies find that interrupts increase the stability of alleles for several disease-related TRs (40–46). Of particular interest in the polyglutamine diseases is the synonymous point mutation G→A, which changes the CAG codon into the CAA codon, both coding for glutamine. Interruptions of a $(CAG)_n$ repeat tract by mutation to CAA has a stabilizing effect in the polyglutamine diseases (47) such as Huntington's disease (51,52), SCA17 (46) and SCA2 (13,48–50). For instance, the instability of expanded CAG repeats in SCA17 is up to 3-fold higher in patients with uninterrupted CAG repeats than in those with CAA interrupts (46). In this work we have shown that CAA interrupts in a CAG sequence can dramatically reduce strand slipping in the corresponding hairpins. In particular, when the mutation takes place in the second base in the tetraloop, i.e. the original 5′-A**G**CA-3′ tetraloop mutates into the 5′-A**A**CA-3′ tetraloop, the stacking previously described for the AGCA tetraloop is preserved, and the same stability considerations apply. This results in the highly populated states of the -1 slip in the $(CAG)_7$-CAA-$(CAG)_7$ hairpin (70% population) and the +1 slip in the $(CAG)_6$–CAA–$(CAG)_8$ hairpin (85% population). However, when this mutation takes place anywhere else in or close to the loop in the hairpin, it either adds mismatches or destabilizes the stacking, therefore resulting in much less stable hairpins.

Interestingly, a recent smFRET study (80) of a pentanucleotide hairpin $(TGGAA)_n$ found similar slipping by one repeat unit as we characterize for CAG TRs. The (TGGAA) study reported dramatically different behavior for even and odd numbers of repeats. That study concluded that the stem energies dominated the behaviors and formulated a local seeding model where the kinked GGA in the stem is key to the slipping transitions. Our results contrast with this result by highlighting the importance of the loop and the nuanced balance in energies between the loop and stem interactions for the CAG hairpins. Intriguingly, both sequences exhibit pathological expansion phenomena, which suggests that the mechanisms of the expansion might be similar after a hairpin slips, even if the molecular details driving the slip are different.

In summary, we have confirmed that the strands of a CAG TR hairpin slip with respect to each other by an integer number of CAG units, most commonly two, using smFRET, which is capable of directly resolving temporal kinetics of the slipping in comparison to previous, indirect approaches to assess slipped hairpins with single strand overhangs (28–30,38). We also present the first atomic structural and stability analysis of the possible CAG hairpin conformations. Our results are supported by previous thermodynamic studies of CAG hairpins showing that there are even/odd differences in unfolding energetic parameters (31,81). Our data suggest that slippage away from triloops is triggered by an instability in the A-A mismatch closest to the hairpin loop, giving rise to the formation of a temporary tetraloop.

Strand slippage may be crucial for the expansion of TRs. Although the *in vivo* context is long double-stranded DNA molecules, the results of our *in vitro* studies can inform aspects of a simple model that have been suggested previously (28). This model suggests that if strand separation were to occur in the region of the initial complementary (CAG)·(CTG) duplex, then CAG and CTG hairpins could form opposite each other in a cruciform-like stem structure. Strand slippage would thus allow these hairpins to travel apart in a soliton-like wave or 'rollamers' (28,82). This simple model proposes that single-stranded cleavage in one strand of a hairpin (nicking) would permit the hairpin be stretched open, leaving a single-stranded gap in the local duplex (28). The subsequent filling of this gap in the nicked strand would then result in TR expansion. Hairpin migration apart would enhance expansion because directly opposed, complementary hairpins (like the cruciform-like structure) might be more likely to collapse back to the (CAG)·(CTG) duplex, minimizing expansion. Notably, TR hairpins contain multiple mismatched bases, which are recognized by DNA mismatch repair proteins. Activation of the latent endonuclease activity of MutLα as part of the mismatch repair cascade upon detection of a mismatch in the hairpin could be the source of the nick required to permit the TR hairpins to open in this model. Such a mechanism would agree with the observation that suppressing DNA mismatch repair activity can reduce TR expansion (3,83). Within this model, suppression of strand-slipping when an interrupting trinucleotide selects for a specific hairpin configuration could suppress possible hairpin migration and reduce this pathway of TR expansion.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr. Dorothy Erie for discussions.

FUNDING

National Institute of Health [NIH-R01GM118508]; National Science Foundation (NSF) [SI2-SEE-1534941]; Extreme Science and Engineering Discovery Environment (XSEDE) [TG-MCB160064]. Funding for open access charge: National Institute of Health [NIH-R01GM118508]; National Science Foundation (NSF) [SI2-SEE-1534941].
Conflict of interest statement. None declared.

REFERENCES

1. Ellegren,H. (2004) Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.*, **5**, 435–445.
2. Orr,H.T. and Zoghbi,H.Y. (2007) Trinucleotide repeat disorders. *Annu. Rev. Neurosci.*, **30**, 575–621.
3. McMurray,C.T. (2010) Mechanisms of trinucleotide repeat instability during human development. *Nat. Rev. Genet.*, **11**, 786–799.
4. Giunti,P., Sweeney,M.G., Spadaro,M., Jodice,C., Novelletto,A., Malaspina,P., Frontali,M. and Harding,A.E. (1994) The trinucleotide repeat expansion on chromosome 6p (SCA1) in autosomal dominant cerebellar ataxias. *Brain*, **117**, 645–649.
5. Wells,R.D. and Warren,S.T (1998) In: *Genetic Instabilities and Hereditary Neurological Diseases*. Academic Press, San Diego, CA.
6. Pearson,C. and Sinden,R. (1998) Slipped strand DNA (S-DNA and SI-DNA), trinucleotide repeat instability and mismatch repair: A short review. In: Sarma,RH and Sarma,MH (eds). *Structure, Motion, Interaction and Expression of Biological Macromolecules*. SUNY Albany. Vol. **2**, pp. 191–207.
7. Pearson,C.E., Edamura,K.N. and Cleary,J.D. (2005) Repeat instability: mechanisms of dynamic mutations. *Nat. Rev. Genet.*, **6**, 729–742.
8. Mirkin,S.M. (2007) Expandable DNA repeats and human disease. *Nature*, **447**, 932–940.
9. Kovtun,I.V. and McMurray,C.T. (2008) Features of trinucleotide repeat instability in vivo. *Cell Res.*, **18**, 198–213.
10. Oberle,I., Rouseau,F., Heitz,D., Devys,D., Zengerling,S. and Mandel,J. (1991) Molecular-basis of the fragile-X syndrome and diagnostic applications. *Am. J. Hum. Genet.*, **49**, 76.
11. Campuzano,V., Montermini,L., Moltò,M.D., Pianese,L., Cossée,M., Cavalcanti,F., Monros,E., Rodius,F., Duclos,F., Monticelli,A. *et al.* (1996) Friedreich's Ataxia: Autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science*, **271**, 1423–1427.
12. Mirkin,S.M. (2006) DNA structures, repeat expansions and human hereditary disorders. *Curr. Opin. Struct. Biol.*, **16**, 351–358.
13. Yu,Z., Zhu,Y., Chen-Plotkin,A.S., Clay-Falcone,D., McCluskey,L., Elman,L., Kalb,R.G., Trojanowski,J.Q., Lee,V. M.-Y., Van Deerlin,V.M. *et al.* (2011) PolyQ repeat expansions in ATXN2 associated with ALS are CAA interrupted repeats. *PloS one*, **6**, e17951.
14. Zoghbi,H.Y. and Orr,H.T. (2000) Glutamine repeats and neurodegeneration. *Annu. Rev. Neurosci.*, **23**, 217–247.
15. Buchanan,L.E., Carr,J.K., Fluitt,A.M., Hoganson,A.J., Moran,S.D., de Pablo,J.J., Skinner,J.L. and Zanni,M.T. (2014) Structural motif of polyglutamine amyloid fibrils discerned with mixed-isotope infrared spectroscopy. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 5796–801.
16. Man,V.H., Roland,C. and Sagui,C. (2015) Structural determinants of polyglutamine protofibrils and crystallites. *ACS Chem. Neurosci.*, **6**, 632–645.
17. Zhang,Y., Man,V., Roland,C. and Sagui,C. (2016) Amyloid properties of asparagine and glutamine in Prion-like proteins. *ACS Chem. Neurosci.*, **7**, 576–587.
18. Wells,R.D., Dere,R., Hebert,M.L., Napierala,M. and Son,L.S. (2005) Advances in mechanisms of genetic instability related to hereditary neurological diseases. *Nucleic Acids Res.*, **33**, 3785–3798.
19. Kim,J.C. and Mirkin,S.M. (2013) The balancing act of DNA repeat expansions. *Curr. Opin. Genet. Dev.*, **23**, 280–288.
20. Cleary,J.P., Walsh,D.M., Hofmeister,J.J., Shankar,G.M., Kuskowski,M.A., Selkoe,D.J. and Ashe,K.H. (2005) Natural oligomers of the amyloid-β protein specifically disrupt cognitive function. *Nat. Neurosci.*, **8**, 79–84.
21. Dion,V. and Wilson,J.H. (2009) Instability and chromatin structure of expanded trinucleotide repeats. *Trends Genet.*, **25**, 288–297.

22. McMurray,C.T. (2008) Hijacking of the mismatch repair system to cause CAG expansion and cell death in neurodegenerative disease. *DNA Repair*, **7**, 1121–1134.
23. Kim,J.C., Harris,S.T., Dinter,T., Shah,K.A. and Mirkin,S.M. (2017) The role of break-induced replication in large-scale expansions of (CAG) n/(CTG) n repeats. *Nat. Struct. Mol. Biol.*, **24**, 55.
24. Schmidt,M.H. and Pearson,C.E. (2016) Disease-associated repeat instability and mismatch repair. *DNA Repair*, **38**, 117–126.
25. Mitas,M., Yu,A., Dill,J. and Haworth,I. (1995) The trinucleotide repeat sequence d(CGG)15 forms a Heat-Stable hairpin containing G(syn).G(anti) base Pairs. *Biochemistry*, **34**, 12803–12811.
26. Gacy,A.M., Goellner,G., Juranić,N., Macura,S. and McMurray,C.T. (1995) Trinucleotide repeats that expand in human disease form hairpin structures in vitro. *Cell*, **81**, 533–540.
27. Darlow,J.M. and Leach,D.R. (1995) The effects of trinucleotide repeats found in human inherited disorders on palindrome inviability in escherichia coli suggest hairpin folding preferences in vivo. *Genetics*, **141**, 825–832.
28. Petruska,J., Hartenstine,M.J. and Goodman,M.F. (1998) Analysis of strand slippage in DNA polymerase expansions of CAG/CTG triplet repeats associated with neurodegenerative disease. *J. Biol. Chem.*, **273**, 5204–5210.
29. Hartenstine,M.J., Goodman,M.F. and Petruska,J. (2000) Base stacking and even/odd behavior of hairpin loops in DNA triplet repeat slippage and expansion with DNA polymerase. *J. Biol. Chem.*, **275**, 18382–18390.
30. Figueroa,A.A., Cattie,D. and Delaney,S. (2011) Structure of even/odd trinucleotide repeat sequences modulates persistence of Non-B conformations and conversion to duplex. *Biochemistry*, **50**, 4441–4450.
31. Huang,J. and Delaney,S. (2016) Unique length-dependent biophysical properties of repetitive DNA. *J. Phys. Chem. B*, **120**, 4195–4203.
32. Cleary,J.D., Nichol,K., Wang,Y.H. and Pearson,C.E. (2002) Evidence of cis-acting factors in replication-mediated trinucleotide repeat instability in primate cells. *Nat. Genet.*, **1**, 37–46.
33. Napierala,M., Bacolla,A. and Wells,R.D. (2005) Increased negative superhelical density in vivo enhances the genetic instability of triplet repeat sequences. *J. Biol. Chem.*, **280**, 37366–37376.
34. 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.
35. Krzyzosiak,W.J., Sobczak,K., Wojciechowska,M., Fiszer,A., Mykowska,A. and Kozlowski,P. (2012) Triplet repeat RNA structure and its role as pathogenic agent and therapeutic target. *Nucleic Acids Res.*, **40**, 11–26.
36. Liu,G., Chen,X., B.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.
37. Axford,M.M., Wang,Y.-H., Nakamori,M., Zannis-Hadjopoulos,M., Thornton,C.A. and Pearson,C.E. (2013) Detection of slipped-DNAs at the trinucleotide repeats of the myotonic dystrophy type I disease locus in patient tissues. *PLoS genetics*, **9**, e1003866.
38. Pearson,C.E. and Sinden,R.R. (1996) Alternative structures in duplex DNA formed within the trinucleotide repeats of the myotonic dystrophy and fragile X loci. *Biochemistry*, **35**, 5041–5053.
39. Ni,C.-W., Wei,Y.-J., Shen,Y.-I. and Lee,I.-R. (2019) Long-range hairpin slippage reconfiguration dynamics in trinucleotide repeat sequences. *J. Phys. Chem. Lett.*, **10**, 3985–3990.
40. Kraus-Perrotta,C. and Lagalwar,S. (2016) Expansion, mosaicism and interruption: mechanisms of the CAG repeat mutation in spinocerebellar ataxia type 1. *Cerebellum Ataxias*, **3**, 20.
41. Chung,M.-Y., Ranum,L.P., Duvick,L.A., Servadio,A., Zoghbi,H.Y. and Orr,H.T. (1993) Evidence for a mechanism predisposing to intergenerational CAG repeat instability in spinocerebellar ataxia type I. *Nat. Genet.*, **5**, 254.
42. Chong,S.S., McCall,A.E., Cota,J., Subramony,S., Orr,H.T., Hughes,M.R. and Zoghbi,H.Y. (1995) Gametic and somatic tissue–specific heterogeneity of the expanded SCA1 CAG repeat in spinocerebellar ataxia type 1. *Nat. Genet.*, **10**, 344.
43. Eichler,E.E., Holden,J.J., Popovich,B.W., Reiss,A.L., Snow,K., Thibodeau,S.N., Richards,C.S., Ward,P.A. and Nelson,D.L. (1994) Length of uninterrupted CGG repeats determines instability in the FMR1 gene. *Nat. Genet.*, **8**, 88.
44. Kunst,C.B. and Warren,S.T. (1994) Cryptic and polar variation of the fragile X repeat could result in predisposing normal alleles. *Cell*, **77**, 853–861.
45. Zhong,N., Yang,W., Dobkin,C. and Brown,W.T. (1995) Fragile X gene instability: anchoring AGGs and linked microsatellites. *Am. J. Hum. Genet.*, **57**, 351.
46. Gao,R., Matsuura,T., Coolbaugh,M., Zühlke,C., Nakamura,K., Rasmussen,A., Siciliano,M.J., Ashizawa,T. and Lin,X. (2008) Instability of expanded CAG/CAA repeats in spinocerebellar ataxia type 17. *Eur. J. Hum. Genet.*, **16**, 215.
47. Massey,T.H. and Jones,L. (2018) The central role of DNA damage and repair in CAG repeat diseases. *Dis. Models Mech.*, **11**, dmm031930.
48. Charles,P., Camuzat,A., Benammar,N., Sellal,F., Destee,A., Bonnet,A., Lesage,S., Le Ber,I., Stevanin,G., Dürr,A. *et al.* (2007) Are interrupted SCA2 CAG repeat expansions responsible for parkinsonism? *Neurology*, **69**, 1970–1975.
49. Elden,A.C., Kim,H.-J., Hart,M.P., Chen-Plotkin,A.S., Johnson,B.S., Fang,X., Armakola,M., Geser,F., Greene,R., Lu,M.M. *et al.* (2010) Ataxin-2 intermediate-length polyglutamine expansions are associated with increased risk for ALS. *Nature*, **466**, 1069.
50. Choudhry,S., Mukerji,M., Srivastava,A.K., Jain,S. and Brahmachari,S.K. (2001) CAG repeat instability at SCA2 locus: anchoring CAA interruptions and linked single nucleotide polymorphisms. *Hum. Mol. Genet.*, **10**, 2437–2446.
51. Wright,G.E., Collins,J.A., Kay,C., McDonald,C., Dolzhenko,E., Xia,Q., Bečanović,K., Drögemöller,B.I., Semaka,A., Nguyen,C.M. *et al.* (2019) Length of uninterrupted CAG, independent of polyglutamine size, results in increased somatic instability, hastening onset of Huntington disease. *Am. J. Hum. Genet.*, **104**, 1116–1126.
52. Lee,J.-M., Correia,K., Loupe,J., Kim,K.-H., Barker,D., Hong,E.P., Chao,M.J., Long,J.D., Lucente,D., Vonsattel,J.P.G. *et al.* (2019) CAG repeat not polyglutamine length determines timing of Huntington's disease onset. *Cell*, **178**, 887–900.
53. Pearson,C.E., Eichler,E.E., Lorenzetti,D., Kramer,S.F., Zoghbi,H.Y., Nelson,D.L. and Sinden,R.R. (1998) Interruptions in the triplet repeats of SCA1 and FRAXA reduce the propensity and complexity of slipped strand DNA (S-DNA) formation. *Biochemistry*, **37**, 2701–2708.
54. Dorsman,J., Bremmer-Bout,M., Pepers,B., Ommen,G.-J.V. and Dunnen,J.D. (2002) Interruption of perfect CAG repeats by CAA triplets improves the stability of glutamine-encoding repeat sequences. *Biotechniques*, **33**, 976–978.
55. Rolfsmeier,M.L. and Lahue,R.S. (2000) Stabilizing effects of interruptions on trinucleotide repeat expansions in Saccharomyces cerevisiae. *Mol. Cell. Biol.*, **20**, 173–180.
56. Kiliszek,A., Kierzek,R., Krzyzosiak,W.J. and Rypniewski,W. (2010) Atomic resolution structure of CAG RNA repeats: structural insights and implications for the trinucleotide repeat expansion diseases. *Nuclear Acids Res.*, **38**, 8370–8376.
57. Yildirim,I., Park,H., Disney,M.D. and Schatz,G.C. (2013) A dynamic structural model of expanded RNA CAG Repeats: A refined X-ray structure and computational investigations using molecular dynamics and umbrella sampling simulations. *JACS*, **135**, 3528–3538.
58. Tawani,A. and Kumar,A. (2015) Structural insights reveal the dynamics of the repeating r(CAG) transcript found in Huntington's Disease (HD) and Spinocerebellar Ataxias (SCAs). *PLoS One*, **10**, e0131788.
59. Guo,P., Chan,H.Y.E. and Lam,S.L. (2017) Conformational flexibility in the RNA stem-loop structures formed by CAG repeats. *FEBS Lett.*, **591**, 1752–1760.
60. Pan,F., Man,V.H., Roland,C. and Sagui,C. (2017) Structure and dynamics of DNA and RNA double helices of CAG and GAC trinucleotide repeats. *Biophys. J.*, **113**, 19–36.
61. Tsukanov,R., Tomov,T.E., Berger,Y., Liber,M. and Nir,E. (2013) Conformational dynamics of DNA hairpins at millisecond resolution obtained from analysis of Single-Molecule FRET histograms. *J. Phys. Chem. B*, **117**, 16105–16109.
62. Tsukanov,R., Tomov,T., Masoud,R., Drory,H., Plavner,N., Liber,M. and Nir,E. (2013) Detailed study of DNA hairpin dynamics using single-molecule Fluorescence assisted by DNA origami. *J. Phys. Chem. B*, **117**, 11932.

63. Sass,L.E., Lanyi,C., Weninger,K. and Erie,D.A. (2010) Single-Molecule FRET TACKLE reveals highly dynamic mismatched DNA- MutS complexes. *Biochemistry*, **49**, 3174–3190.
64. McCann,J.J., Choi,U.B., Zheng,L., Weninger,K. and Bowen,M.E. (2010) Optimizing methods to recover absolute FRET efficiency from immobilized single molecules. *Biophys. J.*, **99**, 961.
65. LeBlanc,S.J., Gauer,J.W., Hao,P., Case,B.C., Hingorani,M.M., Weninger,K.R. and Erie,D.A. (2018) Coordinated protein and DNA conformational changes govern mismatch repair initiation by MutS. *Nucleic Acids Res.*, **46**, 10782–10795.
66. Case,D.A., Betz,R.M., Cerutti,D.S., Cheatham,T.E. III, Darden,T.A., Duke,R.E., Giese,T.J., Gohlke,H., Goetz,A.W., Homeyer,N. *et al.* (2016) *AMBER 16*. University of California, San Francisco.
67. Ivani,I., Dans,P.D., Noy,A., Pérez,A., Faustino,I., Hopsital,A., Walther,J., Andrio,P., Goñi,R., Balaceanu,A. *et al.* (2016) Parmbsc1: a refined force field for DNA simulations. *Nat. Methods*, **13**, 55–58.
68. Jorgensen,W.L., Chandrasekhar,J., Madura,J.D., Impey,R.W. and Klein,M.L. (1983) Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, **79**, 926–935.
69. Joung,I.S. and Cheatham,T.E. III (2008) Determination of Alkali and Halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B*, **112**, 9020–9041.
70. Essmann,U., Perera,L., Berkowitz,M.L., Darden,T., Lee,H. and Pedersen,L.G. (1995) A smooth particle mesh ewald method. *J. Chem. Phys.*, **103**, 8577–8593.
71. Gopich,I.V. and Szabo,A. (2010) FRET efficiency distributions of multistate single molecules. *J. Phys. Chem. B*, **114**, 15221–15226.
72. Darlow,J.M. and Leach,D.R. (1998) Secondary structures in d(CGG) and d(CCG) repeat tracts. *J. Mol. Biol.*, **275**, 3–16.
73. Darlow,J.M. and Leach,D.R. (1998) Evidence for two preferred hairpin folding patterns in d(CGG).d(CCG) repeat tracts in vivo. *J. Mol. Biol.*, **275**, 17–23.
74. Erie,D.A., Suri,A.K., Breslauer,K.J., Jones,R.A. and Olson,W.K. (1993) Theoretical predictions of DNA hairpin loop conformations: correlations with thermodynamic and spectroscopic data. *Biochemistry*, **32**, 436–454.
75. El Amri,C., Mauffret,O., Monnot,M., Tevanian,G., Lescot,E., Porumb,H. and Fermandjian,S. (1999) A DNA hairpin with a single residue loop closed by a strongly distorted Watson-Crick G x C Base-pair. *J. Mol. Biol.*, **294**, 427–442.
76. Wu,H., Yang,P., Butcher,S., Kang,S., Chanfreau,G. and Feigon,J. (2001) A novel family of RNA Tetraloop structure forms the recognition site for saccharomyces cerevisiae RNase III. *EMBO J.*, **20**, 7240–7249.
77. Butcher,S., Allain,F. and Feigon,J. (1999) Solution structure of the Loop B domain from the hairpin ribozyme. *Nat. Struct. Biol.*, **6**, 212–216.
78. Lebars,I., Lamontagne,B., Yoshizawa,S., Aboul-Elela,S. and Fourmy,D. (2001) Solution structure of conserved AGNN Tetraloops: Insights into Rnt1p RNA processing. *EMBO J.*, **20**, 7250–7258.
79. Mitchell,M.L., Leveille,M.P., Solecki,R.S., Tran,T. and Cannon,B. (2018) Sequence-Dependent effects of monovalent cations on the structural dynamics of Trinucleotide-Repeat DNA hairpins. *J. Phys. Chem. B*, **122**, 11841–11851.
80. Huang,T.-Y., Chang,C.-K., Kao,Y.-F., Chin,C.-H., Ni,C.-W., Hsu,H.-Y., Hu,N.-J., Hsieh,L.-C., Chou,S.-H., Lee,I.-R. *et al.* (2017) Parity-dependent hairpin configurations of repetitive DNA sequence promote slippage associated with DNA expansion. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, 9535–9540.
81. Volle,C.B., Jarem,D.A. and Delaney,S. (2011) Trinucleotide repeat DNA alters structure to minimize the thermodynamic impact of 8-oxo-7, 8-dihydroguanine. *Biochemistry*, **51**, 52–62.
82. Völker,J., Gindikin,V., Klump,H.H., Plum,G.E. and B.K,J. (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.
83. Iyer,R., Pluciennik,A., Napierala,M. and Wells,R. (2015) DNA triplet repeat expansion and mismatch repair. *Ann. Rev. Biochem.*, **84**, 199.