Emerging Accessibility Patterns in Long Telomeric Overhangs

Sajad Shiekh, Golam Mustafa, Mohammed Enamul Hoque, Eric Yokie, John J. Portman, Hamza Balci

Abstract: We present single molecule experimental and computational modeling studies investigating the accessibility and folding landscape of human telomeric overhangs of physiologically relevant lengths. The overhangs contain 4-28 repeats of GGGTTA (G-Tract) sequence and accommodate 1-7 tandem G-quadruplex (GQ) structures. Using FRET-PAINT, we probed the distribution of accessible sites via a short imager strand, which is complementary to a G-Tract and transiently binds to unfolded sites. We report accessibility patterns that periodically change with overhang length and provide insights about the underlying folding frustration. Overhangs that have 4n G-Tracts, (12, 16…), demonstrate maximum frustration, while those with 4n+2 G-Tracts, (14, 18…), have minimal frustration. We also developed a computational model that suggests positive folding cooperativity between neighboring GQs is required for persistence of such patterns. Our experimental and computational studies suggest lower folding stability at the junction between single and double stranded telomeric DNA, which has implications for Shelterin complex formation.

INTRODUCTION

Telomeres are specific DNA-protein complexes, located at the ends of linear chromosomes and play a vital role in promoting their integrity. They are involved in safeguarding the chromosome ends by differentiating them from DNA double-strand breaks, which would otherwise trigger unwanted DNA damage response [1,2]. Human telomeric DNA is composed of tandem repeats of hexanucleotide d(GGGTTA) arranged into a long double-stranded region followed by a 50-300 nucleotide (nt) long 3′ single-stranded overhang [3]. These unique telomeric repeats are conserved across vertebrates [4]. The length of telomeres is important for their protective functions [5,6]; however, the inability of DNA replication machinery to complete the replication and processing of chromosome ends causes progressive telomere shortening during successive cell divisions [7]. When a critical telomere length is reached, apoptosis or cell senescence is activated in somatic cells [8], but certain oncogenic cells continue to proliferate by overexpressing telomerase ribonucleoprotein complex which elongates telomeres, eventually leading to development of malignant tumors [6,9]. Telomerase includes a reverse transcriptase enzyme that synthesizes telomeric DNA repeats d(GGGTTA) using an internal RNA template [3,10].

The G-rich overhang allows human telomeric DNA to fold into compact non-canonical structures called G-quadruplex (GQ). GQ is formed when four repeats of d(GGGTTA) are arranged into planar quartet geometry, stabilized by Hoogsteen hydrogen bonding and centrally located monovalent cation. For brevity, the d(GGGTTA) sequence will be referred to as a ”G-Tract”. Potentially GQ forming sequences are concentrated in certain regions of human genome, including telomeres and promoters, and GQs have been confirmed to form in vitro [11,12] and in vivo [13–15]. GQs are implicated in crucial biological processes, such as transcription, recombination, and replication [16,17]. Telomeric GQs are reported to inhibit telomerase activity, preventing telomere elongation [18].

Most studies on GQs have typically focused on understanding the structure and function of “single” DNA or RNA GQ molecules [19–21] and their interactions with DNA binding proteins [22,23] and helicases [24–27]. However, the presence of many G-Tracts in human telomeric overhang allows formation of 5-10 tandem GQs with varying separations. How these tandem GQs interact with each other and their folding characteristics are important to understand their role in protecting chromosome ends. The studies on DNA constructs that form multiple telomeric GQs have not reached a consensus yet and concluded negligible interactions, stabilizing (positive cooperativity) stacking interactions, or destabilizing interactions (negative cooperativity) between neighboring GQs [28–33]. Therefore, systematic investigations on these critical genomic sequences are needed.

In this study, we investigated the accessibility patterns of telomeric single-stranded DNA (ssDNA) molecules that contain 4-28 G-Tracts, a total of 25 constructs, using single molecule FRET-PAINT (Förster Resonance Energy Transfer-Point Accumulation for Imaging in Nanoscale Topography) method [33]. The telomeric ssDNA was part of a partial-duplex DNA (pdDNA) construct that contained a double-stranded DNA (dsDNA) stem and single stranded telomeric overhang. This also mimics the physiological structure where the single stranded telomeric

overhang is connected to the double stranded telomeric region. The sequences we studied can form 1-7 tandem GQs or could have multiple unfolded G-Tracts at different segments of the overhang. Binding of a Cy5-labeled short complementary Peptide Nucleic Acid (PNA) strand to an unfolded G-Tract results in a FRET signal that depends on the position of the binding site within the overhang and the folding pattern between the binding site and the ssDNA/dsDNA junction, where Cy3 is located. The resulting FRET distributions represent the accessibility patterns of the overhang, which depend on the underlying GQ folding landscape.

We also developed a computational model that captured similar folding patterns when neighboring GQs interact with high positive cooperativity and the G-Tracts at the ssDNA/dsDNA junction have lower folding stability compared to intermediate G-Tracts and the G-Tracts at the free 3´-end of the overhang, which is in the vicinity of ssDNA/dsDNA junction.

RESULTS

We investigated the accessibility patterns of 4-28 G-Tract long telomeric overhangs using FRET-PAINT assay (Fig. 1A). FRET-PAINT combines attributes of super-resolution microscopy technique DNA-PAINT [34] with those of single-molecule FRET (smFRET) [35]. This methodology takes advantage of transient binding of a labeled probe (called imager strand) to a target strand. As the imager strand, we used an acceptor (Cy5) labeled PNA that is complementary to a G-Tract. The imager strand is introduced to a microfluidic chamber that contains surface-immobilized, donor (Cy3) labeled partial duplex DNA (pdDNA) constructs that have a single stranded telomeric overhang (Fig. 1A). The surface immobilized DNA is annealed and slowly cooled to room temperature to ensure the folding pattern reaches a thermodynamic steady-state. A FRET signal that depends on the position of the binding site is observed when Cy5-PNA transiently binds to an accessible (unfolded, not part of a GQ) G-Tract (Fig. 1B).

When fully folded, the longest telomeric overhang we studied (28G-Tract) can form up to seven tandem GQs. However, the number of folded GQs and how the unfolded G-Tracts are distributed would vary depending on the folding pattern. Fig. 2 shows the FRET distributions obtained from transient bindings of Cy5-PNA to unfolded G-Tracts in telomeric overhangs. The data are grouped in sets of four histograms, with the shortest construct in each set having an integer multiple of 4 G-Tracts, [4n], while the longest construct has an additional three G-Tracts, [4n+3]. The constructs that remain fully-folded during the measurement time and those whose unfolded sites are not accessible to the probe do not contribute to these histograms.

These FRET distributions reveal surprising accessibility patterns that emerge when the telomeric overhang reaches a length of about 10 G-Tracts. The patterns have a periodicity of 4 G-Tracts and persist in the 10-28 G-Tract range. While the [4n]G-Tract constructs have the broadest distributions, the [4n+2]G-Tract constructs are typically the narrowest. These narrower distributions are concentrated at high-FRET levels, which represent binding to sites closer to the ssDNA/dsDNA junction (5´-side). The broad distributions of [4n]G-Tract constructs suggest the binding sites, i.e. unfolded G-Tracts, are distributed throughout the overhang. Fig. 3A shows a contour plot of the combined data for all constructs that illustrates alternating broad (marked with

white dashed lines) and narrow (marked with white dotted lines) distributions. The same bin-size was used in this contour plot as that used in the histograms in Fig. 2. To quantify the broadness of FRET distributions, we defined an S-parameter which is analogous to Shannon entropy. The S-parameter describes the uncertainty associated with determining where the imager strand would bind (Fig. 3B). This uncertainty would be greater for broad distributions since the imager strand has similar probabilities to bind to different sites. Therefore, the maxima of S-parameter should occur at [4n]G-Tract constructs and minima at [4n+2]G-Tract constructs.

The accessibility patterns presented in Fig. 2 depend on the folding landscape of the telomeric overhang. We interpret the broad FRET distributions as an indicator for high folding frustration. Narrow distributions suggest different overhangs fold in a relatively consistent manner and the unfolded sites are concentrated at particular regions of the overhang, which are consistent with lower folding frustration. Our results show that [4n]G-Tract constructs have significantly higher frustration compared to other constructs. Adding 1-3 G-Tracts resolves this frustration to different extents, resulting in a concentration of unfolded sites in the vicinity of ssDNA/dsDNA junction, on the 5′-side. This implies the G-Tracts at 3′-end or intermediate regions have higher folding stability compared to those at 5′-side. The proximity of these sites to a duplex DNA, which mimics the way telomeric DNA is in a physiological setting, could be the reason for their reduced stability.

The [4n]G-Tract constructs have an exact number of G-Tracts that can accommodate n GQs. Therefore, attaining complete folding requires a perfect progression of the folding throughout the overhang. For these constructs, nucleation of GQ folding from multiple sites will likely result in leaving unfolded G-Tracts between neighboring GQs. Also, skipping of even one G-Tract during progression of the folding (two neighboring GQs separated by a single unfolded G-Tract) would result in three more unfolded G-Tracts at different regions of the overhang. We assume GQs with long loops that incorporate one or more G-Tracts (9 nt or longer loops instead of the canonical 3 nt long loops) are highly unlikely due to entropic considerations [32]. On the other hand, constructs with additional G-Tracts ([4n+1], [4n+2] and [4n+3]) are more tolerant for shifts in the G-Tract register during folding or initiation of folding from different sites. Assuming lower folding stability at 5′-side, nucleation of folding from 3′-end or intermediate regions would more likely result in unfolded G-Tracts at the 5′-side. Once a particular folding landscape is established, it persists for long periods of time due to the high stability of associated GQ structures, which prevents resolving the frustration in folding [32]. The observed patterns also imply relatively efficient progression of folding towards 5′-side without leaving many unfolded G-Tracts between neighboring GQs, which can be achieved with positive cooperativity between neighboring GQs during the folding process.

These observations have physiologically significant implications as the free 3′-ends of telomeres are more prone to degradation by exonuclease activity and folding into GQ helps protecting these ends [36]. In addition, both telomerase and alternative lengthening of telomere (ALT) mechanisms require unfolded 3′-ends for the extension of telomeres [37] and presence of such secondary structures could potentially render telomeres inaccessible. The ssDNA/dsDNA junction that is more likely to be unfolded is the region where POT1-TPP1 connect to members of Shelterin

complex that are localized on duplex DNA. Therefore, having unfolded G-Tracts at this region would facilitate binding of POT1 and establishing this connection. Interestingly, the observed accessibility pattern onsets when telomeric overhang reaches ~10G-Tract, which corresponds to ~60 nt. This is similar to the minimum telomere length (~50 nt) that is required before the cells undergo senescence or apoptosis. It is not currently known whether these repeating accessibility and folding patterns persist in the presence of Shelterin complex or play any role in telomere maintenance or organization; however, we speculate that they might be involved in sensing telomere length and maintaining a repeating pattern of telomere organization as the overhang shortens with successive cell divisions or as it is elongated by telomerase.

As described in the Methods section, we used a model [32] that estimates unfolding probability (P_i) of a G-Tract based on folding cooperativity (K_c) between neighboring GQs and the position of the G-Tract, which influences both combinatorial considerations and the statistical weight of being in the folded state, i.e. being one of the four G-Tracts that form a GQ. For a telomeric overhang with N G-Tracts, the stabilities of a GQ in the interior ($K_f = 60$), a GQ ending at the 3´-end ($K_{f,N} = 229$), and a GQ ending at the neighboring G-Tract ($K_{f,N-1} = 80$), were kept similar to those reported in a recent study by Mittermaier *et al.* due to similarity of our pdDNA construct to their ssDNA construct in the middle and 3´-ends [32]. Varying these parameters from these values resulted in similar results as long as $K_{f,N-1}$ was greater than K_f and $K_{f,N}$ was significantly greater than both $K_{f,N-1}$ and K_f. On the 5´ side, our constructs have a ssDNA/dsDNA junction rather than a free end as in ssDNA, which resulted in significantly lower stability for G-Tracts in this region. A similar destabilization due to flanking duplex regions was recently reported for the GQs that form in KIT promoter [38]. We estimated the folding parameters associated with the G-Tracts in the vicinity of ssDNA/dsDNA junction ($K_{f,1}$ and $K_{f,2}$) and the cooperativity parameter (K_c) by comparing the patterns observed for the experimentally measured and computationally calculated S-parameters, as described below.

The FRET signal measured in the experiment probes the location of unfolded G-Tracts. To compare the outcomes of the model with the experimental distribution measured as a function of telomere length, we calculate how delocalized the distribution of unfolded G-Tracts is for a given length via the S-parameter, as defined in the Methods section. We find that the model accommodates a variety of patterns of the calculated S-parameter as a function of telomere length depending on folding and cooperativity parameters in the model. Fig. 4 shows a phase diagram as a function of the stability for a GQ starting with the 5´ G-Tract, $K_{f,1}$, and the cooperativity K_c. In this figure, we choose the weight for a GQ starting at the second G-Tract from the junction to be twice that of starting at the first G-Tract, i.e. $K_{f,2} = 2K_{f,1}$. Similar patterns are observed as long as $K_{f,2}$ is greater than $K_{f,1}$ and significantly smaller than K_f (Fig. S2). For simplicity, we focus on a window composed of N=20 to N=24 G-Tracts to characterize the pattern of the S-parameter and monitor if the S-parameter increases or decreases as N increases within this range. With positive cooperativity, four patterns are observed for the S-parameters. The pattern observed in the FRET efficiency (colored green in Fig. 4) occurs within a specific (though rather broad) range for parameters $K_{f,1}$, $K_{f,2}$, and K_c. For $K_c > 6$, the stabilities of GQ starting near the ssDNA/dsDNA junction must be below an upper threshold to observe the appropriate pattern, and for smaller cooperativity, these stabilities must lie between a lower and upper bound. The S-parameter as a

function of N, shown in Fig. 4, is similar to that experimentally observed. Because of this similarity to the experimentally measured S-parameter, we will use parameters ($K_{f,1}$ = 5, $K_{f,2}$ = 10, and K_c = 5) as an illustration in what follows.

To get a feel for the stabilities suggested by this model, at T = 25 C, the stability of an interior GQ relative to an unfolded G-Tract is given by ΔG = -2.43 kcal/mol and a GQ involving the first G-Tract at the ssDNA/dsDNA junction is ΔG = -0.95 kcal/mol. The free energy of two adjacent GQ has free energy ΔG = -5.81 kcal/mol, of which -0.95 kcal/mol is associated with cooperative stabilizing interactions between the GQ's. Destabilization of GQs at the ssDNA/dsDNA junction and positive cooperativity are consistent with the high FRET peak observed in many constructs. The free energies for GQ stability calculated from our model are consistent with those reported in literature [39]; although, there is no consensus about whether the interaction between neighboring GQs is stabilizing or destabilizing. Several studies have investigated the interactions between neighboring GQs and possible higher-order structures that might form in telomeric overhangs using different constructs and experimental methods. However, these studies have not reached a consensus with conclusions varying from negligible stacking interactions (beads on a string type arrangement) [28] and destabilizing interactions (negative cooperativity) [32,40] to higher order structures with multiple GQs condensing into compact structures [29,33,41,42], which would suggest positive cooperativity. Our results are more consistent with these latter studies that propose positive cooperativity.

Fig. 5 shows that the parameters identified by comparison of the computational and experimental S-parameters correspond to very stabilized GQs. Although the strong stability weakens as the number of G-Tracts increases, the average number of folded GQs ($\langle n_{GQ} \rangle$) is nearly maximal for G-Tracts ranging from N = 4 to 28 (Fig. 5A). Nevertheless, constructs with 4n and 4n+1 G-Tracts have somewhat higher average number of unfolded G-Tracts ($\langle n_{uf} \rangle$) than the minimum associated with fully folded GQs (Fig. 5B). This is consistent with the experiment where most of telomeric constructs do not show any binding events during experimental observation time (~3 min). Also, constructs with 4n and 4n+1 G-Tracts have large relative fluctuations so that $\langle n_{uf} \rangle$ is not a sharp value for these constructs, suggesting that the number of sites available to PNA binding can be significantly larger than indicated by $\langle n_{uf} \rangle$. Even though the unfolded G-Tracts are only a small fraction of the telomeric overhang, they are nevertheless physiologically very significant as they might serve as the nucleation sites for triggering DNA damage response, nuclease activity, or telomerase-mediated telomere extension. Therefore, understanding how they are distributed, and their relative concentration is important.

Patterns in the S-parameter can be understood as balancing the population of unfolded G-Tracts within the destabilized junction region, and the rest of the telomeric G-Tracts given the underlying frustration of the landscape and the minimum number of unfolded G-Tracts when the length of the construct differs from 4n. Because the first two repeats near the junction at the 5′ side are destabilized compared to others, the unfolded distribution is shifted towards the first two repeats. As shown in Fig. 5C-D, the G-Tracts are mostly folded for N = 4n, with a weak pattern in unfolding propensity along the chain. For N = 4n+1, which has at least one unfolded G-Tract, the first G-Tract is much more likely unfolded, and the pattern for the unfolded G-Tract along the chain is more prominent. This localization of the unfolded G-Tracts is reflected in a decrease in

the S-parameter. For $N = 4n+2$, the first two G-Tracts are unfolded with high probability and the remaining unfolding propensity decreases along the chain. Thus, the distribution of P_i is more localized and results in a local minimum of the S-parameter. For $N = 4n + 3$, unfolding of the first three G-Tracts are more likely, though the propensity of unfolding for other G-Tracts increases, giving rise to a more delocalized distribution along the chain, and hence an increase in the S-parameter.

A similar analysis of the patterns of the S-parameters shows that the pattern observed in the experiment does not appear for negative cooperativity (Supplement Fig. S3). Also, looking at a broader window $N = 12 – 24$, we see that the pattern can shift from one type to another as a function of N near the boundaries of phases of the S-parameter (Supplementary Fig. S4). Furthermore, these kinds of patterns occur not only in the S-parameter which explicitly depends on the distribution of unfolded G-Tracts, but also in the entropy as a function of N (Supplementary Fig. S5).

Our results demonstrate repeating accessibility and folding frustration patterns for telomeric overhangs of physiological lengths. While overhangs with [4n]G-Tracts demonstrate elevated levels of frustration, overhangs with two additional G-Tracts, [4n+2]G-Tracts, show minimal frustration where most unfolded sites are concentrated in the vicinity of ssDNA/dsDNA junction. These folding patterns would help with protecting the free 3′-end against exonuclease activity and inhibit telomerase-mediated elongation while facilitating binding of POT1/TPP1 in the vicinity of ssDNA/dsDNA junction and connecting with the rest of the Shelterin complex.

METHODS

Nucleic Acid Constructs: DNA strands were purchased either from Eurofins Genomics or IDT DNA and their sequences are given in Table S1 and Table S2. The oligonucleotides were purified in-house using denaturing polyacrylamide gel electrophoresis (PAGE). The corresponding gel images are shown in Supplementary Fig. S1. HPLC purified Cy5-PNA strand was purchased from PNA-Bio Inc. The pdDNA constructs (Fig. 1A) were formed by annealing a long strand that contains the telomeric overhang (4-28 G-Tracts) with a short strand (18, 24, or 30nt) that has a biotin at 3′-end and Cy3 at the 5′-end (Table S2). The two strands were annealed in a thermal cycler in 150 mM KCl and 10 mM $MgCl_2$ at 95°C for 3 minutes followed by a slow cooling to 30 °C (5°C decrease every 3 minutes). This slow cooling process ensures attaining a thermodynamic steady state for the folding pattern. During annealing, excess long strand (500 nM for 100 nM of short strand) was used to ensure all biotinylated and Cy3-labeled short strands have a matching long strand. The unpaired long strands are washed out of channel after the DNA is immobilized on surface via biotin-streptavidin linker.

The smFRET assay and Imaging Setup: A home-built prism-type total internal reflection fluorescence (TIRF) microscope was used for these measurements following protocols described in earlier work [43]. The slides and coverslips were initially cleaned with 1M potassium hydroxide (KOH) and acetone, followed by piranha etching, surface functionalization by amino silane and surface passivation by polyethylene glycol (PEG). To reduce nonspecific binding, the surfaces were treated with a mixture of m-PEG-5kDa and biotin-PEG-5kDa in the ratio 40:1 followed by

another round of passivation with 333 Da PEG to increase the density of the PEG brush. The microfluidic chamber was created by sandwiching a PEGylated slide and a coverslip with double sided tape, followed by sealing the chamber with epoxy. The chamber was treated with 2% (v/v) Tween-20 to further reduce non-specific binding.

After washing the excess detergent from the chamber, 0.01 mg/mL streptavidin was incubated in the chamber for 2 minutes. The pdDNA samples were diluted to 10 pM in 150 mM KCl and 2 mM $MgCl_2$ and added to the chamber for 2-5 minutes resulting in surface density of ~300 molecules per imaging area (~50 µm × 100 µm). The excess or unbound DNA was removed from the chamber with 150 mM KCl + 2 mM $MgCl_2$. Unless otherwise specified, the imaging buffer contained 50 mM Tris-HCl (pH 7.5), 2 mM Trolox, 0.8 mg/mL glucose, 0.1 mg/mL glucose oxidase, 0.1 mg/mL bovine serum albumin (BSA), 2 mM $MgCl_2$, 150 mM KCl and 40 nM Cy5-PNA. The Cy5-PNA strand was heated to 85 °C for 10 minutes prior to adding it to the imaging buffer. 1500-2000 frame long movies with 100 ms frame integration time were recorded using an Andor Ixon EMCCD camera. A green laser beam (532 nm) was used to excite the donor fluorophore and the fluorescence signal was collected by an Olympus water objective (60x, 1.20 NA).

Data Analysis: A custom written C++ software was used to record and analyze the movies, and to generate single molecule time traces of donor and acceptor intensities. The time traces of each molecule were further analyzed by a custom MATLAB code to select the molecules that passed a single molecule screening test. The background was subtracted for each of these molecules based on the remnant donor and acceptor intensities after donor photobleaches. The FRET efficiency (E_{FRET}) was calculated using E_{FRET} = Acceptor Intensity/(Acceptor Intensity + Donor Intensity). The molecules that did not show any binding event formed a donor-only (DO) peak at FRET efficiency E_{FRET}=0.06, which was used as a reference to rescale the FRET distribution such that DO peak corresponds to E_{FRET}=0.00. After this rescaling, the DO peak was subtracted from the histograms. Therefore, the FRET histograms presented in Fig. 2 were generated from single molecules that show at least one Cy5-PNA binding event. The numbers of molecules that contributed to each histogram in Fig. 2 are presented in Table S3, but typically each histogram was constructed from binding events to 100-150 DNA molecules, resulting in several hundred binding events in each histogram.

The FRET distributions were normalized such that each molecule contributes equally to the histogram and the total population across the entire FRET range was normalized to 100%, i.e. the population of a particular FRET bin represents it's percent population. Because of this normalization, dividing the FRET distribution by 100 gives a distribution of probabilities for different FRET levels where the total probability is normalized to 1.0. Therefore, the probability p_i of a particular FRET level i (a particular FRET bin of width 0.02 FRET) is obtained by dividing population at that FRET level by 100. To quantify the spread in the FRET histograms, we define an S-parameter : $S = -\sum_i p_i \ln p_i$ where the summation is carried over the entire FRET range and p_i is the unfolded probability, normalized so that $\sum_{i=1}^{N} p_i = 1$. (Fig. 3). To calculate the S-

parameter predicted by the computational model, p_i is taken as the unfolded probability of ith G-Tract and the total probability is normalized so that $\sum_{i=1}^{N} p_i = 1$. The S-parameter is analogous to Shannon entropy of a particular FRET distribution, with broader distributions resulting in a larger S-parameter which represents greater uncertainty in determining the position of the binding site. These calculations were performed in Origin 2015.

Computational Model: We consider a simple model introduced by Mittermaier and co-workers consisting of *N* G-Tracts, each of which can either be unfolded or folded with three neighboring G-Tracts in a GQ [32]. The statistical weight for a state α with N_f^{α} folded GQs and N_{int}^{α} interfaces between nearest neighbor GQs is given by $\omega_\alpha = K_f^{N_f^{\alpha}} K_c^{N_{int}^{\alpha}}$, where K_f is the weight associated with a folded GQ, and K_c models interactions between adjacent GQs. Folding cooperativity between neighboring GQs can be positive ($K_c > 1$), negative ($K_c < 1$), or neutral ($K_c = 1$). In Mittermaier's study, the DNA constructs were in the form of single stranded DNA with symmetrical 3′ and 5′ ends. This study demonstrated higher folding stability for GQs beginning or ending on the two terminal G-Tracts on each end: initiating on the 1st and 2nd G-tract at the 5'-end, or ending on the (N-1)th and Nth G-Tract at the 3'-end. Compared to the internal G-Tracts: $K_{f,1}, K_{f,2}, K_{f,N-1} and\ K_{f,N}$ were greater than K_f for internal G-Tracts. In our study, we used a partial duplex DNA, which is a better representative of physiological telomeric DNA that includes a dsDNA/ssDNA junction at the 5′ side. Since the free 3′-end and interior regions of our constructs are identical to those of Mittermaier study, we used similar statistical weights for $K_f, K_{f,N-1} and\ K_{f,N}$ as those estimated in that study. However, the vicinity of the dsDNA/ssDNA junction was significantly different from the rest in our constructs, which required using much lower folding parameters. These considerations are implemented as $K_{f,N} > K_{f,N-1} > K_f > K_{f,2} > K_{f,1}$ in our study.

The probability for a state α is given by $p_\alpha = \omega_\alpha / Z$ where $Z = \sum_\alpha \omega_\alpha$ is the partition function. The total number of states can be written as $\Omega(N) = \sum_{N_f=0}^{\lfloor N/4 \rfloor} \frac{(N_f + N_{uf})!}{N_f!\, N_{uf}!}$, where the combinatorial factor is the number of states with N_f folded GQs and $N_{uf} = N - 4N_f$ unfolded G-Tracts. For telomeric overhangs of interest, $\Omega(N)$ is relatively small. For example, a telomere of 32 G-Tracts has only 16493 possible states, assuming long loops that contain one or more G-Tracts (such loops would be 9, 15, or 21 nt long compared to the canonical 3 nt loops) are not allowed due to their lower stability [32]. This allows the partition sum to be computed explicitly by generating all states and evaluating the statistical weight of each state in the sum. The probability that the ith G-Tract is unfolded is given by $P_i = \sum_\alpha \delta_{i,\alpha} p_\alpha$, where $\delta_{i,\alpha} = 1$ if the ith G-Tract is unfolded and $\delta_{i,\alpha} = 0$ if the G-Tract is in a folded GQ. Overall, average properties of the telomere can be computed, such as the average number of folded GQs, $\langle n_{GQ} \rangle = \sum_\alpha N_f^{\alpha} p_\alpha$, the average number of unfolded G-Tracts, $\langle n_{uf} \rangle = N - 4\langle n_{GQ} \rangle$, and the fluctuations of the unfolded G-Tracts, $\langle \Delta^2 n_{uf} \rangle = \langle n_{uf}^2 \rangle - \langle n_{uf} \rangle^2$.

For short chains, this model predicts a non-uniform unfolding probability along the sequence of G-Tracts under conditions strongly favoring folding [32]. Here, combinatorial factors associated

different ways in which specific G-Tracts can participate in GQ formation, as well as competing states with similar free energy conspire to produce patterns in P_i as a function of G-Tract index i; that is, some G-Tracts are more likely to be unfolded than others depending on their position. The experimental FRET distribution reflects binding of PNA to unfolded G-Tracts modeled by the distribution P_i which depends on the location of the G-Tract along the chain and the total number of repeats. This degree of localization of the unfolded G-Tracts is characterized by computing the S-parameter: $S = -\sum_{i=1}^{N} \tilde{P}_i \ln \tilde{P}_i$, where $\tilde{P}_i$ is the unfolded probability normalized to $\sum_{i=1}^{N} \tilde{P}_i = 1$.

References

1. Blackburn, E. H. & E.H., B. Structure and function of telomeres. *Nature* **350**, 569–573 (1991).
2. Blackburn, E. H., Greider, C. W. & Szostak, J. W. Telomeres and telomerase: the path from maize, Tetrahymena and yeast to human cancer and aging. *Nat Med* **12**, 1133–1138 (2006).
3. Lim, C. J. & Cech, T. R. Shaping human telomeres: from shelterin and CST complexes to telomeric chromatin organization. *Nature Reviews Molecular Cell Biology* vol. 22 283–298 (2021).
4. Meyne, J., Ratliff, R. L. & Moyzis, R. K. Conservation of the human telomere sequence (TTAGGG)(n) among vertebrates. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 7049–7053 (1989).
5. Cech, T. R. Beginning to Understand the End of the Chromosome. *Mod. Biopharm. Des. Dev. Optim.* **1**, 36–48 (2008).
6. Shay, J. W. Role of Telomeres and Telomerase in Aging and Cancer. *Cancer Discov.* **6**, 584–593 (2016).
7. Harley, C., Futcher, B. & Greider, C. Telomeres shorten during ageing of human fibroblasts. *Nature* **345**, 458–460 (1990).
8. Stewart, S. A. *et al.* Erosion of the telomeric single-strand overhang at replicative senescence. *Nat. Genet.* **33**, 492–496 (2003).
9. Wu, R. A., Upton, H. E., Vogan, J. M. & Collins, K. Telomerase mechanism of telomere synthesis. *Annual Review of Biochemistry* vol. 86 439–460 (2017).
10. Jafri, M. A., Ansari, S. A., Alqahtani, M. H. & Shay, J. W. *Roles of telomeres and telomerase in cancer, and advances in telomerase-targeted therapies*. vol. 8 69 (BioMed Central Ltd., 2016).
11. Henderson, E., Hardin, C. C., Walk, S. K., Tinoco, I. & Blackburn, E. H. Telomeric DNA oligonucleotides form novel intramolecular structures containing guanine·guanine base pairs. *Cell* **51**, 899–908 (1987).
12. Phan, A. T. & Mergny, J. L. Human telomeric DNA: G-quadruplex, i-motif and Watson-Crick double helix. *Nucleic Acids Res* **30**, 4618–4625 (2002).
13. Henderson, A. *et al.* Detection of G-quadruplex DNA in mammalian cells. *Nucleic Acids Res.* **42**, 860–869 (2014).
14. Di Antonio, M. *et al.* Single-molecule visualisation of DNA G-quadruplex formation in live cells. *Nat. Chem.* 1–6 (2020) doi:10.1038/s41557-020-0506-4.
15. Biffi, G., Tannahill, D., McCafferty, J. & Balasubramanian, S. Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat. Chem.* **5**, 182–186 (2013).
16. Maizels, N. G4-associated human diseases. *EMBO Rep* **16**, 910–922 (2015).
17. Neidle, S. Quadruplex Nucleic Acids as Novel Therapeutic Targets. *Journal of Medicinal Chemistry* vol. 59 5987–6011 (2016).
18. Zaug, A. J., Podell, E. R. & Cech, T. R. Human POT1 disrupts telomeric G-quadruplexes

allowing telomerase extension in vitro. *Proc. Natl. Acad. Sci. U. S. A.* **102**, (2005).

19. Banco, M. T. & Ferré-D'Amaré, A. R. The emerging structural complexity of G-quadruplex RNAs. *RNA* vol. 27 390–402 (2021).
20. Lane, A. N., Chaires, J. B., Gray, R. D. & Trent, J. O. Stability and kinetics of G-quadruplex structures. *Nucleic Acids Res* **36**, 5482–5515 (2008).
21. Tran, P. L., Mergny, J. L. & Alberti, P. Stability of telomeric G-quadruplexes. *Nucleic Acids Res* **39**, 3282–3294 (2011).
22. Ray, S. *et al.* RPA-mediated unfolding of systematically varying G-quadruplex structures. *Biophys. J.* **104**, 2235–2245 (2013).
23. Ray, S., Bandaria, J. N., Qureshi, M. H., Yildiz, A. & Balci, H. G-quadruplex formation in telomeres enhances POT1/TPP1 protection against RPA binding. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 2990–5 (2014).
24. Budhathoki, J. B. *et al.* ATP-dependent G-quadruplex unfolding by Bloom helicase exhibits low processivity. *Nucleic Acids Res.* **43**, 5961–5970 (2015).
25. Maleki, P. *et al.* Quantifying the impact of small molecule ligands on G-quadruplex stability against Bloom helicase. *Nucleic Acids Res.* **47**, 10744–10753 (2019).
26. Chen, M. C. *et al.* Structural basis of G-quadruplex unfolding by the DEAH/RHA helicase DHX36. *Nature* **558**, 465–469 (2018).
27. Mendoza, O., Bourdoncle, A., Boule, J. B., Brosh Jr., R. M. & Mergny, J. L. G-quadruplexes and helicases. *Nucleic Acids Res* **44**, 1989–2006 (2016).
28. Yu, H. Q., Miyoshi, D. & Sugimoto, N. Characterization of structure and stability of long telomeric DNA G-quadruplexes. **128**, 15461–15468 (2006).
29. Petraccone, L. *et al.* Structure and stability of higher-order human telomeric quadruplexes. *J. Am. Chem. Soc.* **133**, 20951–20961 (2011).
30. Kar, A., Jones, N., Ozlem Arat, N., Fishel, R. & Griffith, J. D. Long repeating (TTAGGG)n single-stranded DNA self-condenses into compact beaded filaments stabilized by G-quadruplex formation. *J. Biol. Chem.* **293**, 9473–9485 (2018).
31. Abraham Punnoose, J. *et al.* Interaction of G-quadruplexes in the full-length 3' human telomeric overhang. *J Am Chem Soc* **136**, 18062–18069 (2014).
32. Carrino, S., Hennecker, C. D., Murrieta, A. C. & Mittermaier, A. Frustrated folding of guanine quadruplexes in telomeric DNA. *Nucleic Acids Res.* **49**, 3063–3076 (2021).
33. Mustafa, G., Shiekh, S., Gc, K., Abeysirigunawardena, S. & Balci, H. Interrogating accessibility of telomeric sequences with FRET-PAINT: Evidence for length-dependent telomere compaction. *Nucleic Acids Res.* **49**, 3371–3380 (2021).
34. Jungmann, R. *et al.* Single-molecule kinetics and super-resolution microscopy by fluorescence imaging of transient binding on DNA origami. *Nano Lett.* **10**, 4756–4761 (2010).
35. Auer, A., Strauss, M. T., Schlichthaerle, T. & Jungmann, R. Fast, Background-Free DNA-PAINT Imaging Using FRET-Based Probes. *Nano Lett.* **17**, 6428–6434 (2017).

36. Tang, J. *et al.* G-quadruplex preferentially forms at the very 3′ end of vertebrate telomeric DNA. *Nucleic Acids Res.* **36**, 1200–1208 (2008).

37. Wang, Q. *et al.* G-quadruplex formation at the 3' end of telomere DNA inhibits its extension by telomerase, polymerase and unwinding by helicase. *Nucleic Acids Res* **39**, 6229–6237 (2011).

38. Vesco, G. *et al.* Double-stranded flanking ends affect the folding kinetics and conformational equilibrium of G-quadruplexes forming sequences within the promoter of KIT oncogene. *Nucleic Acids Res.* **49**, 9724–9737 (2021).

39. Chaires, J. B. Human telomeric G-quadruplex: thermodynamic and kinetic studies of telomeric quadruplex stability. *FEBS J* **277**, 1098–1106 (2010).

40. Vorlíčková, M., Chládková, J., Kejnovská, I., Fialová, M. & Kypr, J. Guanine tetraplex topology of human telomere DNA is governed by the number of (TTAGGG) repeats. *Nucleic Acids Res.* **33**, 5851–5860 (2005).

41. Xu, Y., Ishizuka, T., Kurabayashi, K. & Komiyama, M. Consecutive formation of G-Quadruplexes in human telomerieoverhang DNA: A protective capping structure for telomere ends. *Angew. Chemie - Int. Ed.* **48**, 7833–7836 (2009).

42. Wang, H., Nora, G. J., Ghodke, H. & Opresko, P. L. Single molecule studies of physiologically relevant telomeric tails reveal POT1 mechanism for promoting G-quadruplex unfolding. **286**, 7479–7489 (2011).

43. Maleki, P., Budhathoki, J. B., Roy, W. A. & Balci, H. A practical guide to studying G-quadruplex structures using single-molecule FRET. *Mol. Genet. Genomics* **292**, 483–498 (2017).

Acknowledgements: This work was supported by NIH (1R15GM123443 to H.B.). We thank Dr. Soumitra Basu for use of his lab to perform the PAGE experiments.

Author Information:

These authors contributed equally: Sajad Shiekh and Golam Mustafa.

Affiliations:

Department of Physics, Kent State University, Kent, Ohio, United States;
Sajad Shiekh, Golam Mustafa, Eric Yokie, John J. Portman, Hamza Balci

Department of Chemistry & Biochemistry, Kent State University, Kent, Ohio, United States;
Mohammed Enamul Hoque

Contributions: S.S., J.J.P. and H.B. designed research. S.S. and G.M. performed FRET-PAINT experiments and analyzed the data. M.E.H. purified the oligonucleotides. E.Y. wrote a Matlab code for data analysis. J.J.P. performed computational research and analyzed data. S.S., J.J.P, and H.B. wrote the manuscript.

Corresponding authors:
Correspondence to Hamza Balci or John J. Portman.

Ethics declarations
Competing Interests
The authors declare no competing interests.

Figures

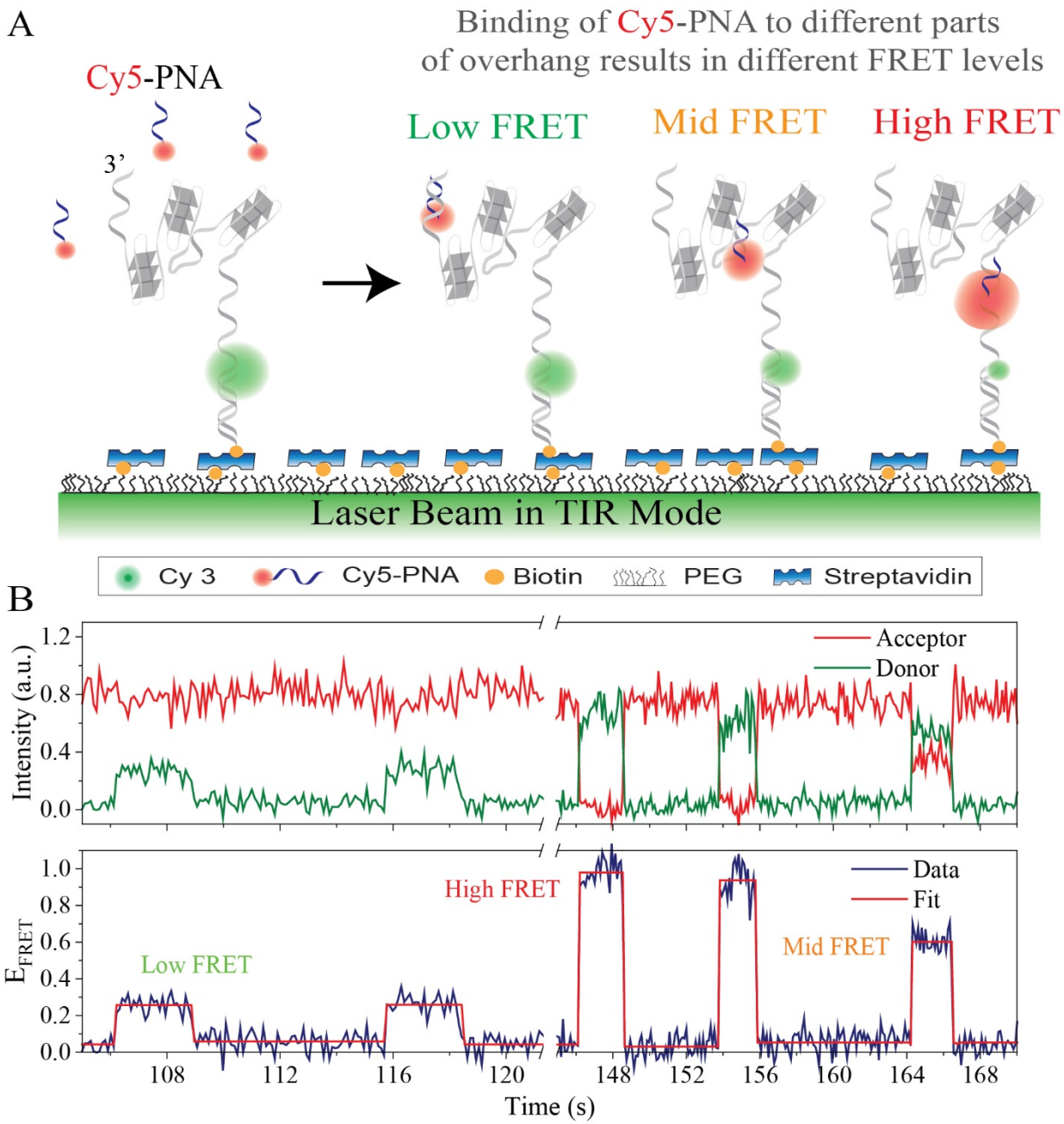

Figure 1: Schematic of FRET-PAINT assay. (A) Partial duplex DNA (pdDNA) construct with donor fluorophore (Cy3) is immobilized on the PEGylated surface via biotin-streptavidin attachment. The pdDNA construct has a telomeric overhang with multiple GGGTTA (G-Tract) repeats, which can form GQs separated from each other with unfolded regions of varying length. FRET signals with varying levels are observed when the imager strand (Cy5-PNA), which is complementary to a G-Tract, transiently binds to different segments of the overhang. (B) An example smFRET time trace showing five binding events with low, mid, and high FRET levels. The red line is a fit to the FRET trace to indicate the corresponding FRET levels.

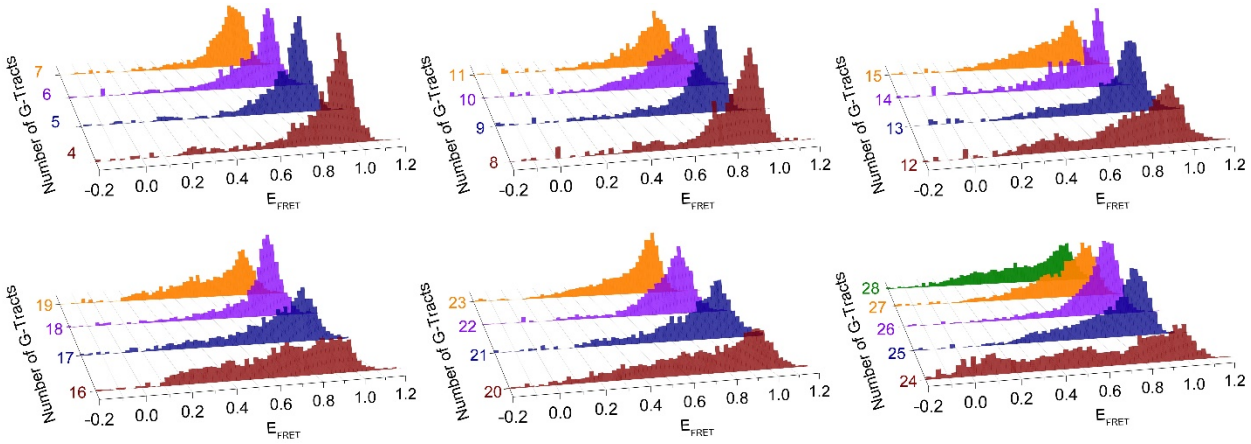

Figure 2: Normalized FRET histograms for telomeric overhangs with 4-28 G-Tracts. The histograms are grouped in sets of 4 with [4n]G-Tracts in brown, [4n+1]G-Tracts in blue, [4n+2]G-Tracts in violet and [4n+3]G-Tracts in orange. The 28G-Tract construct is added to the last histogram in green. The FRET histograms, which represent accessibility patterns, are broadest in [4n]G-Tract constructs and narrowest in [4n+2]G-Tract constructs.

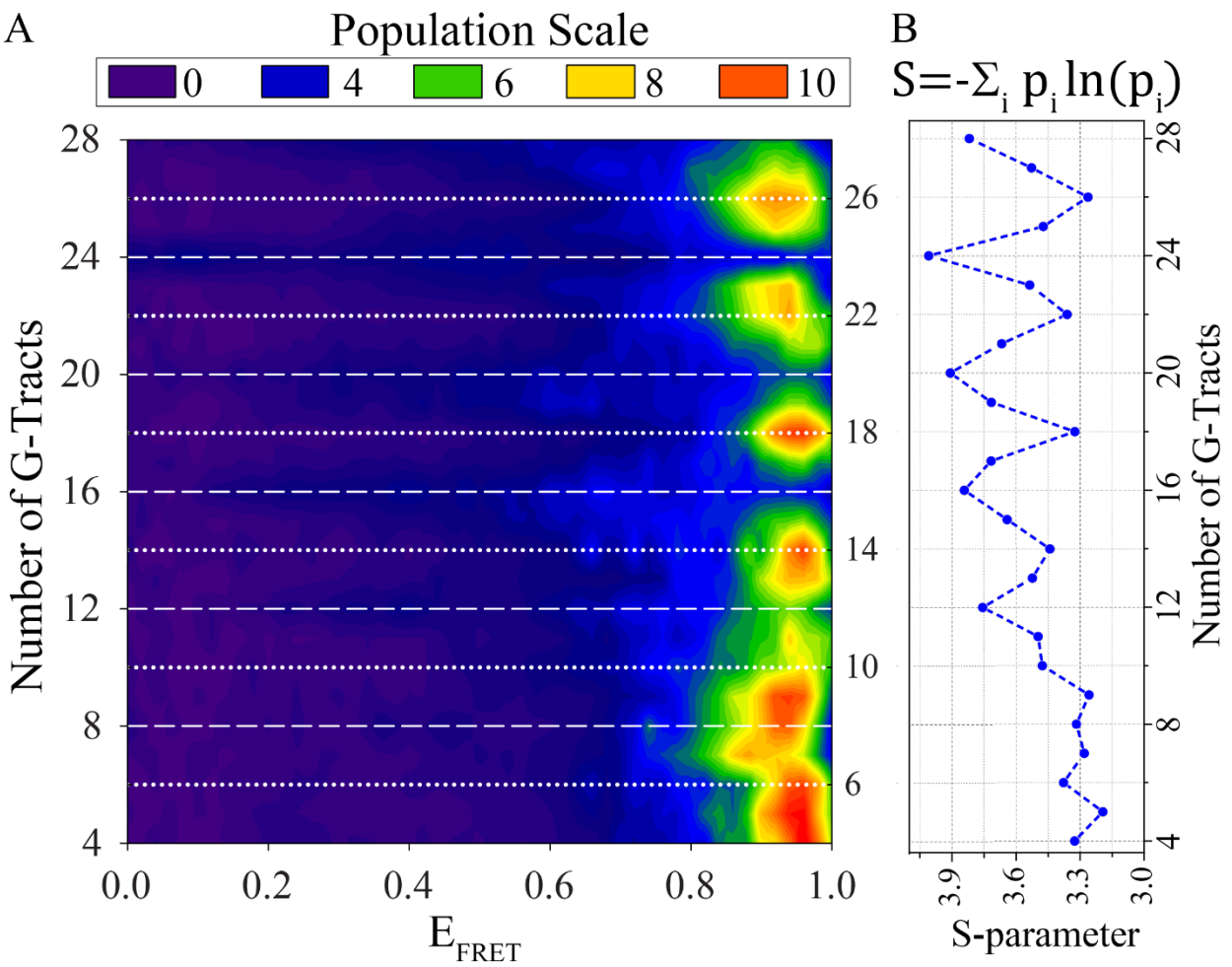

Figure 3: Contour plot of accessibility patterns and variation of S-parameter with overhang length. (A) For each of the histograms in Fig. 2, the FRET range was divided into five segments each of width 0.2 FRET and the cumulative population in each segment is demonstrated in the contour plot. The overhangs with [4n]G-Tracts are marked with dashed white lines while those with [4n+2]G-Tracts are marked with dotted white lines. (B) To quantify the broadness of the histograms in Fig. 2, an S-parameter (defined on top) is introduced. S-parameter is analogous to Shannon entropy and is a measure of folding frustration: the larger the S-parameter the greater the folding frustration. At lengths longer than 12G-Tract, a pattern is established where the maxima of S-parameter occur at [4n]G-Tract constructs (maximum frustration) while minima are at [4n+2]G-Tract (minimum frustration).

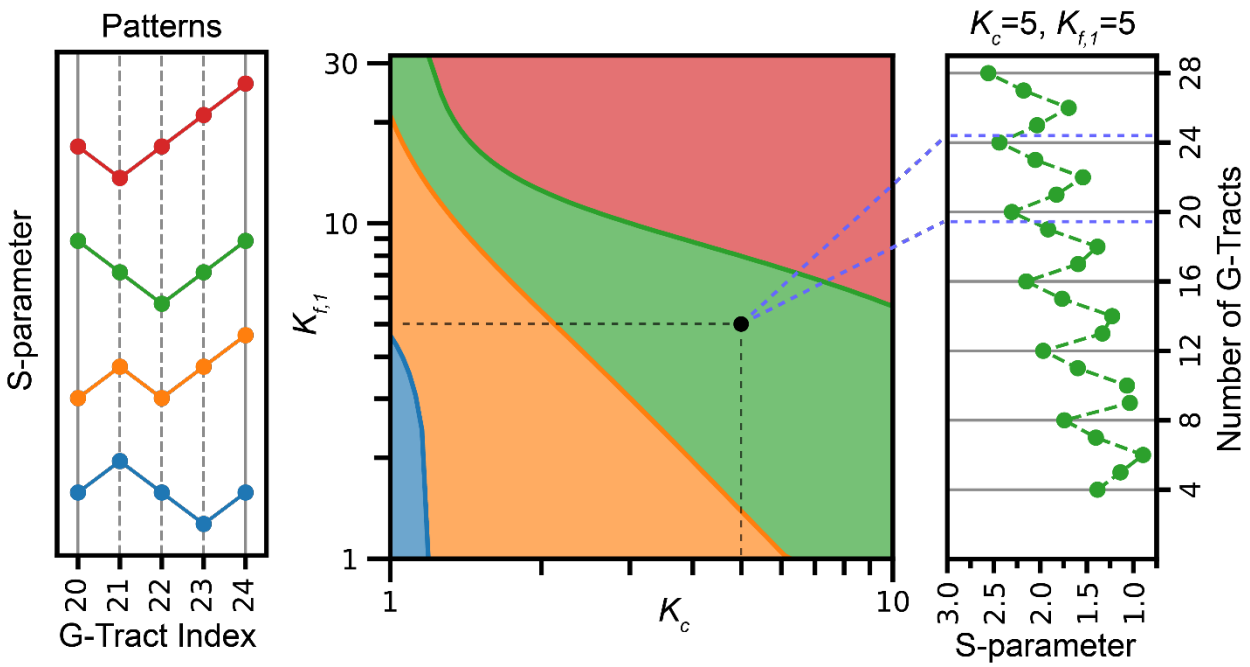

Figure 4. Computational patterns observed for the S-parameter for N=20-24 G-Tracts for varying $K_{f,1}$ and K_c parameters. In these calculations, $K_{f,1}$ and K_c are varied while $K_{f,2}=2K_{f,1}$, $K_f=60$, $K_{f,N-1}=80$, and $K_{f,N}=229$. The left panel shows the four patterns observed depending on whether the S-parameter increases or decreases as the length of the overhang is increased by one G-Tract. The phase diagram in the middle shows the parameter range where different patterns are observed in colors that match the patterns on the left panel. The green range corresponds to the pattern observed in the experiment. The right panel shows the calculated S-parameters for $K_c=5$ and $K_{f,1}=5$, which can be compared to the experimentally observed pattern in Fig. 3B.

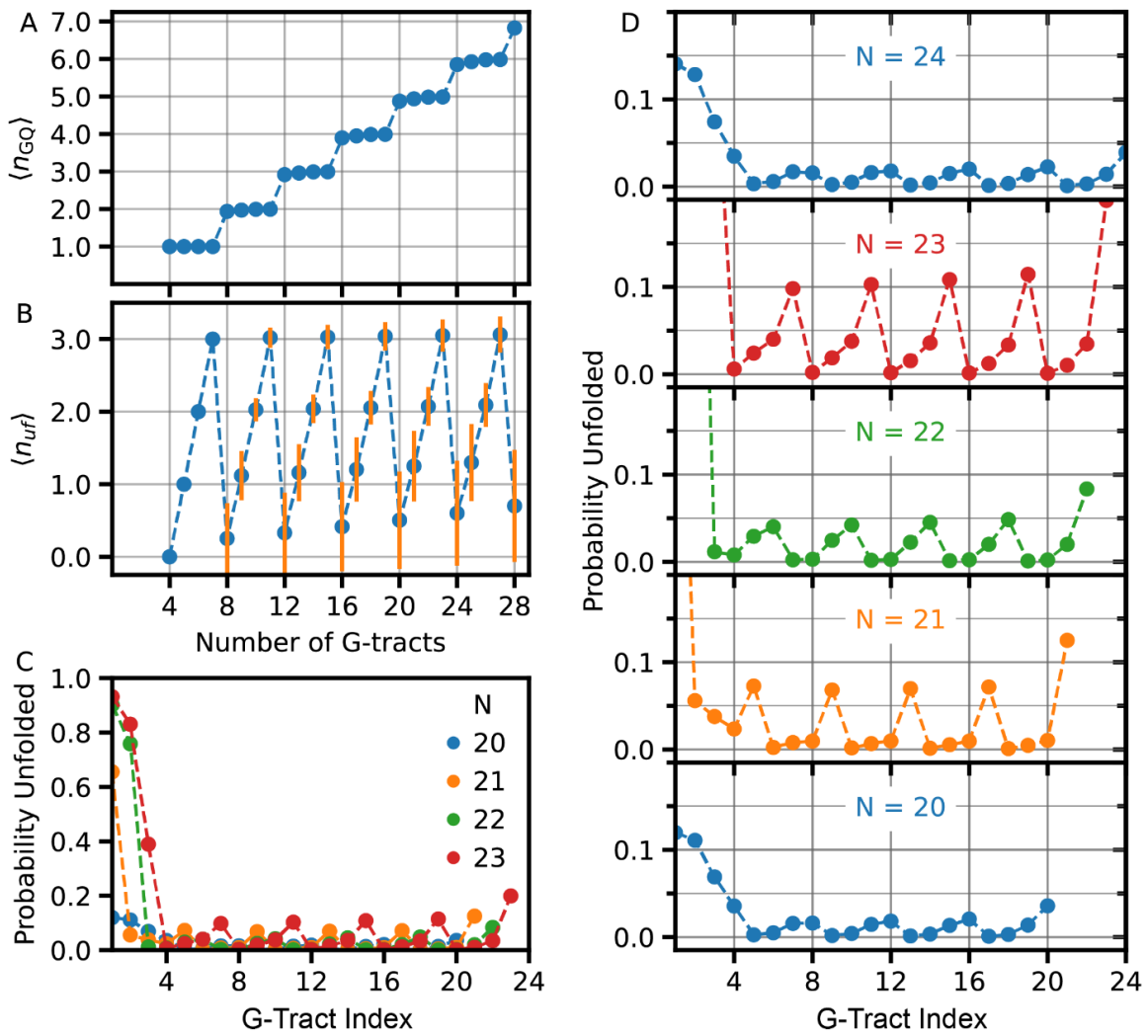

Figure 5. Folding levels for the set of parameters identified in Fig. 4: $K_{f,1}$=5, $K_{f,2}$=10, K_f=60, $K_{f,N-1}$ =80, $K_{f,N}$=229 and K_c=5. These parameters result in high levels of folding for all constructs investigated although periodic patterns were observed for probability of being unfolded depending on the position of the G-Tract and length of overhang. (A) For the set of parameters used in these studies, the average number of GQ ($\langle n_{GQ} \rangle$) is close to the maximum possible throughout 4-28 G-Tract overhang length. (B) Average number of unfolded G-Tracts ($\langle n_{uf} \rangle$) as a function of overhang length. As expected, constructs with 4n+1, 4n+2, and 4n+3 G-Tracts have at least 1, 2, and 3 unfolded G-Tracts, respectively. As can be most clearly seen for 4n constructs, $\langle n_{uf} \rangle$ increases as the overhang length increases. The orange vertical lines are standard deviation in $\langle n_{uf} \rangle$. (C) Probability of ith G-Tract being unfolded as a function its position for overhangs with N=20-24 G-Tracts. In this diagram i=1 for the G-Tract at ssDNA/dsDNA junction and i=*N* for the G-Tract at 3′-end. (D) The plots that are overlaid in (C) are separated for each N, and the y-axis scale is limited to unfolding probability in 0.0-0.2 range ($0.0<P_i<0.2$) to better illustrate the oscillations in the unfolding probability. The amplitude of oscillations in P_i vary depending overhang length, which can be related to the patterns observed for S-parameter.