Published online 19 April 2016

Nucleic Acids Research, 2016, Vol. 44, No. 10 ***4807–4817***
doi: 10.1093/nar/gkw272

Quantitative analysis and prediction of G-quadruplex forming sequences in double-stranded DNA

Minji Kim[1,2,†], Alex Kreig[3,†], Chun-Ying Lee[4], H. Tomas Rube[2,5], Jacob Calvert[3,6], Jun S. Song[2,3,7,8,*] and Sua Myong[2,3,4,8,*]

[1]Department of Electrical and Computer Engineering, University of Illinois; 306 N. Wright St. Urbana, IL 61801, USA, [2]Institute for Genomic Biology; 1206 Gregory Drive, Urbana, IL 61801, USA, [3]Department of Bioengineering, University of Illinois; 1304 W. Springfield Ave. Urbana, IL 61801, USA, [4]Department of Biophysics, Johns Hopkins University; 3400 N. Charles St. Baltimore, MD 21218 USA, [5]Department of Biological Sciences, Columbia University, New York, New York 10027, USA, [6]School of Mathematics, University of Bristol; University Walk, Bristol BS8 1TW, UK, [7]Department of Physics, University of Illinois; 1110 West Green Street, Urbana, IL 61801-3080, USA and [8]Physics Frontier Center (Center for Physics of Living Cells), University of Illinois, 1110 W. Green St. Urbana, IL 61801, USA

Received February 19, 2016; Revised March 30, 2016; Accepted April 05, 2016

ABSTRACT

G-quadruplex (GQ) is a four-stranded DNA structure that can be formed in guanine-rich sequences. GQ structures have been proposed to regulate diverse biological processes including transcription, replication, translation and telomere maintenance. Recent studies have demonstrated the existence of GQ DNA in live mammalian cells and a significant number of potential GQ forming sequences in the human genome. We present a systematic and quantitative analysis of GQ folding propensity on a large set of 438 GQ forming sequences in double-stranded DNA by integrating fluorescence measurement, single-molecule imaging and computational modeling. We find that short minimum loop length and the thymine base are two main factors that lead to high GQ folding propensity. Linear and Gaussian process regression models further validate that the GQ folding potential can be predicted with high accuracy based on the loop length distribution and the nucleotide content of the loop sequences. Our study provides important new parameters that can inform the evaluation and classification of putative GQ sequences in the human genome.

INTRODUCTION

The G-quadruplex (GQ) is a noncanonical DNA secondary structure arising from two or more stacked sets of four guanine (G) nucleotides (G-tetrads) interacting in a plane (Figure 1A), although three G-tetrads comprise the most common form in which the four sets of guanine triplets form a four-stranded structure through Hoogsteen base pairing coordinated by monovalent cations. GQ DNA can assume various folding configurations including parallel, antiparallel and hybrid conformations dictated by ion conditions and loop sequence compositions (1–4). A surge of interest in the GQ structure has followed the recent findings, suggesting its multifaceted role in key processes within the central dogma of biology (5–12). In particular, it is hypothesized that the formation of GQs modulates gene expression through a physical interaction between the GQ structure and transcription-related protein complexes (13). In support, recent work has confirmed the capability of GQs to form stably within the genome (14,15). Thus, GQs may prove to be an important component in the regulation of specific genes and, as such, may serve as an effective pharmaceutical target for a wide range of diseases (16–19). Putative GQ forming sequences are unevenly distributed throughout the human genome, with their presence increased in select gene regulatory regions, such as promoters of oncogenes and immunoglobulin switch regions (20,21). This irregular distribution highlights the challenge in identifying functional sequences that can actually form GQ structures *in vivo*.

GQ forming sequences are frequently modeled following the pattern $GGGN_{L1}GGGN_{L2}GGGN_{L3}GGG$, where N can be adenine (A), cytosine (C) or thymine (T), and $L1$, $L2$ and $L3$ are positive integers indicating the lengths of the intervening sequences that correspond to loops in the folded GQ structure (Figure 1A) (4). We note that loops can contain G bases, although we do not consider this possibility

*To whom correspondence should be addressed. Sua Myong. Tel: +1 410 5165122; Fax: +1 470 5164118; Email: smyong@jhu.edu
Correspondence may also be addressed to Jun S. Song. Tel: +1 217 2447750; Fax: +1 217 2442496; Email: songj@illinois.edu
† These authors contributed equally to the paper.

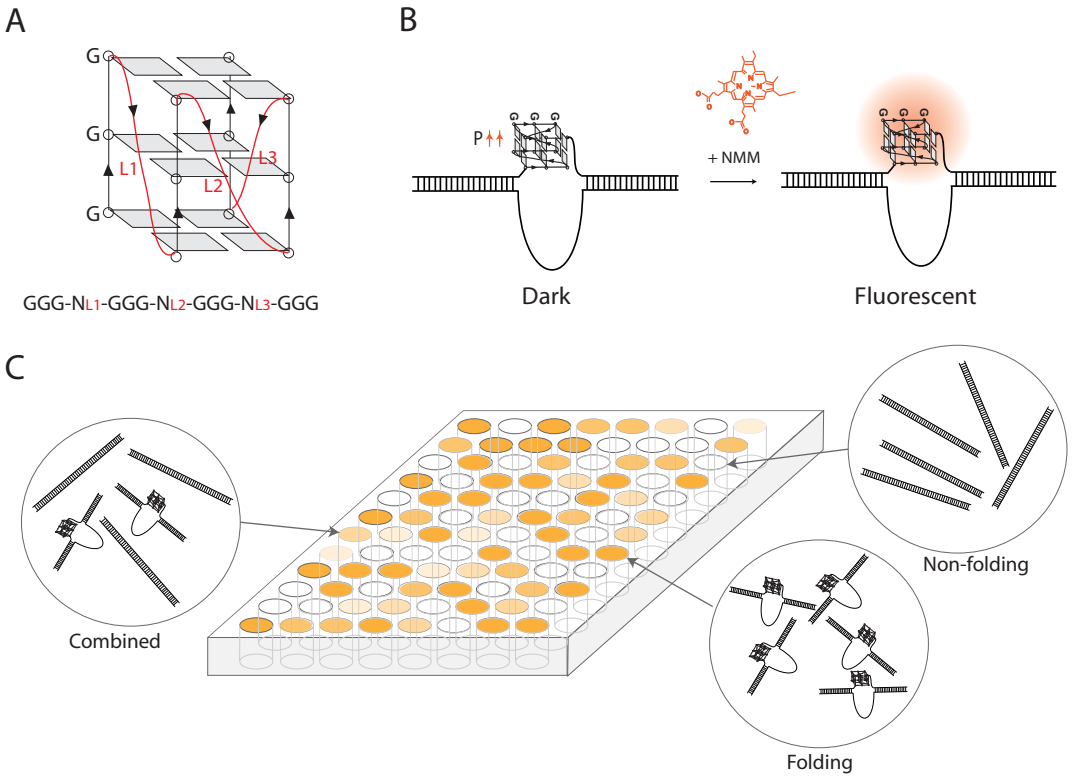

Figure 1. An overview of G-quadruplex structure and the NMM technique. (**A**) A schematic of a parallel GQ structure is depicted. The guanine–guanine Hoogsteen base pairing between each guanine triplet is shown for the sequence $GGGN_{L1}GGGN_{L2}GGGN_{L3}GGG$, where N denotes the nucleotide component and L1, L2, L3 are the three loop lengths. (**B**) GQ folding propensity is investigated through an induced fluorescence based assay. The molecule NMM shows a specific increase in fluorescence signal upon binding to a parallel GQ sequence. (**C**) A plate is filled with strong folding sequences in high intensity, combined folding and nonfolding sequences in a lower intensity, and nonfolding sequences in low intensity.

in our current study. Typical upper limits on loop length have been suggested to be between 7 and 9 bases within a single-stranded DNA (ssDNA) context, but a maximal loop length has not yet been established in a double-stranded DNA (dsDNA) context (22–25). Even with such restricted pattern assumptions, determining how nucleotide content and intervening loop lengths control the GQ formation potential of more than 400,000 candidate genomic sequences remains a challenging task. This ambiguity in GQ characterization complicates the identification of true GQ forming sequences implicated in essential biological activities.

The discovery of stable genomic GQ formation coupled with the significant number of potential GQ sequences located within the human genome underscores the need for new tools that can accurately predict folding propensity. Owing to the seemingly regular pattern found in GQ forming sequences, many bioinformatics studies have been conducted on putative GQ sequences (26–29). Generally, these studies simply searched for recurring patterns of putative GQs or developed models describing folding propensity based on GQ experiments in ssDNA. As a result, the methods may be biased toward known patterns and miss novel GQ folding sequences. Previously, we showed that the GQ folding propensity is substantially diminished in dsDNA and that, unlike ssDNA, dsDNA has limited ability to form only into parallel GQs (30). These considerations highlight the need for a new model that can predict GQ folding propensity specifically in a dsDNA context, which is more representative of genomic DNA than ssDNA.

We performed a survey of systematically designed GQ forming sequences to identify folding propensity within a dsDNA context. The survey contained more than four hundred putative GQ forming sequences with loops composed entirely of A, C or T with total loop length ranging up to 12 bp. Quantitative measurement of parallel GQ formation was obtained by N-methyl mesoporphyrin IX (NMM) fluorescence assay that was established in our previous work (30). The NMM intensity measurements were complemented by single-molecule fluorescence resonance energy transfer (smFRET) experiments, which enable direct quantitation of molecules comprising both the GQ-folded and unfolded populations (Figure 1B). We utilized these complementary methods to categorize each sequence as one of 'strongly folding,' 'nonfolding' or 'combined' classes, providing a simple metric for comparing the folding propensities of specific putative GQ sequences. Furthermore, by analyzing the impact of loop lengths and compositions on the NMM intensity measurement, we identified GQ-driving loop parameters. These results were combined in regression models that can predict GQ folding propensity with high accuracy. Our GQ folding experimental platform and computational models will serve as a useful reference that facilitates the investigation of potential genomic GQs in the future.

MATERIALS AND METHODS

Preparation of DNA

The oligonucleotide for GQ DNA strands and their complements were purchased unmodified from Integrated DNA Technologies (IDT). Each GQ strand was constructed with a unique 18 mer overhang present on both the 5′ and 3′ ends of the GQ. Annealing of complementary DNA pairs was conducted in a 1:1 molar ratio at 10 μM concentration for the GQ strand and its complement. Standard GQ DNA buffer containing 20 mM Tris-HCl pH 7.5, 100 mM KCl was supplemented with 40% (v/v) PEG 200 (Sigma Aldrich) to induce GQ formation within the dsDNA construct. Annealing was initiated by incubating samples at 95°C for 5 min and then cooling 2°C per min to room temperature (24 ± 1°C). For single molecule experiments, the same sequences as above were purchased containing an amine-modified thymine located 3 or 4 bases from the GQ forming region. Constructs were labeled by incubating 10 mM Cy3 or Cy5-NHS ester (GE Lifesciences) with 0.1 mM DNA in 100 mM sodium bicarbonate pH 8.5 buffer for 4–5 hours.

NMM GQ measurements

A final concentration of NMM 1μM (Frontier Scientific) was mixed with 500 nM dsDNA samples in standard GQ buffer. Final imaging conditions contained 4% PEG 200 (v/v). Samples were loaded into an optically clear 96-well plate (Nunc), and fluorescence measurements were conducted on a Gemini EM microplate reader (Molecular Devices). Emission measurements were taken at 609 nm while being excited at 570 nm.

Single-molecule imaging

Single-molecule fluorescence experiments were performed in channels made from glass coverslips on quartz slides (Finkenbeiner). To prevent DNA–surface interactions, slides and coverslips were coated with 97% methyl-PEG (m-PEG-5000, Laysan Bio, Inc.) and 3% biotin PEG (biotin-PEG-5000, Laysan Bio, Inc.). Biotinylated single-molecule DNA constructs were immobilized to the slide surface through biotin–neutravidin interactions (31). Imaging buffer was flowed through the chamber to wash out unbound molecules and remove residual PEG 200. Total internal reflection microscopy (TIRF) was utilized to collect single-molecule FRET traces. The evanescent field of illumination was created with a 532-nm Nd:YAG laser. Signals were collected by a water-immersed objective with a 550 nm long pass filter to remove the scattered light. Donor dye signals were collected using a 630 nm dichroic mirror and a charge coupled device camera.

smFRET traces were recorded with a 100 ms time resolution and analyzed with Interactive Data Language (IDL) to give single-molecule traces of fluorescence intensity over time. Outputs from IDL were processed with custom MATLAB scripts, which are available to download from https://physics.illinois.edu/cplc/software/. Efficiency of FRET was calculated as the acceptor channel intensity divided by the sum of donor and acceptor channel intensities. Folding populations were calculated through the removal of donor only (Cy3) containing traces and by applying a Gaussian fit to the peaks of FRET histograms generated from 20 fields of view.

Experimental data

For a given sequence, three readings of NMM measurements were recorded and the average intensity value was used throughout the analysis. We represented the loop components of a GQ sequence using the length vector ($L1$,$L2$,$L3$) and nucleotide content N. For instance, (4,1,2) and N = A encodes the sequence GGGAAAAGGGAGGGAAGGG. We only considered the cases where all nucleotides in the loops are the same, in order to fully characterize the rules governing these simple, yet poorly understood cases. The total length of intervening sequences is denoted as $L = L1 + L2 + L3$. We considered combinations of $L1$, $L2$ and $L3$ such that $L \leq 12$, and N is allowed to be A, C or T. For each N, there are four sequences corresponding to $L1 = L2 = L3$ and 26×3 sequences corresponding to the case where exactly two of the lengths are equal, accounting for $4 + 26 \times 3 = 82$ total points in which at least two of the intervening sequences are repeated. There are a total of 138 possible combinations of loop lengths, such that $L1$, $L2$ and $L3$ are distinct and $L \leq 12$, but we subsampled 64 cases for our measurements in order to reduce the dimension, as explained in Supplementary Table S1. Thus, we have a total number of $(82 + 64) \times 3 = 438$ readings, corresponding to 146 combinations of loop lengths for three different nucleotides.

We fitted the histogram of intensity values to a mixture of two or three Gaussian distributions by using the Expectation-Maximization algorithm ('mixtools' package in R) and plotted individual values using the 'colorRamps' and 'calibrate' packages in R. Categorical histograms based on the nucleotide composition or the minimum loop length composition were plotted, and the distribution of a given subset of categories was compared to the rest of the categories via the one-sided unpaired Wilcoxon rank sum test. Finally, we applied the two-sided Kolmogorov–Smirnov test to compare the distributions of T, C and A pairwise.

Linear regression

We first applied a linear regression model of the NMM intensity against the predictor variables $L1$, $L2$, $L3$, $seqT$, $seqC$ and an intercept term, where $seqT$ and $seqC$ are indicator variables for T and C nucleotides, respectively. Note that $seqA$ was omitted due to the linear constraint $seqA = 1 - seqC - seqT$. We then examined an alternative model by replacing $L1$, $L2$, $L3$ with $minL$, $medL$, $maxL$, where $minL$, $medL$ and $maxL$ correspond to the minimum, median and maximum of the three loop lengths. We trained both models on all 438 sequences' NMM intensities to obtain interpretable coefficients and model prediction. This analysis showed that the second model outperformed the first approach, and we thus used the predictor variables $minL$, $medL$ and $maxL$ thereafter. Subsequently, we performed 6-fold cross-validation to demonstrate that

our model is robust. We randomly partitioned the population into 6 groups, each group containing 73 points. Using one group as test data and the remaining five groups as training data, we computed the average coefficient of determination for both test and training data. We adopted the following definition of the coefficient of determination: $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$, where $\hat{y}_i$ is the predicted value and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the mean of n samples used for calculating R^2. For example, $n = 365$ for training data, and $n = 73$ for test data. Likewise, the residual for each sample i is defined as $y_i - \hat{y}_i$, i.e., the difference between the observed and predicted values. The linear regression method has many advantages such as its simplicity and the interpretability of the coefficients. However, it has the limitation of assuming linearity of the response in predictor variables.

Gaussian process regression

Gaussian process regression (GPR) is a flexible nonparametric regression method that does not assume linearity of the response in predictor variables (32). A Gaussian process f is defined on a set X by specifying that the values of f on any finite number of points in X form random variables following a joint Gaussian distribution, with mean 0 and fixed covariance $k(x, x')$ at $x,x' \in X$. Thus, we only need to define the covariance function $k(x, x')$ in order to specify a Gaussian process; $k(x, x')$ is a kernel that measures the similarity between inputs x and x'. The choice of covariance function plays an important role in model prediction, and a popular choice is the squared exponential function: $k_{SE}(x, x') = \sigma_f^2 e^{[\frac{-(x-x')^2}{2\ell^2}]} + \sigma_n^2\delta(x, x')$, where the hyperparameters σ_f^2 and σ_n^2 are the variance of the process and experimental measurement, respectively, ℓ is the length scale of fluctuation and $\delta(x, x')$ is the Kronecker delta function. For n training data points (x_i, y_i), $i = 1, \ldots, n$, we construct an n by n covariance matrix $\mathrm{K} = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{bmatrix}$. For a test data point x_*, we define $K_* = [k(x_*, x_1)\, k(x_*, x_2) \ldots\, k(x_*, x_n)]$ and $K_{**} = k(x_*, x_*)$. Then, the joint distribution of the observed output $\mathbf{y}$ and predicted output y_* is assumed to be $\begin{bmatrix} \boldsymbol{y} \\ y_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K & K_*^T \\ K_* & K_{**} \end{bmatrix}\right)$, and the predictive distribution is $y_*|\boldsymbol{y} \sim \mathcal{N}(K_* K^{-1} y,\ K_{**} - K_* K^{-1} K_*^T)$. We subsequently obtain our prediction as the mean $\overline{y_*} = K_*\, K^{-1}\boldsymbol{y}$. The above methods were all implemented using the GPML MATLAB package (33).

Choice of covariance functions. A valid covariance function $k(x, x')$ requires the function to be symmetric and positive semidefinite. In addition, many of the widely used kernels are stationary, i.e., it is a function of only the distance $r = |x - x'|$. Two examples of stationary covariance functions are a noiseless squared exponential $k_{SE}(r) = e^{-\frac{r^2}{2l^2}}$, with length-scale parameter l, and a Matérn class $k_{Mat,\nu}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)}(\frac{\sqrt{2\nu}r}{l})^{\nu} K_{\nu}(\frac{\sqrt{2\nu}r}{l})$, with positive parameters ν and l, and a modified Bessel function of the second kind K_{ν}. For half-integer ν, the Matérn function K_{ν} is a product of an exponentially decaying function and a polynomial, with $\nu = \frac{1}{2}$ giving a nonsmooth process. As $\nu \to \infty$, the Matérn function behaves similarly to the squared exponential function, which is smooth. We used different length parameters for each predictor variable, adding flexibility to the input space.

Denoting our predictors (*minL*, *medL*, *maxL*, *seqA*, *seqC*, *seqT*) as (x_1, x_2, x_3, x_4, x_5, x_6), we defined our noiseless covariance function as $k((x_1,\ldots,x_6), (x'_1,\ldots,x'_6)) = x_4 \cdot x'_4 \cdot \sigma^2_{f,A} \cdot k_{Mat,\nu=\frac{5}{2}}((x_1, x_2, x_3), (x'_1, x'_2, x'_3)) + x_5 \cdot x'_5 \cdot \sigma^2_{f,C} \cdot k_{SE}((x_1, x_2, x_3), (x'_1, x'_2, x'_3)) + x_6 \cdot x'_6 \cdot \sigma^2_{f,T} \cdot k_{Mat,\nu=\frac{3}{2}}((x_1, x_2, x_3), (x'_1, x'_2, x'_3))$, where $k_{Mat,\nu}(\cdot)$ is the Matérn kernel with specific ν, $k_{SE}(\cdot)$ is the squared exponential kernel, and $\sigma^2_{f,A}$, $\sigma^2_{f,C}$, $\sigma^2_{f,T}$ each corresponds to the variance of the process for *seqA*, *seqC* and *seqT*, respectively. This combination has been derived by testing the squared exponential and Matérn class with $\nu = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}$ separately for *seq A*, *seq C* and *seq T*, and choosing the best function for each nucleotide.

Estimation of hyperparameters. There are four hyperparameters, $\sigma_{f,N}$,$l_{1,N}$, $l_{2,N}$, $l_{3,N}$ (the length scale for *minL*, *medL*, *maxL*, respectively) for each nucleotide N, summing to a total number of 12. A common method to estimate a set of hyperparameters $\boldsymbol{\theta}$ is by maximizing the marginal log-likelihood $\log p(\boldsymbol{y} \mid X, \boldsymbol{\theta}) = -\frac{1}{2}\boldsymbol{y}^T K_y^{-1}\boldsymbol{y} - \frac{1}{2}log|K_y| - \frac{n}{2}log2\pi$, where $K_y = K + \sigma_n^2 I$ and x and y are predictor and response variables for the training data. We also adopted this method, but implemented in two steps. First, for each individual nucleotide, we initialized each of $l_{1,N}$,$l_{2,N}$, $l_{3,N}$ and $\sigma_{f,N}$ to be 5, and obtained an estimate by using a conjugate gradient method. Note that there are three separate estimates for $\sigma_{f,N}$, obtained for each of $l_{1,N}$,$l_{2,N}$ and $l_{3,N}$, and that we let the final estimate be the average of the three. We then initialized all 12 hyperparameters with the values obtained from the previous step and maximized the marginal log-likelihood over all lengths and nucleotides. This approach allows for more flexibility in each length scale than treating each loop length with an equal weight. Finally, we estimated $\sigma_n^2 = 18$ as the empirical covariance of our replicate experimental NMM intensity measurements. Supplementary Table S2 contains the estimated hyperparameters used for fitting the entire population, and the same estimation method was repeated for each cross-validation set.

RESULTS

Pilot study establishes cut-off for GQ folding

We designed a series of GQ forming dsDNA constructs by following the conventional pattern, [GGGN$_{L1}$GGGN$_{L2}$GGGN$_{L3}$GGG] as defined above (Figure 1A). The GQ formation in dsDNA was performed in 40% PEG condition used previously (30,34). We have ex-

cluded the loop lengths that would not support GQ folding based on our previous study that revealed a significantly diminished GQ folding potential in dsDNA compared to ssDNA (30). As a pilot study, we designed 246 sequences that satisfied the following three conditions. First, the total loop length, $L1 + L2 + L3$, was restricted to be 12 bases or less. Second, all loops consisted entirely of only one nucleotide, A, C or T. Third, at least two loop lengths were of equal length. NMM was applied to each DNA in 96 well plates, and the induced fluorescence from NMM was measured to assess the GQ folding potential (Figure 1B and C). The NMM measurement was repeated three times per DNA and the results were highly reproducible (average standard deviation = 18; Supplementary Data). The NMM-based fluorescence assay allows detection of parallel GQ structure, which is the only form of GQ that can form in dsDNA. In our previous work, we used single molecule FRET, Circular Dichroism (CD), NMM and Crystal Violet-induced ensemble fluorescence measurements to demonstrate that only parallel GQ can form in the context of dsDNA. The Crystal Violet (CV) fluorescence selectively measures antiparallel GQ formation. The GQ folding probed by smFRET matched closely with the NMM fluorescence, whereas the CV fluorescence showed no signal in all sequences tested (Supplementary Figure S1), indicating that only parallel GQ conformation can be supported in the context of dsDNA (30). We chose NMM over other GQ ligands, NMP, NMMDE and BRACO19 due to the lowest K_d (dissociation constant) exhibited by NMM, although all four are highly specific to parallel GQ structure (Supplementary Figure S2) (25). Therefore, the NMM signal induced by potential GQ-dsDNA indicates the degree of its GQ folding. We expect a high NMM signal for DNA that primarily forms into a GQ, intermediate intensity for a combined population of folded and nonfolded GQs, and no signal if all DNA molecules become duplexed (Figure 1C).

Based on the NMM intensity, we roughly categorized the folding propensity of the 246 sequences into folding (> 254) and nonfolding (< 254) classes by using a Gaussian mixture model (Figure 2A). The Kolmogorov–Smirnov test did not detect a statistically significant deviation of the model from the data (two-sided p-value = 0.315), supporting the goodness of fit. The NMM intensity cut-off of 254, estimated from the transition point in the ratio of posterior class probabilities, corresponded to 52 and 48% of the sequences as folding and nonfolding, respectively. In order to check whether all three nucleotide types yield similar NMM intensity distributions, we grouped the data by the nucleotide content of loop sequences and plotted the empirical cumulative distribution for each group (Figure 2B). The distribution for T was clearly shifted to the right, strongly suggesting that T loops induce a stronger GQ folding potential than A and C loops. This effect is further analyzed and discussed below. For control measurement, the same series of GQ sequences measured in ssDNA displayed overall enhanced folding potential probed by NMM (Supplementary Figure S3).

To test the validity of NMM intensity as an accurate measurement of GQ folding propensity, we performed a smFRET assay on a selected subset of GQ DNA constructs

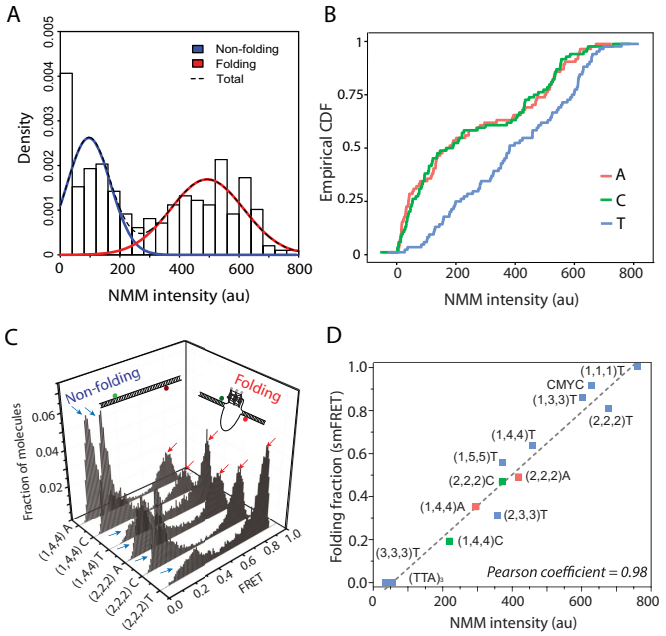

Figure 2. Pilot study of NMM fluorescence data points and relationship with smFRET scores. (**A**) The population of 246 sequences is separated into nonfolding (blue) and folding (red) classes via the Gaussian mixture model. Dotted line shows the marginal (total) distribution of NMM intensities in the fitted mixture model. (**B**) The empirical cumulative distribution functions (CDF) are plotted for three nucleotides, A (red), C (green) and T (blue). (**C**) GQ folding percentage is verified through smFRET analysis for the loop lengths (1,4,4) and (2,2,2) in all three bases, where ($L1$, $L2$, $L3$) denotes the three loop lengths. High FRET populations (>0.7) correspond to GQ folding, while low FRET populations (<0.3) represent nonfolding sequences. (**D**) The graph shows the linear relationship between NMM intensity (x-axis) and GQ folding percentage (y-axis) for a wide range of 13 sequences that are composed of A (red), C (green) and T (blue). Pearson coefficient of 0.98 indicates a strong correlation between the two measurements.

that represent low to high ranges of NMM intensity. Our previous study demonstrated that NMM intensity is highly correlated with GQ folding fraction measured by smFRET when the loop is composed of T bases. In order to further test whether such correlation still holds for other bases, we performed NMM intensity measurements and smFRET analysis on the loop sequences including the ones composed of either A or C. Due to the two fluorophores attached at the boundary of a GQ forming sequence, high FRET is expected only when the GQ is folded, whereas low FRET indicates duplexed DNA without GQ folding (Figure 2C). The histograms built from FRET values of over 3000 molecules yielded two FRET peaks corresponding to folded and nonfolded (duplexed) GQ structures represented by high FRET (0.8) and low FRET (0.2) peaks, respectively (30). The folded fraction was calculated by obtaining an area under the Gaussian-fitted curves on the FRET histogram. The resulting plot showed that the NMM intensity was highly correlated with the smFRET-based folding estimation (Pearson Coefficient of 0.98), even for the A and C containing loops, validating the NMM as a reliable folding probe for GQ DNA regardless of the loop sequence (Figure 2D). Consistent with the above finding, the loops composed of A and C displayed substantially less folding for

both (1,4,4) and (2,2,2) than the T loop, strongly suggesting sequence-dependent GQ-folding propensity (Figure 2B).

Expansive coverage of candidate sequences identifies loop length and composition dependence of folding trends

In order to further investigate the dependence of GQ-folding trends on loop lengths and the nucleotide content, we visualized our initial data by constructing color-weighted NMM intensity graphs. For a clear illustration of the previously observed GQ-folding pattern, we first partitioned the data into three groups according to the loop length composition. The loop lengths ($L1$, $L2$, $L3$) were encoded in a two-dimensional space, instead of three dimensions, by defining the variable Z to denote the length that is repeated at least twice, and V the remaining length. Using these two variables, the three possible permutations of loop lengths considered were coded as (Z,Z,V), (Z,V,Z) and (V,Z,Z) (Figure 3A). Each of these three groups were then further partitioned into three classes based on the loop sequence, T, C and A, thereby visually capturing the experimental NMM intensities of all 246 sequences via nine different subgraphs (Figure 3B). High GQ-induced NMM fluorescence levels were displayed in red (warm) colors, while low intensities were shown in blue (cool) colors. The sequences with nucleotide T and loop pattern (V,Z,Z) are shown in Figure 3A. This representation clearly demonstrates an inverse relation between the intensity and minimum length (*minL*), as shown by the similar colors for sequences with the same *minL* and the color gradient with respect to increasing *minL* (red and yellow for the 14 sequences with $minL = 1$, mostly green for the 10 sequences with $minL = 2$ and dark blue for $minL>2$). By contrast, the correlation between intensity and total length (L) remained weak, as shown by the wide fluctuation of colors for sequences with the same L. For example, the sequences in each group with $7 \leq L \leq 12$ displayed colors ranging from red to blue, providing little insight on the likelihood of a particular group of sequences to fold.

We compared the subgraphs to further investigate the effect of nucleotide content and length distributions on the GQ folding intensity (Figure 3B). Comparing the three rows pairwise revealed that C and A loop compositions generally showed a lower folding pattern than T, consistent with our previous observation (Figure 2B and C). For example, in all three permutations of the loop lengths (3, 2, 2), T exhibited yellow or green colors (in the range 400 to 500), whereas C and A displayed light or dark blue (less than 250); according to the NMM intensity cut-off value of 254 derived in the previous section, only the T-containing sequences were folding in these cases, thus exemplifying the overall diminished GQ folding for C and A compared to T. Examining the effect of loop lengths on folding, we found that the inverse relation between the minimum loop length and intensity observed in Figure 3A was present in all groups: sequences with $minL = 1$ generally displayed high intensity, whereas the intensity values rapidly dropped as *minL* increased. Furthermore, even though the intensities were generally not affected by the ordering of loop lengths, we noticed that for T and A, the sequence arrangements of (1, *maxL*, 1) were less likely to fold than (1, 1, *maxL*) or (maxL, 1, 1), as the 10 data points along the left-most vertical line in the (Z, V, Z) column exhibited cooler colors than those in the (V, Z, Z) and (Z, Z, V) columns. Likewise, the diminished intensity of five data points along the bottom-most horizontal line of the (Z, V, Z) column indicated that (*maxL*, 1, *maxL*) was less likely to fold than (1, *maxL*, *maxL*) and (*maxL*, *maxL*, 1) for all T, C and A loops.

In order to test and validate our observations from the initial data, we next expanded the study design to include sequences with unique loop lengths in all three positions, while keeping the total loop length at 12 base pairs or less. Of 138 such possible combinations, we chose 64 combinations for each nucleotide by selecting every other point in each of 7 unique combinations of minL and medL in the ordered list (Supplementary Table S1). This choice allowed us to reduce the number of new cases by roughly half, yielding a total of $246 + 64 \times 3 = 438$ sequences. When applied to the NMM fluorescence assay, the new 192 data points with unique loop lengths yielded an intensity distribution pattern that differed from the first 246 pilot DNA sequences tested above. Instead of the bimodal distribution seen in the previous pilot data (Figure 2A), the new set of DNA displayed a broad single peak centered around 300 (Figure 4A). This difference is likely due to the change in the distribution of loop lengths for the new sets of DNA. The loops in the pilot DNA were constrained to possess at least two repeated lengths, while the loops in the new design had unique lengths in the three positions. As a result, the two sets had similar minimum loop length distributions and significantly different median and maximum loop length distributions (Two-sided Kolmogorov–Smirnov test p-value = 0.0044, 2.4×10^{-11}, $5.66 \times 10^{-15}$for *minL*, *medL*, *maxL*, respectively; Supplementary Figure S4). Compared to the pilot data, the new set contained a substantially higher fraction of sequences with long *medL* and *maxL* loop lengths, likely contributing to the broad peak in the mid-to-low range of NMM intensity. We subsequently confirm this hypothesis using regression models. When the data were grouped by individual bases, we again observed the highest GQ folding potential for T, followed by C and A (Figure 4B). The same set of data analyzed by the colorimetric mapping still followed the same trend as previously observed: short *minL* and nucleotide T both led to high folding propensity (Supplementary Figure S5). Hereafter, we used this comprehensive data set to verify our observations using rigorous statistical methods and to devise predictive regression models applicable to a general set of sequences.

GQ folding depends on minimum loop length and nucleotide T

As an initial means to understand the combined data set, we first categorized the 438 experimentally generated NMM intensity values into three classes based on a mixture of three Gaussian distributions fitted via the Expectation-Maximization algorithm (Figure 5A). This partitioning was based on the two peaks observed in pilot data (Figure 2A) and the third peak in the second data set (Figure 4A), and the two-sided Kolmogorov-Smirnov test p-value of 0.88 confirmed a good model fit. Comparing the ratios of posterior class probabilities suggested the following three GQ

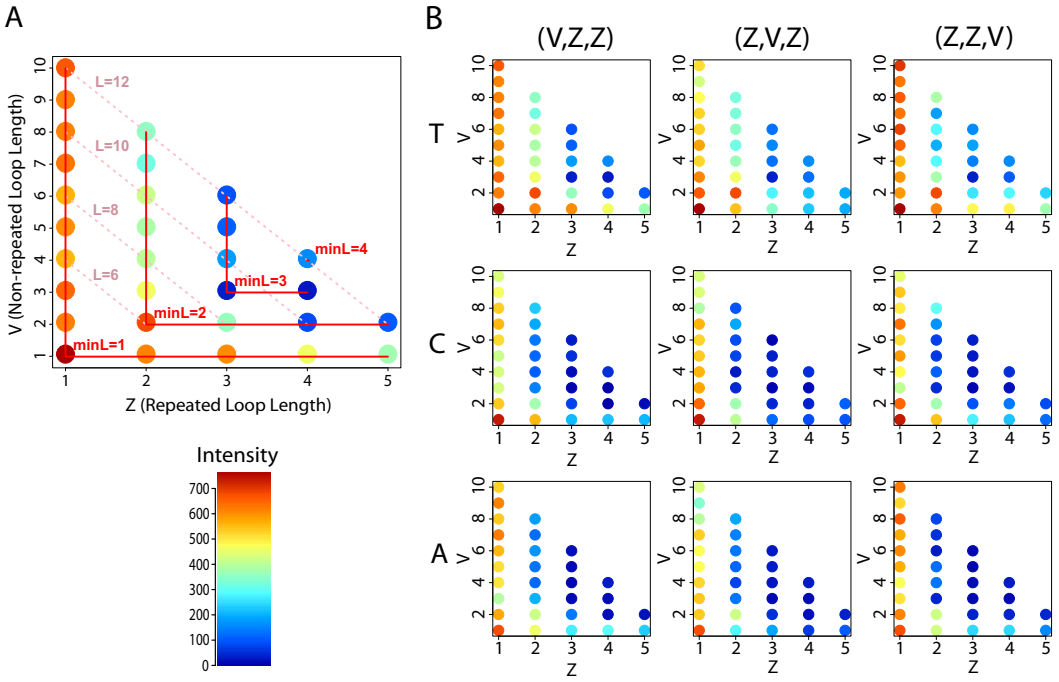

Figure 3. Visualization of the NMM intensity data used in pilot study. (**A**) Intensities of 30 data points are plotted for nucleotide T and permutation (V, Z, Z), where Z (x-axis) is the repeated loop length and V (y-axis) is the remaining length. The minimum (*minL*) and sum (*L*) of three loop lengths are indicated in red and pink lines, respectively, and each data point is colored according to the intensity color bar plotted in the bottom. (**B**) Intensities of all 246 sequences used for our pilot study are plotted, in nine subgraphs that have similar structure as Figure 3A. Rows indicate N = T, C or A, in order, and columns indicate the three possible permutations for each Z and V, e.g., (V, Z, Z), (Z, V, Z) and (Z, Z, V).

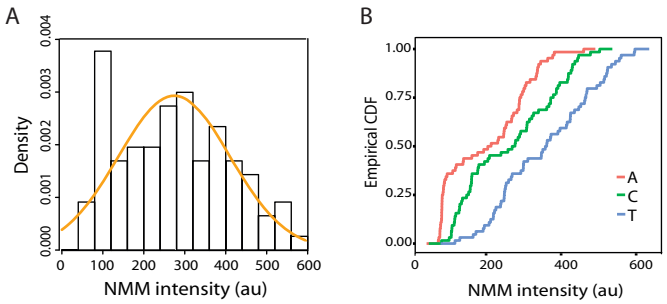

Figure 4. Overview of new set of 192 sequences tested. (**A**) The density of the new set is plotted and the Gaussian distribution is overlaid in orange, where mean and variance are calculated from the 192 intensity values. (**B**) The empirical cumulative distribution functions (CDF) are plotted for sequences in A (red), C (green) and T (blue).

folding categories: (1) Intensity < 151 for nonfolding, (2) 151 < Intensity < 412 for combined folding and nonfolding, and (3) Intensity > 412 for strong folding. Each of the nonfolding, combined and strong folding categories contained 31, 39 and 30% of the data, respectively.

Using the above threshold values as a guideline, we investigated the role of loop nucleotide content on folding. The three nucleotide-specific histograms of NMM intensity clearly showed that sequences containing T had a greater tendency to fold than those containing C or A (one-sided unpaired Wilcoxon rank sum test p-value = 4.3×10^{-13} for sequences containing T versus those containing C or A; Figure 5B). Moreover, the overall distribution for T was significantly different from that for C or A (two-sided Kolmogorov-Smirnov (KS) test p-value = 2.8×10^{-7} and 2.008×10^{-9} for C and A, respectively; Supplementary Figure S6), while the distribution for C was not significantly different from that for A (two-sided KS test p-value = 0.13; Supplementary Figure S6). We note that the sequence dependence shown here is not due to the thermal stability of the GQ forming fragments with different length of GC content. Examining the melting temperature (T_m) for all 438 GQ sequences used in this study shows that the sequences containing C have, on average, approximately 15°C higher T_m than the sequences containing either A or T, demonstrating that the similar level of GQ forming potential between A and C-containing sequences and the increased GQ forming potential seen only in the T-containing sequences cannot be explained by the duplex stability of GQ DNA (Supplementary Figure S7).

The three loop lengths *L*1, *L*2 and *L*3 have been previously proposed to modulate GQ folding, but the rule governing their effect remains unknown (6,23–25,35–36). Inspection of the intensity plots in Figure 3 revealed that an informative feature was the minimum of loop lengths (*minL*). Indeed, the intensity histograms plotted for different *minL* values showed that the sequences with *minL* = 1 spanned all three folding categories, although slightly skewed towards the strong folding region, those with *minL* = 2 were either nonfolding or combined, and those with *minL* > 2 were mostly nonfolding (one-sided unpaired Wilcoxon rank sum test p-value < 2.2×10^{-16} for *minL* = 1 versus *minL* >1; Figure 5C). The NMM intensity thus decreased dra-

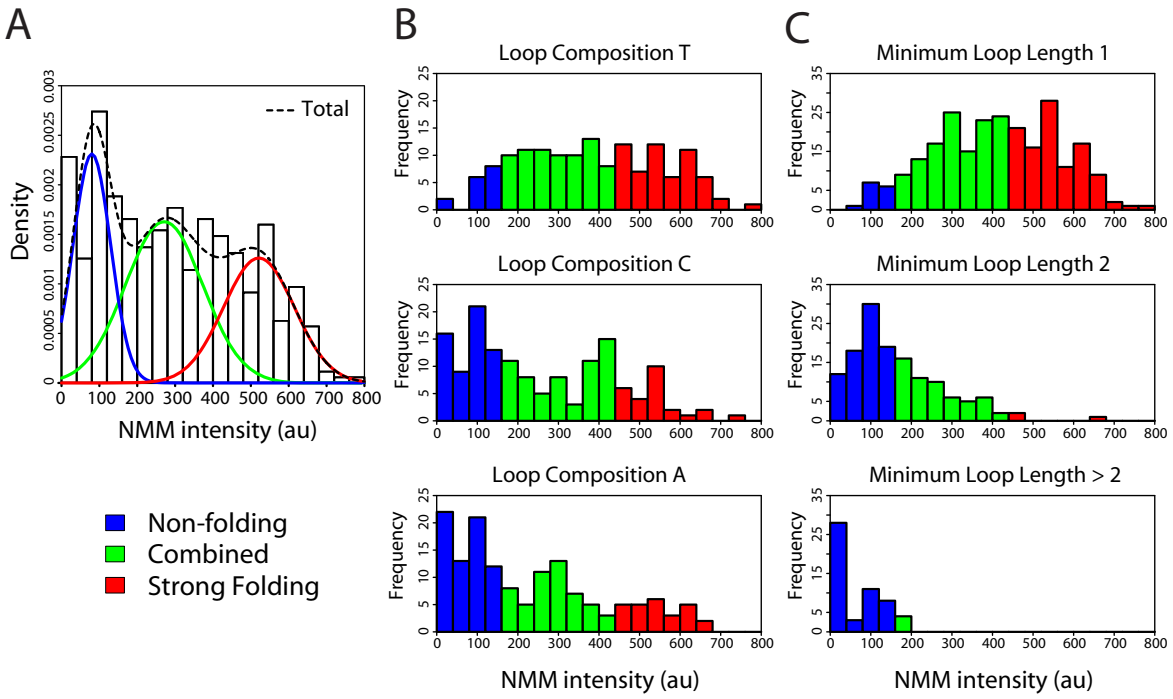

Figure 5. Histograms of 438 comprehensive NMM fluorescence data points. (**A**) The density of NMM intensities is plotted and the Gaussian mixture model separates the population into three separate classes: nonfolding (blue), combined (green) and strong folding GQs (red). Dotted line shows the marginal distribution of NMM intensities in the fitted mixture model. (**B**) Three independent histograms of the NMM intensities are provided for each loop composition T, C and A. Bars are colored according to their GQ classification from Figure 5A: blue if intensity < 151, green if $151 <$ intensity < 412 and red if intensity > 412. (**C**) Histograms of the NMM intensities are provided for sequences with minimum loop length 1, 2 and greater than 2 and the bars are colored according to the GQ class that they belong to.

matically as the minimum loop length increased, suggesting that transforming the loop lengths $L1$, $L2$ and $L3$ to order statistics *minL*, *medL* and *maxL* may help predict GQ intensity. Our regression models in the subsequent section explores this transformation, after attempting a simpler linear fit with $L1$, $L2$ and $L3$.

Regression models predict GQ folding propensity

To learn how GQ folding propensity depends on the characteristic features of intervening loops, we first fitted the experimental NMM intensities using a linear regression model with the following five predictor variables: $L1$, $L2$, $L3$, $seqT$ and $seqC$. The $seqT$ and $seqC$ are indicator variables for the T and C nucleotides, respectively, and $seqA = 1 - seqT - seqC$ is omitted due to its linear dependency on $seqT$ and $seqC$. Training on all 438 sequences, we obtained an R^2 value of 0.35, implying that our model could predict only 35% of the total variance in NMM intensities. By transforming the three loop lengths to the order statistics *minL*, *medL*, *maxL*, our R^2 value significantly improved to 0.80. Thus, our subsequent analyses are based on this transformation. The predicted mean intensity was $\hat{y} = 679 + 149seqT + 27seqC - 147minL - 74medL - 4maxL$. Among the regression coefficients, the two largest magnitudes corresponded to $seqT$ and *minL*, confirming that the two main driving factors of GQ folding are the T loop composition and the minimal loop length. By contrast, $seqC$ and *maxL* had the smallest magnitudes and had the least significant p-values of 0.008 and 0.125, respectively, suggesting that they both do not contribute substantially to folding. The fact that the coefficient for $seqC$ was relatively small also indicated that there was very little difference between A and C nucleotides. By contrast, the effect of T on folding was more than 5-fold greater than that of C. These results are consistent with the similarity in intensity distribution between C and A, and the distinction from T previously detected by the Kolmogorov–Smirnov Test (Supplementary Figure S6).

To test the generalizability of our model, we performed 6-fold cross-validation. The dataset of 438 points was randomly partitioned into 6 groups, and each group was tested using parameters trained from the remaining 5 groups. As a result, we obtained an R^2 value of 0.796 ± 0.005 for the training set and a comparable value of 0.784 ± 0.023 for the test set, supporting that our model is robust. We plotted the average absolute values of residuals, defined as the difference between the observed and the predicted values, in order to visualize how well the model fits each data point (Supplementary Figure S8). Despite the simple nature of our model, most of our predictions did not deviate substantially from the observed true values, as indicated by the overall blue colors ($|\text{residuals}| < 150$). There were, however, some outlier data points for A and C nucleotides showing a poor fit when at least two lengths were repeated. Moreover, the most critical issue for all nucleotides was that the points (1, 1, 1), (2, 2, 2), (3, 3, 3) and (4, 4, 4) had large absolute residuals, most likely due to nonlinear behaviors of their intensities. In

order to improve our prediction accuracy, especially at these outlier points, we developed a Gaussian Process Regression (GPR) model.

Compared to the linear regression model's R^2 value of 0.80, the GPR model trained with the same predictor variables on all 438 sequences (Methods) showed a substantial improvement to $R^2 = 0.92$. Six-fold cross validation using the same partition groups from the linear regression analysis yielded $R^2 = 0.918 \pm 0.002$ for training and $R^2 = 0.878 \pm 0.039$ for test data, which, on average, improved the linear model results by 0.12 and 0.09, respectively. To visualize the overall performance of the GPR method and compare it with that of the linear model, the average absolute values of residuals for GPR were again plotted (Supplementary Figure S9; cf. Figure S8). The plot was generally cooler than Figure S8, especially at the data points that were problematic with the linear regression approach, e.g., the A-containing sequences with loop lengths (1, 1, *maxL*) and (2, 2, *maxL*). Additionally, we observed significant improvements in predicting (1, 1, 1), (2, 2, 2), (3, 3, 3) and (4, 4, 4) for all nucleotides, thus addressing the major difficulties encountered in the linear model. Overall, the only data points with large prediction errors were (2, 2, 2) for sequence T, and (1, 8, 1) and (2, 2, 2) for A, with absolute residuals of ∼200, compared to the rest being less than 100.

Although GPR does not directly provide easily interpretable coefficients as in linear regression, the estimated hyperparameters do confirm our findings from the linear model (Supplementary Table S2). For the squared exponential and Matérn class covariance functions, the length parameter l controls the effect size of the difference in the corresponding predictor variable, and its large value suggests that the response variable is not very sensitive to the corresponding feature. Consistent with the linear regression result, we observed that the length parameters $l_{1,A}$, $l_{1,C}$, $l_{1,T}$ for minL were shorter than those for *maxL*, implying that the intensity depended on *minL* more than on *maxL*, with the effect being most notable for the T nucleotide.

DISCUSSION

We have developed a simple model that can explain the GQ folding potential of a large set of dsDNA sequences. The model is based on studying the distribution of NMM intensity values measured in over 400 putative GQ sequences; this comprehensive sampling spans the potential folding space of loop parameters that cover the generally accepted range of GQ folding sequences. Our results suggest that the most significant composition property that facilitates GQ folding in dsDNA is the minimum loop length. For example, sequences with minimum loop length (*minL*) of 1 constitute 63 and 97% of the combined and strong folding populations, respectively, implying that those with *minL* longer than 1 are not as likely to fold into GQ (Figure 5C). This result is consistent with the finding from a recent *in vivo* study that GQs containing at least one loop length of 1 are preferentially associated with genomic replication errors (36). Furthermore, there is a significant folding propensity bias among base compositions, with T promoting the highest level of GQ formation. Our computational predictive models based on the order statistics of loop lengths and sequence compositions accurately capture these rules, and cross-validation shows that these models can predict unseen GQ forming sequences with high accuracy.

Our regression model is based on the order statistics of loop lengths and thus assumes that the folding propensity is invariant under the permutation of loop lengths. However, a recent study suggests that having a long middle loop may disfavor folding; specifically, it is shown that the (1,*maxL*,1) configuration has reduced GQ folding potential compared to the shuffled configurations (1, 1, *maxL*) and (*maxL*, 1, 1) (36). Our NMM data also exhibits slightly diminished intensities for (1, *maxL*, 1) compared to (1, 1, *maxL*) and (*maxL*, 1, 1) for nucleotides T and A, but not for C. Similarly, in our experiments, the configuration (*maxL*, 1, *maxL*) exhibits lower intensities than (1, *maxL*, *maxL*) and (*maxL*, *maxL*, 1) for all nucleotides. These two cases suggest that our model assumption of permutation symmetry may not hold for some GQ sequences and may lead to prediction errors (Figure 3B). In order to investigate the impact of rearranging loop lengths on folding potential, one can decompose the NMM intensities into Fourier modes that are basis functions defined on the six permutations of (*minL*, *medL*, *maxL*) (Supplementary Method S1); this approach mathematically characterizes the dominant fluctuating behavior of NMM values on permutation elements (Supplementary Figure S10, Supplementary Table S3). Implementing this analysis shows no consistent pattern for 192 sequences containing unique loop lengths, but uncovers the pattern previously observed for sequences with repeated loop lengths (36). That is, the Fourier decomposition of NMM intensities identifies two dominant modes that combine to reduce intensity in the (1, *maxL*, 1) configuration for T and A—but not for C—nucleotides (Supplementary Figure S11; one-sided unpaired Wilcoxon rank sum test for $\{(1, 1, maxL), (maxL, 1, 1)\}$ versus $\{(1, maxL, 1)\}$ p-value = 7.8×10^{-4}, 0.705, 0.003 for T, C, A, respectively). A similar analysis finds reduced folding potential in (*maxL*, 1, *maxL*) compared to its permuted configurations for all nucleotides (Supplementary Figure S12; one-sided unpaired Wilcoxon rank sum test for $\{(1, maxL, maxL), (maxL, maxL, 1)\}$ versus $\{(maxL, 1, maxL)\}$ p-value = 0.002 for all T, C, A). However, our data and mathematical analysis clarify that these patterns of reduced folding potential do not generalize to sequences with minimum loop length greater than 1.

We note that the interpretation of our result may be limited by several factors. In terms of the experimental setup, the DNA constructs used in the study lacks supercoiling that may exist in genomic DNA. Additionally, we employ PEG mediated folding condition for inducing GQ formation in dsDNA (34), which may have promoted or diminished the GQ formation. In terms of computational methods, even though our two regression models can predict GQ folding propensity with high accuracy, both models have limitations. First, our models, as they currently stand, cannot be directly applied to sequences that contain any guanine bases in a loop, because of the ambiguity in assigning guanines to either a loop or G-tetrads. Second, our models have been validated only on sequences with a single uniform base composition in the loops. For sequences containing more than one type of base, it may require modeling not

only the concentration of each nucleotide, but also the specific ordering of the nucleotides. Thus, future research directions include developing a predictive model that can handle sequences with intervening loops consisting of a combination of A, C, G and T. For such a set of complex sequences, the flexibility of Gaussian process regression will likely provide additional advantages over the linear regression approach. As an important step towards achieving these goals, our work provides a reliable experimental and computational framework that greatly reduces the search space for potential GQ forming sequences and quantitatively predicts the likelihood of folding for a broad range of candidate sequences.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

The authors would like to thank Miroslav Hejna, Hu Jin, Wooyoung Moon, Olgica Milenkovic and the Myong lab for helpful discussions.

FUNDING

National Science Foundation Graduate Research Fellowship [DGE-1144245 to M.K.]; National Institutes of Health (NIH) [R01CA163336 to J.S.S]; U.S. National Science Foundation Physics Frontiers Center Program through the Center for the Physics of Living Cells [0822613]; American Cancer Society [RSG-12-066-01-DMC]; NIH [1DP2GM105453 to A.K., C.L., S.M.]. Funding for open access charge: NIH [1DP2GM105453].
Conflict of interest statement. None declared.

REFERENCES

1. Hud,N.V., Smith,F.W., Anet,F.A. and Feigon,J. (1996) The selectivity for K+ versus Na+ in DNA quadruplexes is dominated by relative free energies of hydration: a thermodynamic analysis by 1H NMR. *Biochemistry*, **35**, 15383–15390.
2. Burge,S., Parkinson,G.N., Hazel,P., Todd,A.K. and Neidle,S. (2006) Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res.*, **34**, 5402–5415.
3. Schultze,P., Hud,N.V., Smith,F.W. and Feigon,J. (1999) The effect of sodium, potassium and ammonium ions on the conformation of the dimeric quadruplex formed by the Oxytricha nova telomere repeat oligonucleotide d (G4T4G4). *Nucleic Acids Res.*, **27**, 3018–3028.
4. Todd,A.K., Johnston,M. and Neidle,S. (2005) Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res.*, **33**, 2901–2907.
5. Rhodes,D. and Lipps,H.J. (2015) G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res.*, **43**, 8627–8637.
6. Bochman,M.L., Paeschke,K. and Zakian,V.A. (2012) DNA secondary structures: stability and function of G-quadruplex structures. *Nat. Rev. Genet.*, **13**, 770–780.
7. Besnard,E., Babled,A., Lapasset,L., Milhavet,O., Parrinello,H., Dantec,C., Marin,J.-M. and Lemaitre,J.-M. (2012) Unraveling cell type–specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. *Nat. Struct. Mol. Biol.*, **19**, 837–844.
8. Murat,P. and Balasubramanian,S. (2014) Existence and consequences of G-quadruplex structures in DNA. *Curr. Opin. Genet. Dev.*, **25**, 22–29.
9. Agrawal,P., Lin,C., Mathad,R.I., Carver,M. and Yang,D. (2014) The major G-quadruplex formed in the human BCL-2 proximal promoter adopts a parallel structure with a 13-nt loop in K+ solution. *J. Am. Chem. Soc.*, **136**, 1750–1753.
10. Zhang,C., Liu,H.-H., Zheng,K.-w., Hao,Y.-H. and Tan,Z. (2013) DNA G-quadruplex formation in response to remote downstream transcription activity: long-range sensing and signal transducing in DNA double helix. *Nucleic Acids Res.*, **41**, 7144–7152.
11. Salvati,E., Zizza,P., Rizzo,A., Iachettini,S., Cingolani,C., D'Angelo,C., Porru,M., Randazzo,A., Pagano,B. and Novellino,E. (2013) Evidence for G-quadruplex in the promoter of VEGFR-2 and its targeting to inhibit tumor angiogenesis. *Nucleic Acids Res.*, **42**, 2945–2957.
12. Paeschke,K., Capra,J.A. and Zakian,V.A. (2011) DNA replication through G-quadruplex motifs is promoted by the Saccharomyces cerevisiae Pif1 DNA helicase. *Cell*, **145**, 678–691.
13. Gray,L.T., Vallur,A.C., Eddy,J. and Maizels,N. (2014) G quadruplexes are genomewide targets of transcriptional helicases XPB and XPD. *Nat. Chem. Biol.*, **10**, 313–318.
14. Lam,E.Y.N., Beraldi,D., Tannahill,D. and Balasubramanian,S. (2013) G-quadruplex structures are stable and detectable in human genomic DNA. *Nat. Commun.*, **4**, 1796.
15. Biffi,G., Tannahill,D., McCafferty,J. and Balasubramanian,S. (2013) Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat. Chem.*, **5**, 182–186.
16. Mathad,R.I., Hatzakis,E., Dai,J. and Yang,D. (2011) c-MYC promoter G-quadruplex formed at the 5′-end of NHE III1 element: insights into biological relevance and parallel-stranded G-quadruplex stability. *Nucleic Acids Res.*, **39**, 9023–9033.
17. DeJesus-Hernandez,M., Mackenzie,I.R., Boeve,B.F., Boxer,A.L., Baker,M., Rutherford,N.J., Nicholson,A.M., Finch,N.A., Flynn,H. and Adamson,J. (2011) Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron*, **72**, 245–256.
18. Palumbo,S.L., Ebbinghaus,S.W. and Hurley,L.H. (2009) Formation of a unique end-to-end stacked pair of G-quadruplexes in the hTERT core promoter with implications for inhibition of telomerase by G-quadruplex-interactive ligands. *J. Am. Chem. Soc.*, **131**, 10878–10891.
19. Cogoi,S. and Xodo,L.E. (2006) G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription. *Nucleic Acids Res.*, **34**, 2536–2549.
20. Huppert,J.L. and Balasubramanian,S. (2007) G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.*, **35**, 406–413.
21. Maizels,N. (2006) Dynamic roles for G4 DNA in the biology of eukaryotic cells. *Nat. Struct. Mol. Biol.*, **13**, 1055–1059.
22. Huppert,J.L. and Balasubramanian,S. (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.*, **33**, 2908–2916.
23. Guédin,A., Gros,J., Alberti,P. and Mergny,J.-L. (2010) How long is too long? Effects of loop size on G-quadruplex stability. *Nucleic Acids Res.*, **38**, 7858–7868.
24. Hazel,P., Huppert,J., Balasubramanian,S. and Neidle,S. (2004) Loop-length-dependent folding of G-quadruplexes. *J. Am. Chem. Soc.*, **126**, 16405–16415.
25. Tippana,R., Xiao,W. and Myong,S. (2014) G-quadruplex conformation and dynamics are determined by loop length and sequence. *Nucleic Acids Res.*, **42**, 8106–8114.
26. Eddy,J. and Maizels,N. (2006) Gene function correlates with potential for G4 DNA formation in the human genome. *Nucleic Acids Res.*, **34**, 3887–3896.
27. Kikin,O., D'Antonio,L. and Bagga,P.S. (2006) QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res.*, **34**, W676–W682.
28. Stegle,O., Payet,L., Mergny,J.-L., MacKay,D.J. and Huppert,J.L. (2009) Predicting and understanding the stability of G-quadruplexes. *Bioinformatics*, **25**, i374–i1382.
29. Bedrat,A., Lacroix,L. and Mergny,J.L. (2016) Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res.*, **44**, 1746–1759.
30. Kreig,A., Calvert,J., Sanoica,J., Cullum,E., Tipanna,R. and Myong,S. (2015) G-quadruplex formation in double strand DNA probed by NMM and CV fluorescence. *Nucleic Acids Res.* , **43**, 7961–7970.

31. Roy,R., Hohng,S. and Ha,T. (2008) A practical guide to single-molecule FRET. *Nat. Meth.*, **5**, 507–516.
32. Rasmussen,C.E. and Williams,C.K.I. (2006) Gaussian processes for machine learning. MIT Press, Cambridge.
33. Rasmussen,C.E. and Nickisch,H. (2010) Gaussian processes for machine learning (GPML) toolbox. *J. Mach. Learn. Res.*, **11**, 3011–3015.
34. Zheng,K.W., Chen,Z., Hao,Y.H. and Tan,Z. (2010) Molecular crowding creates an essential environment for the formation of stable G-quadruplexes in long double-stranded DNA. *Nucleic Acids Res.*, **38**, 327–338.
35. Bugaut,A. and Balasubramanian,S. (2008) A sequence-independent study of the influence of short loop lengths on the stability and topology of intramolecular DNA G-quadruplexes. *Biochemistry*, **47**, 689–697.
36. Piazza,A., Adrian,M., Samazan,F., Heddi,B., Hamon,F., Serero,A., Lopes,J., Teulade-Fichou,M.P., Phan,A.T. and Nicolas,A. (2015) Short loop length and high thermal stability determine genomic instability induced by G-quadruplex-forming minisatellites. *EMBO J.*, **34**, 1718–1734.