

# Long promoter sequences form higher-order G-quadruplexes: an integrative structural biology study of *c-Myc*, *k-Ras* and *c-Kit* promoter sequences

Robert C. Monsen<sup>1</sup>, Lynn W. DeLeeuw<sup>1</sup>, William L. Dean<sup>1</sup>, Robert D. Gray<sup>1</sup>, Srinivas Chakravarthy<sup>2</sup>, Jesse B. Hopkins<sup>2</sup>, Jonathan B. Chaires<sup>1,3,4,\*</sup> and John O. Trent<sup>1,3,4,\*</sup>

<sup>1</sup>UofL Health Brown Cancer Center, University of Louisville, Louisville, KY 40202, USA, <sup>2</sup>The Biophysics Collaborative Access Team (BioCAT), Department of Biological, Chemical, and Physical Sciences, Illinois Institute of Technology, Chicago, IL 60616, USA, <sup>3</sup>Department of Medicine, University of Louisville, Louisville, KY 40202, USA and <sup>4</sup>Department of Biochemistry and Molecular Genetics, University of Louisville, Louisville, KY 40202, USA

Received January 31, 2022; Revised March 03, 2022; Editorial Decision March 04, 2022; Accepted March 21, 2022

## ABSTRACT

We report on higher-order G-quadruplex structures adopted by long promoter sequences obtained by an iterative integrated structural biology approach. Our approach uses quantitative biophysical tools (analytical ultracentrifugation, small-angle X-ray scattering, and circular dichroism spectroscopy) combined with modeling and molecular dynamics simulations, to derive self-consistent structural models. The formal resolution of our approach is 18 angstroms, but in some cases structural features of only a few nucleotides can be discerned. We report here five structures of long (34–70 nt) wild-type sequences selected from three cancer-related promoters: *c-Myc*, *c-Kit* and *k-Ras*. Each sequence studied has a unique structure. Three sequences form structures with two contiguous, stacked, G-quadruplex units. One longer sequence from *c-Myc* forms a structure with three contiguous stacked quadruplexes. A longer *c-Kit* sequence forms a quadruplex-hairpin structure. Each structure exhibits interfacial regions between stacked quadruplexes or novel loop geometries that are possible druggable targets. We also report methodological advances in our integrated structural biology approach, which now includes quantitative CD for counting stacked G-tetrads, DNAseI cleavage for hairpin detection and SAXS model refinement. Our results suggest that higher-order quadruplex assemblies may be a common feature within the genome, rather than simple single quadruplex structures.

## INTRODUCTION

DNA and RNA sequences with four runs of typically three consecutive guanine residues separated by 1–7 nucleotides may fold into stable four-stranded structures called G-quadruplexes (G4s) under defined solution conditions (1,2). In G4s, four G residues from different runs assemble to form a planar G-quartet that is stabilized by Hoogsteen hydrogen bonding. These quartets stack to form the G4 core, which is further stabilized by coordination of a monovalent cation to the guanine O6 atoms (reviewed in (3)).

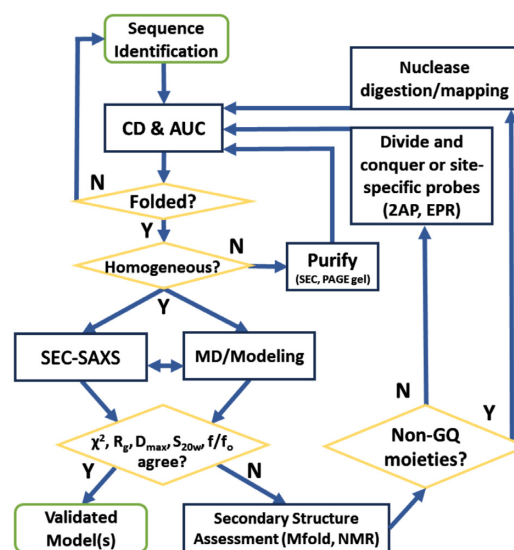
Bioinformatic analysis of a variety of genomes has revealed that oncogenic promoter regions frequently contain tracts of G residues that could potentially fold into quadruplex structures and may regulate adjacent gene transcription (4–6). These findings have been validated by direct *in vitro* and ChIP-sequencing studies (7,8). More recently studies have revealed that promoter quadruplex formation is linked to binding of transcription factors (9) and epigenetic regulation of promoter function in live cells (10). Numerous studies have now validated the concept that ligand-induced G4 stabilization can modulate oncogene expression (reviewed in ref. (5)). For example, the *c-Myc* protein is aberrantly overexpressed in >80% of solid tumors. The *c-Myc* promoter region (NHEIII) contains a potential G4-forming sequence of 27 nucleotides (Pu-27) that contains two G<sub>3</sub> and two G<sub>4</sub> tracts. Ligand stabilization of the *c-Myc* promoter G4 decreases production of *c-Myc* transcripts in cultured cells (11).

To date, most promoter G4 drug discovery efforts have focused on targeting features of the structures of short G-rich sequences that are amenable to characterization by traditional structural biology methods, such as NMR or X-ray crystallography. These sequences are often short (<33 nt),

\*To whom correspondence should be addressed. Tel: +1 502 852 2194; Fax: +1 502 852 7979; Email: john.trent@louisville.edu  
Correspondence may also be addressed to Jonathan B. Chaires. Tel: +1 502 852 1172; Fax: +1 502 852 7979; Email: j.chaires@louisville.edu

heavily modified (e.g. mutations or deletions), and potentially removed from their biological context (e.g. arbitrarily truncated without consideration of adjacent G-tracts, see Supplementary Table S1 (12–43)). Some examples include *c-Myc*(44), *c-Kit* (45), *k-Ras* (46) and *hTERT* (47). It is apparent from the dearth of clinically useful G4 ligands produced by structure-based design that using these simple G4 structures may not be as relevant as drug targets as proposed (48). A possible reason for this limited success is that these ‘well-behaved’ G4s have a paltry repertoire of druggable features. All share a common dominant drug binding site, the terminal G-quartet face. Targeting the G-quartet face often results in selection of planar poly- and hetero-cyclic aromatic compounds that lack optimal drug-like properties, and that bind with high affinity, but little selectivity, to the G-tetrad face (48–50). New avenues for selectively targeting promoter G4s are needed. We hypothesize that longer wild-type promoter sequences can form more complex higher-order structures that might be biologically relevant, and which might contain a richer repertoire of druggable features. These structures might include multiple quadruplexes stacked on one another, or multiple quadruplexes linked by, or including, other secondary structural elements like hairpins. The size of such assemblies is technically challenging for the NMR or X-ray diffraction methods commonly used to determine G-quadruplex structures.

Recent structural studies show that higher-order G4 assemblies exist *in vitro*, and that these more complex structures contain unique binding sites for drug targeting (51–55). It is well established that parallel G4s can stack at the 5′ and 3′ tetrad interfaces (56,57). This arrangement is favorable in the packed conditions of the cell and could be an important regulatory mechanism (58), as these unique structures would provide selective recognition by proteins (59). However, extended G-rich sequences are difficult to study. Often, the number and disposition of G-tracts in promoters suggest the possibility of formation G4 structures with more than the canonical three G-tetrad stack and loop lengths much greater than the traditional 1–4 nucleotides (60). In addition, the presence of multiple G runs can result in formation of G-vacancies or different ‘G-register exchange’ isomers in which different pairings of G residues form a stack (6,61). Such sequence and structural variants, while making structural determination difficult (as there is an ensemble of configurations and potentially topologies), may have biological advantages such as providing an extra G that can substitute for an oxidatively damaged member of a quartet (the ‘spare tire’ hypothesis (62)) or by contributing to the conformational entropy of the folded states, thereby enhancing the probability of G4 formation (61). Further difficulties arise in situations where thermodynamically or kinetically equivalent competing secondary structures exist (22). This plethora of secondary structural possibilities sets the stage for the coexistence of mixtures of G4 topologies. These ensembles of conformers manifest themselves by ill-defined NMR spectra as well as multiple species by SEC, AUC, or electrophoretic experiments (52,60,63–65). Thus, it has been impossible to obtain high-resolution structures of native sequences without resorting to manipulative truncations and mutations to stabilize or create a single conformer at the expense of others (47).



**Figure 1.** Flow chart of the integrative structural biology approach to model higher-order DNA G-quadruplexes. See Table 1 for more description of each technique or experimental property.

Simple G-quadruplexes from short sequences within promoters have been shown by NMR and X-ray crystallography to adopt a variety of topologies including parallel (or ‘propeller’) and antiparallel (‘chair’, ‘basket’ or ‘hybrid’) structures. The parallel conformation facilitates G4 stacking because by necessity the loops project away from the tetrad faces (hence ‘propeller loops’) which allows charge balancing of sugar phosphate backbone repulsion and counterion binding at stacking interfaces (57,66). G4s can also accommodate large loops (>7 nt) (67,68), and when these loops form duplexes they can be stabilizing (69). In their natural context, promoter G4 sequences are flanked at their termini by several nucleotides. It was recently revealed that 5′-flanking bases tend to favor a parallel topology (70). Indeed, nearly all of the high-resolution promoter-derived quadruplex structures flanked at their 5′ ends are parallel (e.g. promoter/PDB IDs: *k-Ras*/5I2V (46), *c-Kit*/2KQG (45), *c-Myc*/2LBY (71), *c-Myc*/6NEB (72), *c-Myc*/1XAV (44), and *VEGF*/2M27 (73)), although exceptions are noted (an *hTERT* promoter G4 with a 5′-A flanking base co-exists as parallel and hybrid 3 + 1 (47)). We have observed this general trend for parallel preference by exhaustively searching the literature for all reported CD spectra of putative promoter G4s (Supplementary Table S1) (12–43). The *in vivo* preference for parallel conformations in promoter G4s is supported by recent ChIP-sequencing studies (74).

To overcome the limitations of high-resolution structural biology techniques in studying extended G-rich sequences, we have developed an integrative structural biology (ISB) platform (75). The integrative approach (Figure 1 and Table 1) uses every available piece of experimental information about a system, in combination with prior structural information and physical theory, to derive self-consistent molecular models that best explain the collective observables. For DNA, each topology comes with a defining spectroscopic signature (NMR, UV, or CD) (76–79) as well as characteristic hydrodynamic and scattering properties including sed-

imentation coefficient ( $S_{20,w}$ ), frictional ratio ( $f/f_0$ ) and radius of gyration ( $R_g$ ) (52,65,80,81). The former informs directly on secondary structural features, while the latter provides coarse grain low- to medium-resolution shape information useful in refining or filtering out inconsistent models. We previously utilized this approach to characterize the human telomerase reverse-transcriptase (*hTERT*) core promoter, a 68-nt sequence with twelve runs of three to five consecutive Gs (52,82). A structure consisting of G4s and WC-hairpin segments was previously proposed based largely on DMS foot-printing experiments (83). Using our ISB approach, however, we demonstrated that such a model was inconsistent with a battery of biophysical measurements, and that the most probable structure for the *hTERT* core promoter was one with three-stacked G4 units, each in a parallel topology (52).

Here we hypothesize that, like *hTERT*, other long promoter sequences may preferentially form stacked parallel G4s. The all-parallel tertiary structure has a distinctive CD signature characterized by a maximum at  $\sim 264$  nm (79,84,85) and we realized that the amplitude of the molar circular dichroism at this wavelength could be used quantitatively to count the number of stacked quartets. We validated and calibrated this by determining the CD spectra of a series of oligonucleotides  $dTG_nT$  ( $n = 3-6$ ) known to form tetrameric, all-parallel G4s in  $K^+$ -containing solutions (86,87). We confirmed that the magnitude of the normalized CD signal at 264 nm for parallel promoters is proportional to the number of stacked G4s. A previous study showed the same relationship for a different set of all-parallel oligonucleotides (88,89). Our calibration curve reveals that the number of stacked parallel G-quartets in an unknown G4 can be determined from its 264 nm CD signal.

We also developed a new independent validation method for the ISB approach by examining the agreement between experimental scattering with theoretical scattering curves calculated from models using the program *CRY SOL*. We measured the solution scattering properties of 14 promoter and artificial G4s (Supplementary Table S4) from the Protein Data Bank (PDB) and compared their measured radii of gyration ( $R_g$ ) and scattering with that of their theoretical values based on their deposited structures. We found an excellent agreement in both cases and subsequently demonstrate how this analysis can and should be used as an additional tool to assess putative molecular models.

Finally, we added the use of secondary structure prediction for the longer promoter sequences, as localized competing structures become more likely with longer sequences and loop regions. To do this we implemented a simple DNaseI cleavage assay to test these predictions.

We used this improved ISB platform to evaluate the extended promoter sequences identified by *Quadparser* (2) in the *c-Kit*, *k-Ras* and *c-Myc* promoters (Table 2). Hydrodynamic and scattering data confirm that all higher-order sequences form secondary structures. All sequences are consistent with preferential formation of parallel stacked topologies. *In silico* construction of models refined by CD, AUC and SAXS data reveals that the three 8-tract sequences form highly compact parallel stacked G4s, consistent with the parallel promoter hypothesis. The *c-Kit* 12-tract sequence, which encompasses its 8-tract counterpart,

exhibits a similar compact parallel G4 region with an extended GC duplex hairpin feature, which was confirmed by SAXS and DNaseI cleavage experiments. Lastly, the 12-tract *c-Myc* sequence predominantly forms a globular parallel stacked structure but has competing hairpin features that obscure its analysis by any singular method. Here we demonstrate how our expanded ISB platform can be used to study even the most recalcitrant higher-order DNA G4 systems, and in doing so reveal novel loop and junctional topologies that might be useful in drug discovery efforts.

## MATERIALS AND METHODS

### Bioinformatic analysis

*Quadparser* software was downloaded from the Balasubramanian group (2). The *Homo sapiens* genome (December 2013 GRCh38/hg38) Eukaryotic Promoter Database was used for the to generate the promoter sequences from  $-499$  to  $+100$  and  $-750$  to  $+100$ , which were searched separately. We set the search parameters to identify four, eight, and twelve runs of two or three guanines with 1–10 loop residues. The results are given in Supplemental Table S2.

### Oligodeoxynucleotides and G4 formation

Oligos (Table 2) were obtained from Eurofins (Louisville, KY) or IDT (Coralville, IA) as lyophilized, desalted powders. Stock solutions of approximately 1 mM were prepared in Milli-Q  $H_2O$ , warmed for  $\sim 30$  min at  $50^\circ C$  to facilitate solubilization, and stored at  $4^\circ C$ . Oligo concentration was estimated from the absorbance at 260 nm determined at either pH 11 or  $90^\circ C$  using their extinction coefficients. Working solutions were prepared by diluting the stock oligo solution to the desired strand concentration in the respective buffer. All extended promoter samples were purified by preparative size-exclusion chromatography (SEC) (Superdex 75 16/600 SEC column, GE Healthcare 28-9893-33, running at 0.5 ml/min, fractions collected every 2 min) and concentrated with Pierce protein concentrators (ThermoFisher, #88515) prior to analysis. Two separate buffers were used throughout: TBAP (tBAP, tetrabutyl ammonium phosphate) and BPEK (potassium phosphate). Buffers contained 1 mM EDTA and had a pH of either 6.8 (TBAP) or 7.2 (BPEK) and were supplemented with varying levels of KCl (25–200 mM) as indicated. Samples were annealed by heating for 10 min. in 1 l boiling water followed by slow cooling overnight to room temperature. To ensure complete and rapid formation of  $d[TG_nT]_4$  tetramers, 1 mM solutions of the  $dTG_nT$  oligos were supplemented with  $10\times$  BPEK to a final buffer concentration of  $1\times$  BPEK followed by overnight incubation at  $4^\circ C$  (90). The samples were not heated. Preliminary experiments revealed that tetramer formation occurred only when folding was initiated by  $K^+$  addition to oligos at high concentration. Once tetramers were formed at mM oligo concentration, the tetrameric aggregation state was stable after dilution to  $\mu M$  working concentrations as verified by analytical ultracentrifugation (65,80).

### CD spectra

Baseline-corrected, normalized CD spectra were recorded in 1-cm quartz cuvettes with Jasco J710 or J810 spectropo-

**Table 1.** Table of experimental techniques used in the integrative approach and corresponding qualitative and quantitative information gained from each type of analysis

Experimental technique (Ref)	Qualitative information	Quantitative information	Model refinement/assessment
CD (78) AUC(80)	Topology Foldedness from frictional ratio ( $f/f_0$ ), oligomeric state(monomer/dimer/oligomer)	# parallel G-tetrad stacks Molecular weight (MW), Sedimentation coefficient ( $S_{20,w}$ ), translational diffusion coefficient ( $D_t$ )	<i>HYDROPRO</i> (104) calculated $S_{20,w}$
SEC (127)	Oligomeric state (monomer/dimer/oligo/aggregate)	Stokes radius ( $R_s$ ) or molecular weight (MW)	
SAXS(102,109)	Foldedness, flexibility, shape	Radius of gyration ( $R_g$ ), maximum particle dimension ( $D_{max}$ ), volume	<i>CRY SOL</i> (102) calculated scattering, reduced $\chi^2$ (Eq. 1 in methods), calculated $R_g$ , <i>HYDROPRO</i> (104) calculated volume & $D_{max}$ , <i>ab initio</i> reconstructions (96–99)
$^1\text{H}$ NMR (128,129)	Qualitative assessment of duplex/quadruplex imino proton shifts	Amount of Watson-Crick or Hoogsteen bonds	
Nuclease Digestion (52)	Topological changes (when monitored by CD)	Changes in MW, $S_{20,w}$ , $f/f_0$ , $R_g$ , $D_{max}$	<i>CRY SOL</i> (102) calculated scattering, reduced $\chi^2$ (Eq. 1 in methods), calculated $R_g$ , <i>HYDROPRO</i> (104) calculated volume & $D_{max}$ , <i>ab initio</i> reconstructions (96–99)
Site-specific Probes (130,131)	Relative solvent exposure	Residue-residue distances	SAS calculations, distance calculations

**Table 2.** Oligonucleotides used in this study with predicted number of G-tetrad stacks based on CD 264 nm amplitude. See Supplementary Table S2 for sequences used in SAXS  $R_g$  analysis

ODN Name	Sequence (5'→ 3')	# nt	# G-Stacks	$\epsilon$ ( $\text{M}^{-1} \text{cm}^{-1}$ )	MW (Da) (strand)	CD $\Delta\epsilon_{264}$ ( $\text{M}^{-1} \text{cm}^{-1}$ )	Predicted # G-stacks
TG3T	TGG GT	5	3	471,500	1534		
TG4T	TGG GGT	6	4	576,200	1863		
TG5T	TGG GGG T	7	5	680,900	2193		
TG6T	TGG GGG GT	8	6	785,600	2522		
1XAV	TGA GGG TGG GTA GGG TGG GTA A	22	3	228,700	6992	190	2.9
2LBY	TAG GGA GGG TAG GGA GGG T	19	3	201,700	6054	206	3.1
2M27	CGG GGC GGG CCT TGG GCG GGG T	22	3	200,400	6905	220	3.2
5I2V	AGG GCG GTG TGG GAA TAG GGA A	22	3	233,100	6970	182	2.8
c-Myc-8	GGG GAG GGT GGG GAG GGT GGG GAA GGT GGG GAG G	34		354,900	10976	361	4.7
c-Myc-12	GGG AAC CCG GGA GGG GCG CTT ATG GGG AGG GTG GGG AGG GTG GGG AAG GTG GGG AGG AGA CTC AGC CGG G	70		702,100	22255	705	8.3
c-Kit-8	GGG CGG GCG CGA GGG AGG GGA GGC GAG GGG CGT GG	38		381,800	12143	317	4.2
c-Kit-12	GGG CGG GCG CGA GGG AGG GGA GGC GAG GGG CGT GGC CGG CGC GCA GAG GGA GGG CGC TGG G	64		626,800	20349	339	4.5
k-Ras-8	GGG AGC GGC TGA GGG CGG TGT GGG AAG AGG GAA GAG GGG GAG <u>G*</u>	43		445,300	13755	276	3.8

\*Underlined region of k-Ras-8 overlaps with sequence studied in ref (76).

larimeters using the protocol outlined by Del Villar (78). DNaseI digestions assays were conducted in 0.5-cm quartz cuvettes with a Jasco J710 as detailed previously (52). In brief, samples at 12  $\mu\text{M}$  were annealed in EDTA-free TBAP buffer with 185 mM KCl, and then mixed with 4 $\times$  DNaseI reaction buffer (80 mM Tris, 8 mM  $\text{MgCl}_2$ , 40 mM KCl, pH 7.2), and MilliQ  $\text{dH}_2\text{O}$  in a 2:1:1 ratio to achieve a final G4 concentration of 3  $\mu\text{M}$  in a total of 500  $\mu\text{l}$  volume. Scans were acquired every 5 min for  $\sim$ 4 h after the addition

of 50  $\mu\text{l}$  of amplification grade DNase (1 unit/ $\mu\text{L}$ ) added immediately after the first scan.

### Analytical ultracentrifugation

Sedimentation velocity measurements were carried out in a Beckman Coulter ProteomeLab XL-A analytical ultracentrifuge (Beckman Coulter Inc., Brea, CA) at 20.0°C and at 50 000 rpm in standard two sector cells. Data

(100 scans collected over an 8-hour centrifugation period) were analyzed using the program *SEDFIT* in the continuous  $c(s)$  model ([www.analyticalultracentrifugation.com](http://www.analyticalultracentrifugation.com)). Buffer density was determined on a Mettler/Paar Calculating Density Meter DMA 55A at 20.0°C and buffer viscosity was measured on an Anton Paar Automated Microviscometer AMVn. For the calculation of frictional ratio and molecular weight, 0.55 ml/g was used for partial specific volume (65).

### Molecular modeling

A generic parallel G-quartet stack was built by superimposing the parallel quadruplex structure 1XAV to build a 12-tetrad stacked parallel G-tetrad model with removal of the loops. The appropriate maximum number of G-tetrads, as determined by CD and the sequence, were used to create the central stacked models. Multiple stacked parallel quadruplexes were created, and the loop sequences were manually inserted to minimize the loop length when contiguous guanines in G-runs were greater than the individual number of G-tetrads of a given quadruplex. Potassium ions were added between quartets, and the initial structures were minimized (implicit water solvation, AMBER\* force field *Macromodel*, Schrodinger Inc., <https://www.schrodinger.com/>), whilst restraining the G-tetrads. The models then underwent full AMBER minimization and molecular dynamics using our standard protocol. The models were imported into the xleap module of *AMBER 2018* with the default force field, ff14SB and OL15 DNA force field, neutralized with  $K^+$  ions, and solvated in a rectangular box of TIP3P water molecules with a 15 Å buffer distance. All simulations were equilibrated using sander using the following steps: (i) minimization of water and ions with restraints of 10.0 kcal/mol/Å on all nucleic acid and amino acid residues (2000 cycles of minimization, 500 steepest decent before switching to conjugate gradient) and 10.0 Å cutoff, (ii) heating from 0 K to 100 K over 20 ps with 50 kcal/mol/Å restraints on all nucleic acid and amino acid residues, (iii) minimization of entire system without restraints (2500 cycles, 1000 steepest decent before switching to conjugate gradient) with 10 Å cutoff, (iv) heating from 100 K to 300 K over 20 ps with restraints of 10.0 kcal/mol/Å on all nucleic acid and amino acid residues and (v) equilibration at 1 atm for 100 ps with restraints of 10.0 kcal/mol/Å on nucleic acids. The output from equilibration was then used as the input file for 100 ns of unrestrained MD simulations using pmemd with GPU acceleration in the isothermal isobaric ensemble ( $P = 1$  atm,  $T = 300$  K). Periodic boundary conditions and PME were used. 2.0 fs time steps were used with bonds involving hydrogen frozen using SHAKE ( $ntc = 2$ ). Trajectories were analyzed using the cpptraj module in the *AmberTools 18* package. Accelerated molecular dynamics under the same conditions were also performed for 100 ns trajectories. All systems were stable throughout the production phase.

Hydrodynamic properties were calculated using 500 equally spaced snapshots across the entire trajectory. This was accomplished using *HYDROPRO10* using an atomic level calculation (INMODE = 1, AER = 2.53) with  $vbar = 0.55$ . All *HYDROPRO* calculations used tem-

perature of 20.0°C, viscosity = 0.0101 poise and density = 1.0092 g/cm<sup>3</sup>.

### SEC-resolved small-angle X-ray scattering (SEC-SAXS)

All samples analyzed by small-angle X-ray scattering (SAXS) were in BPEK buffer supplemented with 185 mM KCl, purified by preparative SEC (Superdex 75 16/600 SEC column, GE Healthcare 28-9893-33, running at 0.5 ml/min, fractions collected every 2 min), concentrated with Pierce protein concentrators (ThermoFisher, #88515), and dialyzed (Spectra/Por Float-A-Lyzers G2 3.5 kDa, Sigma #Z726060) prior to SAXS analysis. SEC-SAXS was performed at the BioCAT beamline (18ID) at the Advanced Photon Source in Chicago, IL. Prepared samples were centrifuged and subsequently loaded onto an equilibrated Superdex 200 Increase 10/300 GL column (Cytiva) maintained at a flow rate of 0.6 or 0.7 ml/min (see Supplementary Tables S3 and S4) using an AKTA Pure FPLC (GE Healthcare Life Sciences). After passing through the UV monitor, the eluate was directed through the SAXS flow cell, which consists of a 1 mm ID quartz capillary with 20 μm walls. A co-flowing buffer sheath was used to separate the sample and the capillary walls, helping to prevent radiation damage (91). Scattering intensity was recorded with a Pilatus3X1M (Dectris) detector placed 3.628 m from the sample, giving access to a  $q$ -range of 0.0044–0.35 Å<sup>-1</sup>. A series of 0.5 s exposures were acquired continuously during elution and the data was reduced using the software *BioXTAS RAW versions 1.6.3, 2.0.3 or 2.1.1* (92). Buffer blanks were created by averaging regions flanking the elution peak and subtracted from exposures selected from the sample elution peak to create the buffer corrected  $I(q)$  vs.  $q$  curves for subsequent analyses. A few of the monomeric G4 sequences (PDB IDs 2KQG, 2LBY, 2M27, 6GH0, 6L92) eluted as oligomeric species and required evolving factor analysis (EFA) (93,94) to retrieve the monomer scattering profile. Singular value decomposition (SVD) and EFA are both standard integrated data deconvolution methods that are integrated in *BioXTAS RAW* (more information on the use of these methods can be found at <https://bioxtas-raw.readthedocs.io/en/latest/>). SAXS sample preparation, data collection, data reduction, analysis, presentation, and interpretation have been done in close accordance with recently published guidelines (95). Tabulated results and elution/ $R_g$  profiles from our SAXS analyses can be found in Supplementary Tables S3, S4 and Supplementary Figures S1–S22. All SAXS data have been deposited in the SASBDB (<https://www.sasbdb.org/>).

Generation of SAXS space-filling envelopes was accomplished using DAMMIF (96) in slow mode with 20 reconstructions (no symmetry or anisometry assumptions) followed by averaging and clustering using DAMAVER (97) and DAMCLUST (98), respectively. The output ‘damstart.pdb’ from DAMAVER was subsequently used as input for a final refinement in DAMMIN (99). The input  $P(r)$  distribution files were generated using the program *GNOM 4.6* (100) in RAW v2.1.1 (92) and truncated to the recommended 0.3  $q$ . See Supplementary Tables S3 and S4 for the normalized spatial discrepancy values (NSDs),  $\chi^2$  values, and resolutions via SASRES (101). The best-fit

molecular models were determined using *CRY SOL* v2.8.3 (102) command line interface to perform calculations across 1000 evenly spaced frames from each 100 ns standard MD trajectory. The best fit structure was determined by minimization of a  $\chi^2$  function:

$$\chi^2(r_o, \delta_\rho) = \frac{1}{N_p} \sum_{i=1}^{N_p} \left( \frac{I_{exp}(q_i) - cI(q_i, r_o, \delta_\rho)}{\sigma(q_i)} \right)^2 \quad (1)$$

where  $I_{exp}(q_i)$  and  $I(q_i)$  are the experimental and computed profiles, respectively,  $\sigma(q_i)$  is the experimental error of the measured profile,  $N_p$  is the number of points in the profile, and  $c$  is the scaling factor. Two other parameters,  $r_o$  and  $\delta_\rho$ , are fitted and represent the effective atomic radius and the hydration layer density, respectively. The solvent electron density used in *CRY SOL* calculations was adjusted for the buffer components from 0.334 to 0.3368  $e^-/\text{\AA}^{-3}$ . Best fit atomistic models were docked to their most probable space-filling envelopes using *SUBCOMB* (103) and visualized in *Chimera* v1.12.

Sedimentation coefficient calculations from SAXS *ab initio* *DAMMIF/N* refined bead models were done in the following way. First, theoretical volumes were calculated for each system with the following equations:

$$V_{anhydrous} (\text{\AA}^3) = M(Da) \times \frac{\left(0.55 \left(\frac{cm^3}{g}\right)\right) \times \left(10^{24} \left(\frac{\text{\AA}^3}{cm^3}\right)\right)}{6.023 \times 10^{23} \left(\frac{Da}{g}\right)}, \quad (2)$$

$$\left(\left(\frac{\delta}{\rho \times v}\right) + 1\right) \times V_{anhydrous} \quad (3)$$

where  $\rho$  is the solution density and the partial specific volume,  $v = 0.55 \text{ cm}^3/\text{g}$  was assumed. Volumes were then adjusted to reflect an assumed hydration of  $\delta = 0.3 \text{ g/g H}_2\text{O}$  (64) (Equation 3). For lower-resolution SAXS models, *HYDROPRO10* (104) recommends that calculations be performed with *INDMODE* = 1 and an AER (hydrodynamic radius of elements in primary model) value that results in agreement between calculated and expected particle volume. The adjusted AER should also yield  $R_g$  in good agreement with that measured by SAXS. Starting values for AER were derived from the *DAMMIF/N* refined model header under 'average volume per atom',  $V_a$ , and the following equation for radius of a sphere:

$$AER = \sqrt[3]{\frac{3 \times V_a}{4 \times \pi}} \quad (4)$$

The resulting AER values were then fed into *HYDROPRO* and, if necessary, adjusted until the calculated volumes were within 5% of expected values. In each case, the resulting  $R_g$  values were within 4% of their measured value. Results are tabulated in Supplementary Table S5.

## RESULTS

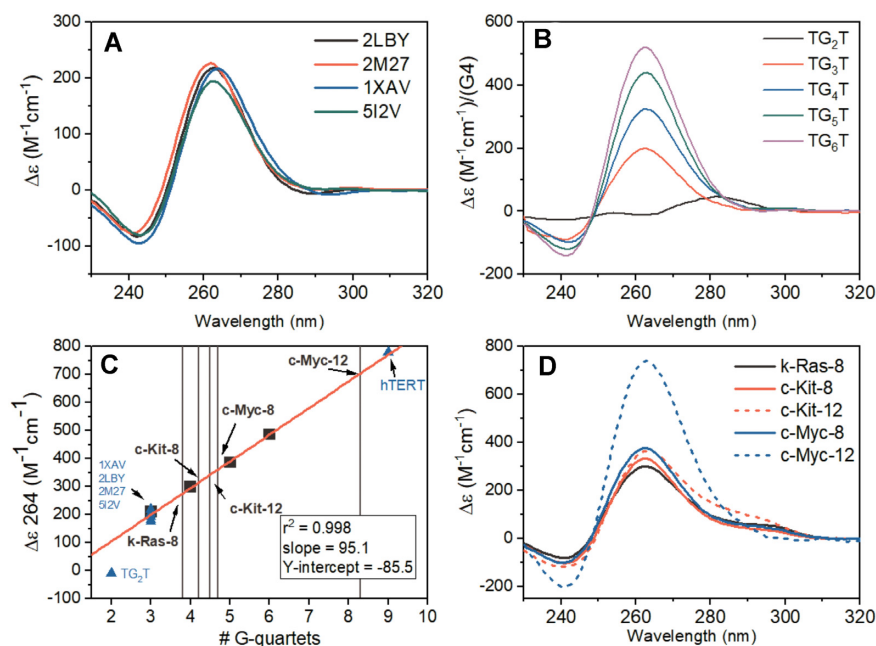
### Bioinformatic queries show that long sequences might form higher-order G4 structures and are abundant in promoter regions of the human genome

We used *Quadparser*(2) to search human promoter sequences between  $-750$  or  $-499$  to  $+100$  relative to the transcriptional start sites for various combinations of two or three contiguous guanines in runs of 4, 8, and 12 tracts interspersed by loop lengths ranging from 1–7 to 1–10 (Supplementary Table S2). We searched for the motifs  $[G_x-L_y-G_x-L_y-G_x]_z$ , where  $x = 2, 3$ ,  $y = 1-7, 1-8, 1-9, 1-10$ ,  $z = 1, 2, 3$  and where an additional loop,  $L_y$  is included between each motif block when  $z = 2$  and 3. Such sequences might form multiple G4 structures. We found hundreds of thousands of potential G4 forming sequences identified with over 56 000 promoter sequences with eight tracts of  $G_{2,3}$  and over 20 000 with 12 tracts of  $G_{2,3}$  with loops of 1–10 bases. From the abundance of these promoter sequences, we chose 8- and 12-tract sequences from the oncogene promoters of *c-Myc* ('c-Myc-8', 'c-Myc-12'), *c-Kit* ('c-Kit-8', 'c-Kit-12'), and *k-Ras* ('k-Ras-8') to characterize on the basis of their disease relevance (Table 2). What follows is our application of our ISB approach to characterize the structure of these sequences.

### The CD 264 nm peak amplitude correlates with G-quartet number in stacked parallel intramolecular quadruplexes

We observed that parallel three-tetrad G4s exhibit similar spectral shapes and CD 264 nm amplitudes ( $\sim 200 \Delta\epsilon$ ) (Figure 2A). The amplitude of a single G4 is approximately 1/3 to 1/4 of the nine-tetrad stack parallel hTERT promoter quadruplex ( $\sim 750 \Delta\epsilon$ ) (52). To determine if the magnitude of the 264 nm CD signal is correlated with the number of stacked G-tetrads, we measured the CD spectra of a series of well-characterized, all-parallel tetrameric G4s with different numbers of stacked quartets that form with the sequences  $d[\text{TG}_n\text{T}]$  ( $n = 3-6$ ) (Figure 2B). To ensure formation of  $d[\text{TG}_n\text{T}]_4$  rather than misfolded structures, it was necessary to initiate G4 formation by adding  $\text{K}^+$  to a concentrated ( $\sim 1 \text{ mM}$ ) solution of monomeric oligonucleotide (90). This procedure ensures rapid, in-register formation of tetrameric G4s rather than G-wires or other misfolded structures. The homogeneity of these tetramolecular structures was confirmed by molecular weights and sedimentation distributions by analytical ultracentrifugation. There is a clear linear relationship between the 264 nm ellipticity normalized with respect to the number of contiguous G-tetrad stacks per tetramer (Figure 2C). In addition, the normalized CD amplitude determined for the 3-tetrad, parallel promoter G4s of known structure all fall on the calibration line. These include: *k-Ras*, 5I2V (46); *c-Myc*, 2LBY (71); *c-Myc*, 1XAV (44); *VEGF*, 2M27 (73) and *hTERT* (52). Linear least-squares fitting of the  $d[\text{TG}_n\text{T}]_4$  data provided a slope  $m = 95.1 \pm 3.8 \text{ M}^{-1} \text{ cm}^{-1}/\text{quartet}$  and an intercept  $b = -85.5 \pm 16.4 \text{ M}^{-1} \text{ cm}^{-1}$ , with a correlation coefficient  $r^2 = 0.998$ .

Figure 2D shows the normalized CD spectra of the five putative higher-order promoter G4 sequences, c-Myc-8, c-



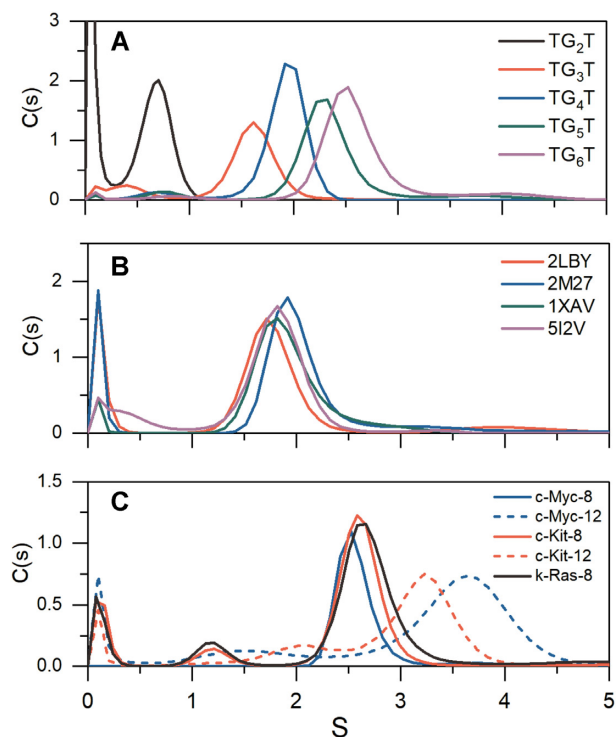
**Figure 2.** (A) CD spectra of parallel promoter G4s with three G-quartet stacks from the literature. (B) CD spectra of d[ $TG_nT$ ]<sub>4</sub> oligos normalized to G4 concentration. (C) Linear relationship of the 264 nm CD signal normalized by G4 concentration of parallel d[ $TG_nT$ ]<sub>4</sub> G4s (black squares) versus number of contiguous stacked G-quartets. The line of best fit is shown in red. The blue triangles show the 264 nm magnitudes for the promoter sequences in (A), the magnitude of the 9-quartet hTERT (52), and the  $TG_2T$  sequence (which does not form a G4). The light gray vertical lines cross the regression at the 264 nm values obtained for the extended promoter 8-tract and 12-tract sequences (shown in D). (D) CD spectra of the 8-tract and 12-tract promoter sequences studied here.

Myc-12, c-Kit-8, c-Kit-12 and k-Ras-8, identified in our bioinformatic inquiry. In each case, the CD spectrum exhibits a maximum near 264 nm and minimum at 240 nm, consistent with parallel G-quadruplex formation. The number of G-tetrads per promoter strand estimated from the slope of the calibration curve in Figure 2C is summarized in Table 2. The presence of a non-integral number of stacks could reflect some slight structural heterogeneity. The 3-tetrad, parallel promoter G4s in Figure 2A show a 264 nm deviation about the linear regression fit of up to 7%. Each of the extended promoter sequences, aside from c-Kit-12, are within 7% of their integral G-tetrad number, i.e. c-Kit-8 and k-Ras-8 have ~4 G-tetrads, c-Myc-8 has ~5 G-tetrads, and c-Myc-12 has ~8 G-tetrads. The 295 nm shoulder in the c-Kit-12 spectrum raises the possibility that its 264 nm value is influenced by the presence of other structures, either an antiparallel quadruplex, a hairpin, or a combination thereof (as shown below). By enumerating the number of stacked quartets in the structures formed by each promoter sequence, these CD results provide quantitative constraints for model building, an important first step in the integrated structural strategy.

#### Analytical ultracentrifugation sedimentation velocity (AUC-SV) studies assess homogeneity, compactness and hydrodynamic shape of G-quadruplexes

After sequence identification and characterization by CD, the next step in our ISB workflow is to assess folding of each sequence into a discrete structure and the homogeneity of samples (Figure 1 and Table 1). AUC-SV is an ideal

tool for this purpose, since it can provide unambiguous evidence of heterogeneity, estimates of the molecular weights of all species present, and low-resolution shape information (105). As demonstrated in Figure 3A, the tetrameric G4s (d[ $TG_nT$ ]<sub>4</sub>) are highly stable, even on 50-fold dilution to working concentrations suitable for spectropolarimetry. The C(s) distribution analysis shows >95% tetramolecular species for each of the d[ $TG_nT$ ]<sub>4</sub> series used for calibration, confirming that the CD signals in Figure 2 arise from the expected tetrameric species and are not influenced by aggregates or unfolded single strands. Similarly, the 3-stack promoter parallel G-quadruplexes *k-Ras* 5I2V (46), *c-Myc* 2LBY (71), *c-Myc* 1XAV (44) and *VEGF* 2M27 (73) (Figure 3B) are all folded and homogeneous. Figure 3C, shows the C(s) distributions of the extended promoter sequences. All AUC-derived molecular weights are within 10% of their true molecular weights. Qualitatively, the distributions show that the 8-tract promoter sequences exhibit sedimentation coefficients close to those of the 5- and 6-tetrad d[ $TG_nT$ ]<sub>4</sub> sequences (which are of similar MW), and much higher than the 3-stack promoter G4s, consistent with compact unimolecular folded species. We note that in Figure 2D, the c-Kit-12 CD spectrum is similar in 264 nm magnitude to the 8-tract promoter G4s, indicating similar number of tetrad stacks, yet it exhibits a sedimentation coefficient that is in between the 8-tract promoters and c-Myc-12. The measured frictional ratios ( $f/f_0$ ) for all sequences are <1.5 confirming that all sequences contain secondary structure and are not random coils (Table 3) (80). The AUC-SV results provide additional quantitative constraints for model building.



**Figure 3.** Sedimentation velocity profiles for G4s. (A) shows  $d[\text{TG}_n\text{T}]_4$  oligos. (B) shows the promoter G4s from Figure 1A, and (C) shows the c-Myc, c-Kit and k-Ras higher-order promoters. Tabulated frictional ratios and sedimentation values corrected to reflect water at 20°C ( $S_{20,w}$ ) are given in Table 3.

### Small-angle X-ray scattering measures the global shape of G4 structures

SAXS is a powerful technique for characterizing complex higher-order G-quadruplex systems (52,81). SAXS provides both qualitative and quantitative structural information, including particle shape, compactness, volume, maximum particle dimension ( $D_{max}$ ) and radius of gyration ( $R_g$ ) (106,107). Scattering data are especially powerful when combined with molecular modeling (95,108), as scattering patterns may be readily computed from *in silico* models with programs such as *CRY SOL* (102). In this way, models can be refined directly against ‘medium’ (~18–30 Å) resolution experimental structural information (107–109).

The most important consideration prior to the interpretation of SAXS data is that there is no aggregation, radiation damage, or interparticle interactions (95). To ensure this, SAXS measurements were made as a function of elution time from an SEC column with a co-flowing buffer sheath to mitigate X-ray damage. The final scattering profiles of the 8- and 12- tract promoter sequences were monodisperse based on linearity of Guinier regression analysis (Supplementary Figures S1–S7), in agreement with AUC analysis (although we note minor amounts of larger and extended species are evident in c-Myc-12 and c-Kit-12 scattering and AUC profiles, respectively) (95). The 8- and 12-tract sequence SAXS results are shown in Figure 4 and a description of the data collection, reduction, and analy-

sis are given in Supplementary Table S3. We included the analysis of 1XAV to serve as a ‘control’ for a globular particle and to aid in comparison of some key features of the results (2LBY, 2M27 and 5I2V were also analyzed but not shown; their scattering results are in Supplementary Table S4 and Supplementary Figures S8, S14, S15, S17). Figure 4A and D shows that each sequence has scattering that is horizontal and parallel to the X-axis at low  $q$ , supporting that the data are free from artifacts due to inter-particle interactions. The data are presented on a log–log scale to highlight the smooth curvature at  $\sim 0.1 q$ . This smooth curvature is characteristic of globular particles (107), and all but c-Kit-12 exhibit a rounded decay at higher values of  $q$ .

The pair distance distribution functions, or  $P(r)$ , plots for the extended promoter sequences, c-Kit-8, c-Myc-8, k-Ras-8, c-Myc-12 and c-Kit-12 are shown in Figure 4B and E. The  $P(r)$  plot is an  $r^2$ -weighted real-space histogram of interatomic distances derived from the scattering by an indirect Fourier transform (107).  $P(r)$  distributions that are symmetric and Gaussian are indicative of globular shapes, as exemplified by 1XAV (Figure 4B, green) (106,107). Deviations from Gaussian shape, such as skewing or multiphasic distributions, indicates deviation from a globular particle, e.g. asymmetric oblate or prolate particles, multi-domains, or regions of disorder (107). Quantitative information is also gained from the  $P(r)$  plot. The maximum dimension of the particle, or  $D_{max}$ , is where the curve intercepts the X-axis. The radius of gyration,  $R_g$ , is an overall size estimate useful for comparisons with calculations from theoretical atomistic models. The  $R_g$  can be determined from the second moment of the  $P(r)$  distribution (which is sometimes more reliable than the Guinier approximation) (106,107).  $D_{max}$  and  $R_g$  values for each sequence are tabulated in Supplementary Table S3. Figure 4B shows the  $P(r)$  distributions, normalized to  $I(0)$ , for 1XAV, c-Myc-8, c-Kit-8 and k-Ras-8. In each case, the 8-tract promoter sequences are globular and of larger dimension than 1XAV, based on their slight positive skew. There is good agreement between Guinier and  $P(r)$ -derived  $I(0)$  and  $R_g$  values for the 8-tract sequences, consistent with a globular and folded particle (Supplementary Table S3). In contrast, the 12-tract promoter sequences (Figure 4E) exhibit significant positive skew with a gradual decline to  $D_{max}$ . The difference of 2–5% in the measured Guinier and  $P(r)$ -derived  $I(0)$  and  $R_g$  values (Supplementary Table S3) indicates some amount flexibility (106). The latter may be attributed to large loop regions, co-existing G-register isomers, or general unstructured regions (61). In the case of c-Kit-12, a distinct shoulder is evident in the  $r$ -range of  $\sim 40$ – $65$  Å which is characteristic of a multi-domain particle (107).

Another qualitative assessment of particle compactness and flexibility is a mathematical transformation of the scattering data into a Kratky plot (Figure 4C and F) (106). Scattering from a spherical particle decays rapidly at large angles,  $I(q) \sim 1/q^4$ , and so by plotting the scattering as  $q^2 \cdot I(q)$  versus  $q$  a Gaussian profile is expected for globular species (106). To make interpretation more semi-quantitative, the Kratky plot can be made ‘dimensionless’ by multiplying  $q$  by  $R_g$ , to account for particle size, and scaling the inten-



**Table 3.** Summary of analytical ultracentrifugation and HYDROPRO results. For the higher-order 8- and 12-tract promoters (c-Kit, k-Ras, and c-Myc) the HYDROPRO calculated results are reflective of only the final best fit all-parallel model builds. SAXS *ab initio* HYDROPRO calculations were done using the final refined models from DAMMIN and an AER that was adjusted until the HYDROPRO calculated volumes and radii of gyration were within ~5% of the expected values (from calculation or SAXS measurements, respectively) (see Supplementary Table S5)

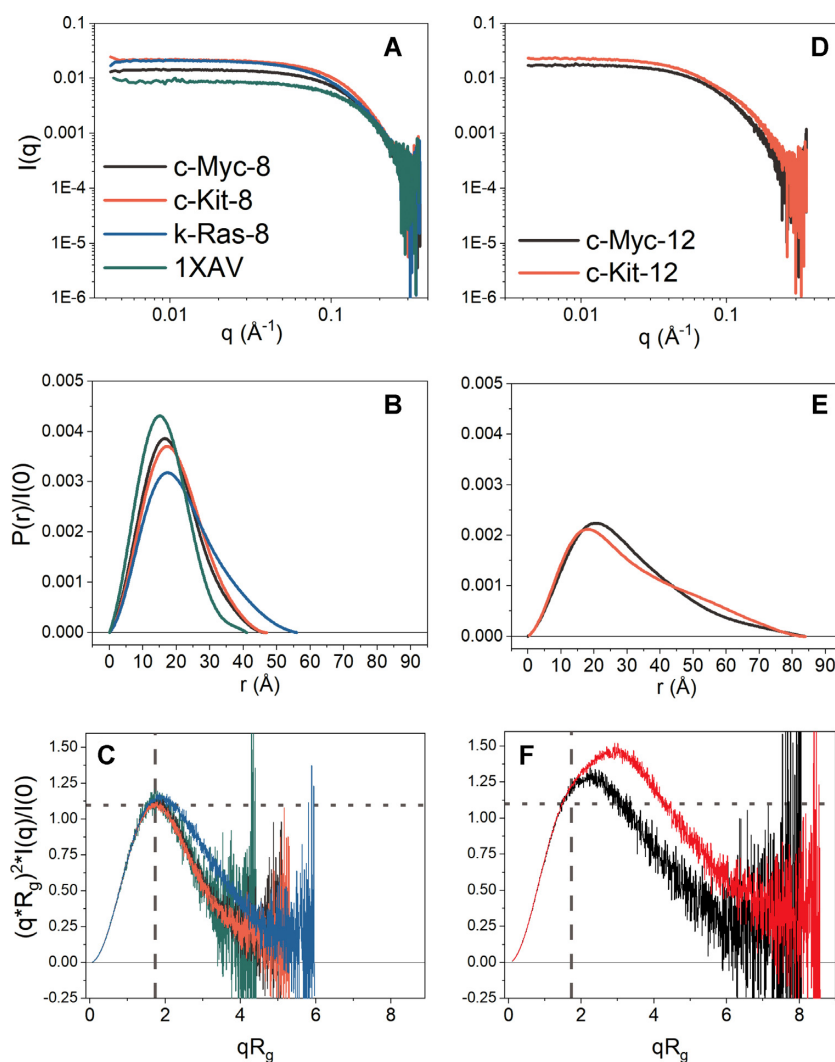
Sequence	G4 molecularity	AUC MW (kDa)	True MW (kDa)	$f/f_0$	$S_{20,w}$ observed	$S_{20,w}$ calculated (all atom)	$S_{20,w}$ calculated (SAXS <i>ab initio</i> )
TG3T	(tetramer)	6.6	6.14	1.32	1.75		
TG4T	(tetramer)	7.7	7.45	1.25	2.05		
TG5T	(tetramer)	10.2	8.77	1.27	2.44		
TG6T	(tetramer)	11.3	10.09	1.24	2.67		
1XAV	(monomer)	8.3	6.99	1.38	1.86		
2LBY	(monomer)	7.1	6.05	1.38	1.72		
2M27	(monomer)	7.9	6.91	1.24	2.10		
5I2V	(monomer)	7.9	6.97	1.49	1.76		
c-Myc-8	(monomer)	12.0	10.98	1.33	2.60	2.16–2.52	2.53
c-Myc-12	(monomer)	22.3	22.25	1.46	3.57	3.68–3.78	3.04
c-Myc-12 (Post DNaseI digestion)	(monomer)	20.6	-	1.41	3.61	-	3.26
c-Kit-8	(monomer)	12.9	12.14	1.34	2.70	2.08–2.61	2.46
c-Kit-12	(monomer)	19.2	20.34	1.42	3.33	3.38–3.73	2.96
c-Kit-12 (Post DNaseI digestion)	(monomer)	14.7	-	1.42	2.85	-	2.82
k-Ras-8	(monomer)	14.6	13.75	1.4	2.83	2.79–2.85	2.56

sity by  $I(0)$ , to account for mass. For a globular particle, this transformation yields a peak at X- and Y-dimensions of  $qR_g = \sqrt{3} = \sim 1.75$ , and  $3/e = 1.104$ , respectively (110). Fully unfolded particles will not return to zero at high  $qR_g$ , but instead continue rising to a plateau at constant value of 2 (110). Structured particles with flexibility are expected to exhibit intermediate curves between these two extremes, e.g. a peak position shifted positively in both X- and Y-directions. Figure 4C and F show the dimensionless Kratky plots for each of the promoter sequences and 1XAV. 1XAV exhibits a Kratky profile with classical symmetric Gaussian shape about  $\sim 1.73 qR_g$ , with peak height of  $\sim 1.1$ , consistent with its globular shape. Similarly, the 8-tract sequences exhibit curves that are bell-shaped, albeit slightly skewed relative to 1XAV (Figure 4C), indicating slight deviation from an entirely globular shape. In contrast, c-Myc-12 exhibits a highly skewed profile that may signify a globular shape with protruding flexible domain (Figure 4F). c-Kit-12 also exhibits a skewed globular profile; however, a notable monotonic increase is evident from  $\sim 1.8$ – $3.5 qR_g$ , signifying that it is multi-domain and flexible (Figure 4F). Altogether, these results are consistent with the AUC analyses in that all 8-tract sequences adopt folded, compact tertiary structures and that the 12-tract sequences are overall globular but more complex than a simple oblate or prolate particle in their tertiary structure.

#### Validation of CRYSOLE $R_g$ calculations as an ISB tool for filtering models of inconsistent size

To date, no studies have systematically evaluated how well the  $R_g$  values of atomistic models of G4s compare with their measured values. In a similar manner to the above CD analysis and our prior hydrodynamic bead modeling studies (65), we sought to determine the degree to which  $R_g$  values calculated from atomic models agree with their SAXS-measured radii of gyration. To examine this, we col-

lected SAXS data on 14 sequences from monomeric G4s deposited in the Protein Data Bank that had been studied in  $K^+$ -containing buffers, including telomere, promoter, and artificial sequences (13 from NMR, 1 from XRD). The scattering, Guinier, Kratky, and  $P(r)$  distributions are shown in Supplementary Figures S8–S21. For  $R_g$  determination, each PDB structure was stripped of ions, waters, and cosolutes, and minimized with  $K^+$  between G-tetrads prior to CRYSOLE calculation (102). In CRYSOLE, obtaining an  $R_g$  from a model first requires calculation of the theoretical scattering,  $I(q)$ , after which the theoretical  $R_g$  can be derived from the slope of the net intensity. If the experimental scattering,  $I_{exp}(q)$ , is supplied, CRYSOLE will fit the theoretical scattering from model to the experimental by adjusting two free parameters that account for the average displaced volume per atomic group (or the effective atomic radius),  $r_o$ , and the contrast of the hydration layer,  $\delta_\rho$ . Without adjusting for hydration, systematic errors can arise leading to large discrepancies between resulting  $R_g$  values. The theoretical scattering is then calculated as a function of both parameters,  $I(q, r_o, \delta_\rho)$ , with initial values of  $r_o = 0.162$  nm and  $\delta_\rho = 30 e^-/nm^3$ . Next, a grid search is performed to identify optimal values of  $r_o$  and  $\delta_\rho$  that minimize the error-weighted  $\chi^2$  discrepancy function in Equation (1). The grid search spans the ranges  $r_o = 0.156 - 0.168$  nm and  $\delta_\rho = 0 - 70 e^-/nm^3$ , as these values encompass what has been experimentally observed (102). The resulting  $R_g$  and  $\chi^2$  values are plotted in Figure 5 and tabulated in Supplementary Tables S3, S4 (see Supplementary Figure S22 for fits). There is a very high correlation between experimental and calculated radii of gyration ( $r^2 = 0.9855$ ), with slope only slightly under 1. Importantly, the  $\chi^2$  values of the known structures ranged from 1.17–3.22 with average of  $2.1 \pm 0.7$ , indicating that each model fit their respective scattering profile well. This range of  $\chi^2$  values is comparable to CRYSOLE benchmarking studies of proteins (although with differences in the number of data points used in the calculations) (111). In-

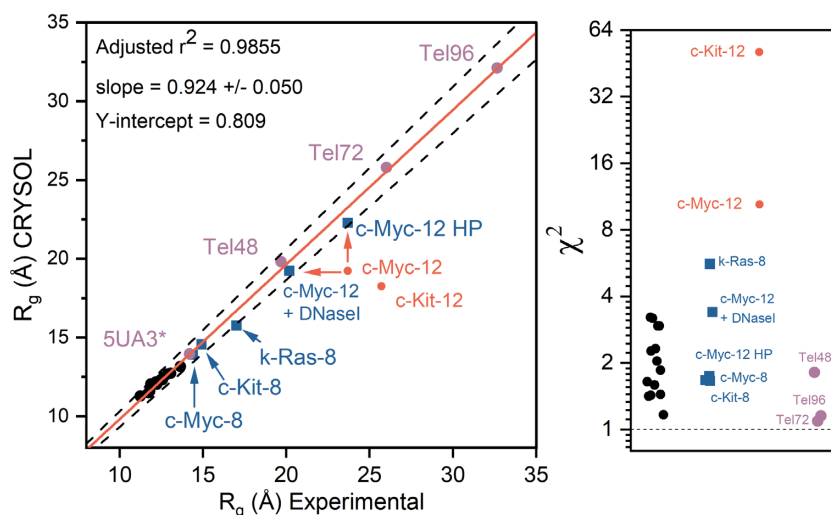


**Figure 4.** SEC-SAXS analysis of 8- and 12-tract promoter G-quadruplex sequences. (A, D) Log-linear plots of buffer subtracted scattering of 8-run (A), and 12-run (D) sequences. (B, E) Pair distance distribution functions,  $P(r)$ , normalized to scattering intensity at zero angle ( $I(0)$ ) for the 8-run (B) and 12-run (E) sequences. Both plots have the same X-axis to emphasize differences in  $D_{\max}$ , the maximum interparticle distances. (C, F) Dimensionless Kratky plots of 8-run (C) and 12-run (F) sequences with grey dashed and dotted lines overlaid to illustrate where on the X- and Y-axis a peak is expected for a globular particle. Included in panels A-C are the SAXS results of the 4-tract parallel c-Myc G-quadruplex 1XAV (in green) to contrast with the higher-order promoter sequences. Tabulated results are given in Supplementary Tables S3 and S4.

cluded in this regression are the higher-order telomere models derived previously (81) and PDB ID 5AU3 (112), the only other available G4s with SAXS data available, for a total of 18 data points. For the 14 PDB G4 structures, the average and standard deviation of  $r_o$  and  $\delta_\rho$  are  $1.64 \pm 0.19 \text{ \AA}$  and  $60 \pm 2 e^-/\text{nm}^3$ , respectively. The former value is consistent with the effective atomic radii reported for proteins and RNA (102,113) and the latter value is within the range observed for protein and RNA systems (with higher values associated with highly charged molecules) (102,113,114). We note here that the hydration density parameter could also be influenced by counterion condensation (114,115); however, this is outside the scope of our present study. The excellent agreement between model and measured  $R_g$  indicates that the radius of gyration is useful as another metric to filter ill-fitting models. However, as we will show, a more rigorous criterion for model inclusion is the combination of  $R_g$  agreement and low  $\chi^2$  value.

#### SAXS-MD model construction and refinement reveal that the 8-tract extended promoters form 4- or 5-quartet stacked parallel G-quadruplexes

The 8-tract structural models were constructed in the following way. First, the number of contiguous G-tetrad stacks formed by each sequence was constrained by values from CD analysis shown in Table 2 (c-Kit-8 and k-Ras-8 with 4 tetrads, c-Myc-8 with 5 tetrads). Next, G-tetrad columns were constructed using the known core G-tetrad stack structure of 1XAV. Each loop was then built into place using known G-quadruplex loop backbone orientations where possible. Adjustments to loop placement and geometry were made iteratively by checking the calculated sedimentation coefficient for the model against those measured by AUC. Models were then optimized and equilibrated as described in the methods section. Each system was then subjected to unrestrained, explicit solvent MD simulations



**Figure 5.** Correlation of experimental and *CRYSOLO* calculated radii of gyration for known G4 structures (left) and corresponding  $\chi^2$  values for each model (right). Black circles represent G4 structures that have been previously solved by either NMR (13/14) or X-ray (1/14) crystallography (tabulated values given in Supplementary Table S4 and *CRYSOLO* fits in Fig. S22). Blue squares represent the G4 models derived from this study that are consistent with the collective experimental data. Red circles indicate parallel stacked models from this study that deviate from one or more biophysical measurements (initially). The red arrows indicate changes in correlation of c-Myc-12 either with DNaseI treatment or by modeling in competing hairpins. Purple circles are values obtained from prior G4 SAXS studies (81,112). In the case of the telomere sequences (Tel48, 72 and 96) the regression values were obtained from EOM analyses (118). The dashed black lines indicate a  $\pm 5\%$  of the line of best fit.

for at least 100 ns. The calculated and experimental hydrodynamic parameters are given in Table 3 for each model. The calculated sedimentation coefficient ranges of the models constructed encompass the experimental values or are at most 0.1 S outside of the range.

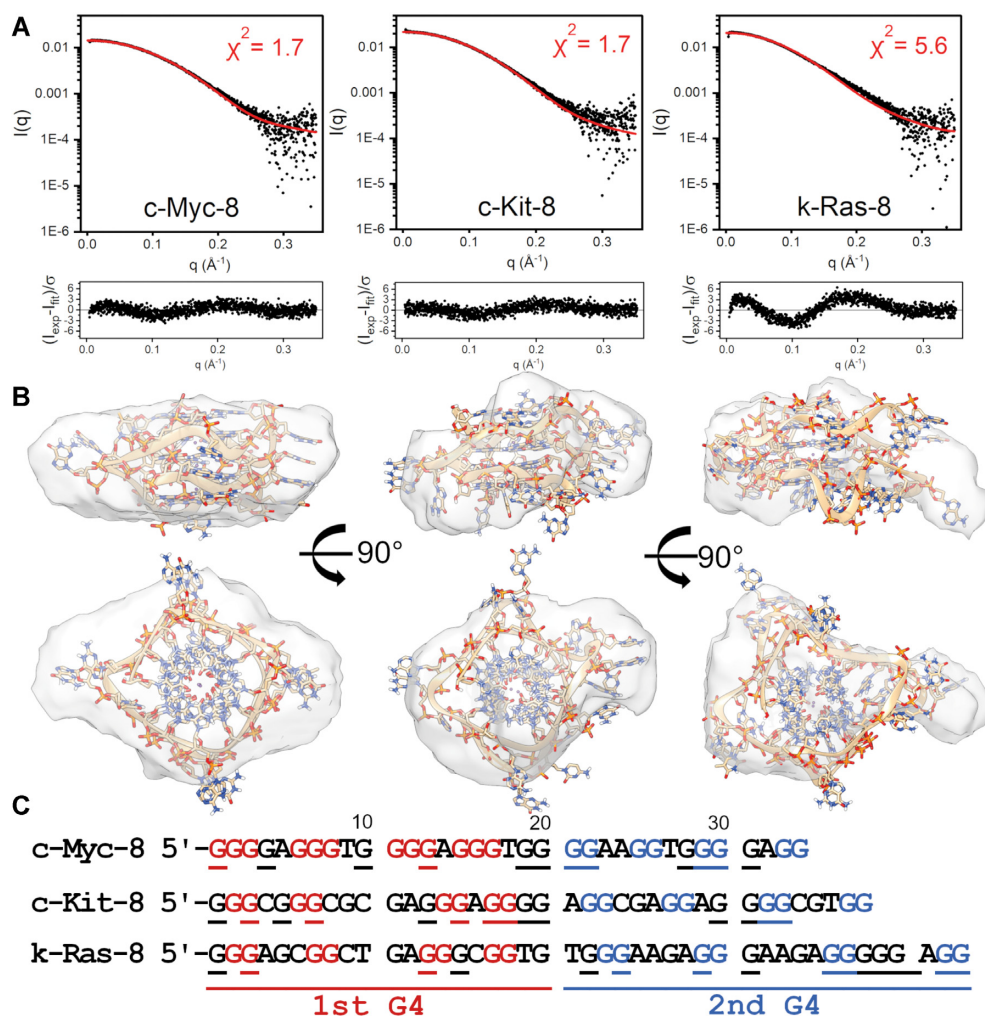
We next used *CRYSOLO* to compute the  $\chi^2$  fit values of 1000 equally spaced PDB frames from across each refined model's trajectory using their experimental scattering curve from Figure 4A. The  $R_g$  and  $\chi^2$  values of the best-fit models are shown in Figure 5. As shown in Figure 6, the atomistic models fit well into their *DAMMIF/N-refined* SAXS envelopes. Ambiguity assessments by *AMBIMETER* (116) gave ambiguity scores of 1.00, 1.08 and 1.53 for c-Myc-8, c-Kit-8 and k-Ras-8, reconstructions, respectively, indicating that the envelopes have little shape ambiguity (Supplementary Table S3). Further, each reconstruction agreed well with their measured  $R_g$  and  $D_{max}$ , had good agreement with known MWs, and normalized spatial discrepancy values (NSDs) of less than 1.0 (Supplementary Table S3). As a check of reconstruction validity, each *ab initio* model was used as input for *HYDROPRO* calculations to see if the shape is consistent with hydrodynamic measurements. Table 3 shows that c-Myc-8's *ab initio* model was within 4% of its measured  $S_{20,w}$ , whereas c-Kit-8 and k-Ras-8 reconstructions are 9% off from their measured value (although c-Kit-8 is well within the  $S_{20,w}$  range estimated from MD simulations of the all-atom model).

The atomistic models fit their scattering with  $\chi^2$  values of 1.7, 1.7 and 5.6 for cMyc-8, cKit-8 and kRas-8, respectively, as determined by *CRYSOLO*. The  $\chi^2$  fit value can be influenced by the noise in the scattering data (see Equation 1). However, since we have data on 14 structural models of G4s under identical buffer conditions, and the scattering in each case has reasonable S/N, we could compare our  $\chi^2$  fit values to those obtained for known structures. The average

and standard deviation of our G4 library  $\chi^2$ s is  $2.1 \pm 0.7$ , indicating that c-Myc-8 and c-Kit-8 models fit their scattering data well relative to known atomic G4 models (see Figure 5). k-Ras-8 has an  $R_g$  that is in good agreement with its measured value, but the model had a poor fit to its scattering, possibly reflecting coexistence of conformational isomers, which will be discussed below. This emphasizes  $R_g$  agreement alone is not sufficient for model validation. The various models built, simulated, and filtered out for the c-Myc-8 and c-Kit-8 G4s had  $\chi^2$  values on the order of 1.8–2.0 (G-register and loop isomers), 2.4–10.0 (fewer G-tetrad stacks) and  $\gg 10.0$  (all others), supporting that SAXS can discriminate G-register or loop isomers, and can easily filter out models with inconsistent topologies.

#### Inclusion of secondary structure prediction to refine models from SAXS data: Fold prediction and DNaseI cleavage analyses reveal that the 12-tract promoter sequences have competition between parallel G-quadruplex and hairpin features

Using the same modeling approach as outlined for the 8-tract sequences, *in silico* all-parallel stacked models were generated for c-Myc-12 and c-Kit-12. Although both models were in near agreement with their sedimentation values (Table 3), both deviated significantly from their experimental radii of gyration (red points in Figure 5). Based on the  $P(r)$  and Kratky analyses in Figure 4, we reasoned that unanticipated competing secondary structures, such as large loop regions, unstructured single-strands, or duplexes (69), could be responsible for the much larger than anticipated  $R_g$ s. To test this possibility, we submitted each sequence, along with the 8-tract sequences, to the Mfold (117) server with 0.2 M monovalent cation concentration and the rest of settings at default values (Supplemental Figures S23–S27). For the three 8-tract sequences, there is minimal



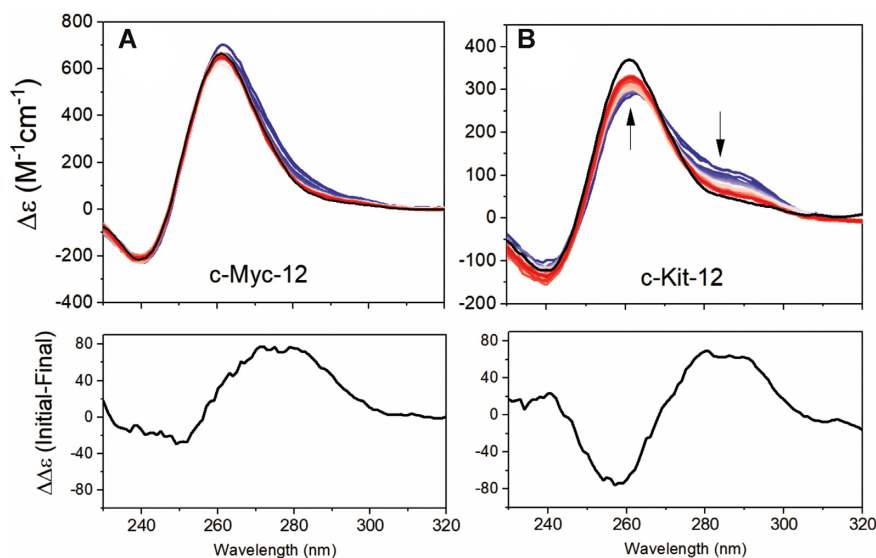
**Figure 6.** SAXS scattering profiles and CRY SOL fits for molecular models of c-Myc-8, c-Kit-8, and K-Ras-8 promoter sequences. (A) For each sequence the scattering is represented on a Log-linear scale in black, with the calculated profile fit from CRY SOL shown in red. Inset is each model's  $\chi^2$  value. See methods for  $\chi^2$  equation and parameter fitting. (B) Below each figure are the best fitting atomic models superimposed in their respective *DAMMIF/N* refined envelopes. (C) Sequences with red and blue coloration to highlight the guanines incorporated in each G4 of each model shown in (B), with underlined residues being those with the potential to be utilized by G-register isomers.

likelihood of competing thermodynamically stable duplex features ( $\Delta G$ s from +3.7 to  $-1.9$  kcal/mol). In contrast, c-Myc-12 and c-Kit-12 are predicted to have stable competing duplex features ( $\Delta G$ s from  $-5.1$  and  $-8.1$  kcal/mol, respectively) that might need to be included in structural models.

DNaseI, which does not cleave parallel G4s, can be used as a probe to test for duplex or hairpin features (52) within c-Myc-12 and c-Kit-12. The results of CD monitored DNaseI cleavage assays are shown in Figure 7. The spectral changes observed for the c-Myc-12 digestion were subtle (Figure 7A), but the CD difference spectrum is qualitatively consistent with digestion of non-G4 elements, as evident by the spectral magnitude, maximum at  $\sim 280$  nm and minimum at 250 nm. The c-Kit-12 digestion showed more pronounced changes in CD (Figure 7B) with a substantial 264 nm increase and reduction at 285 nm, consistent with a conversion to an all-parallel G4. The resulting CD difference spectrum, with  $\sim 280$  nm maximum and  $\sim 255$  nm minimum,

is consistent with B-form hairpin DNA in both shape and magnitude (84).

From the Mfold analyses, both c-Kit-12 and c-Myc-12 were predicted to have hairpin loops that could compete with quadruplex formation. To investigate the extent to which duplex features contributed to each sequence, we analyzed their post-digest products by AUC-SV (Table 3). Surprisingly, c-Myc-12 had only a subtle reduction in molecular weight ( $\sim 2$  kDa), corresponding to just a few nucleotides, yet had an increase in  $S_{20,w}$  with decreased  $f/f_0$ . Combined with the CD digestion analysis, this indicates that the remaining particle is a compact parallel G-quadruplex. The presence of only minor amounts of hairpin formation is consistent with c-Myc-12's proton NMR spectra (Supplemental Figure S28) which shows an 18:1 ratio of G4 imino protons to Watson-Crick imino protons. Conversely, c-Kit-12 had a digestion product that is hydrodynamically and spectroscopically equivalent to the c-Kit-8 G4 (compare in Table 3 and Figure 2). Since c-Kit-12 fully



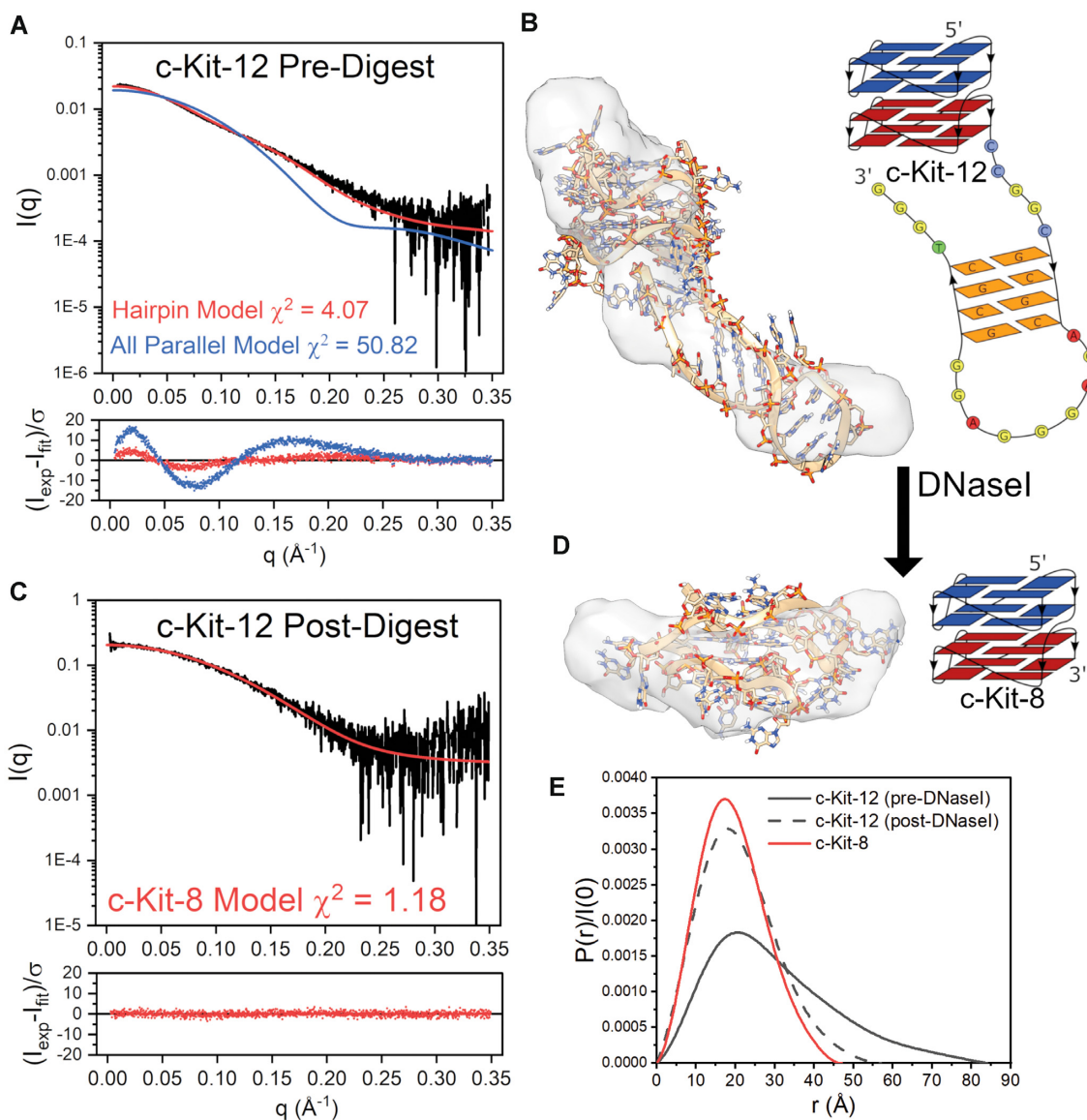
**Figure 7.** DNaseI digestions monitored by CD over 3 hours (blue to red) and overnight (black line) of c-Myc-12 (A) and c-Kit-12 (B) sequences. Shown below each digestion profile is the difference spectrum created by subtracting the final (overnight black line) from the initial spectra.

encompasses the c-Kit-8 sequence, this suggests that the remaining particle from DNaseI cleavage is the c-Kit-8 parallel four stack quadruplex described above (Figure 6).

To further investigate these possibilities, we collected SAXS data on the SEC-purified c-Kit-12 and c-Myc-12 DNaseI digestion products (Figures 8 and 9, respectively). We will first consider c-Kit-12. Figure 8 shows the scattering profiles of c-Kit-12 pre- and post-DNaseI treatment. In both cases the scattering data were determined to be homogeneous and free of interparticle artifacts, as shown by Guinier analysis of the low  $q$  region (Supplementary Figures S5 and S6). Figure 8A shows the scattering of c-Kit-12 pre-digest with overlaid *CRY SOL* fits of the stacked all-parallel model and a four stacked model with a hairpin incorporated as identified by Mfold. The hairpin model, although not in perfect agreement, is a much better representation of the scattering than the condensed globular, all-parallel model (not shown). This point is emphasized by the highly prolate space-filling envelope (Figure 8B). The  $\chi^2$ -value of 4.07 and poor envelope fit indicate that alternative hairpin-G4 isomers may also be contributing to the scattering, as this model is only one of multiple hairpin conformers identified as possible by Mfold. Further, the *ab initio* reconstruction is ambiguous based on AMBIMETER ambiguity score and the number of compatible shape categories (2.301 and 200, respectively) (Supplementary Table S3) and does not reflect the experimental sedimentation coefficient (Table 3). The envelope is shown to emphasize that the scattering is influenced by a highly extended species. More extensive flexible modeling approaches (e.g. EOM (118)) would likely be necessary to arrive at a better solution for the c-Kit-12 hairpin model. Figure 8C shows the post-DNaseI digestion scattering profile with the fit calculated from the four-tetrad c-Kit-8 model (same as in Figure 6). The fit is excellent as determined by its  $\chi^2$ -value of 1.2 and its normally distributed residuals. Visually, there is an imperfect fit with its *DAMMIF/N-refined* envelope (Figure 8D), and

although the reconstruction is also potentially ambiguous (Supplementary Table S3), it does reflect well the measured  $R_g$ ,  $D_{max}$ , and  $S_{20,w}$  (Table 3). c-Kit-12's change in size and shape from an extended prolate particle of  $D_{max} = 86 \text{ \AA}$  and  $R_g = 25.1 \text{ \AA}$ , to a very compact globular particle of  $D_{max} = 54 \text{ \AA}$  and  $R_g = 16.6 \text{ \AA}$  with similar dimensions ( $D_{max} = 48 \text{ \AA}$  and  $R_g = 15.0 \text{ \AA}$ ) and sedimentation coefficient to c-Kit-8 ( $S_{20,w,c-Kit-8} = 2.70$  vs.  $S_{20,w,DNaseI} = 2.85$ ) is reflected in the pair distance distribution functions in Figure 8E.

The c-Myc-12 DNaseI cleavage product is not as easily interpreted as c-Kit-12's (Figure 9). Figure 9A shows the pre-digest scattering of c-Myc-12 with overlaid *CRY SOL* fits of an all-parallel model and a model with a 3 base pair hairpin incorporated at the 5' end. The hairpin model fits the scattering data well ( $\chi^2 = 1.74$ , Figure 9A), visually fits the *DAMMIF/N* envelope (Figure 9B), and has an  $R_g$  that is in good agreement with what was originally measured ( $R_{g,exp} = 23.7 \text{ \AA}$  vs.  $R_{g,HP} = 22.3$  in Figure 5). However, based on our CD regression analysis, the incorporation of a B-form hairpin is expected to reduce the 260 nm CD signature by  $\sim 200 \Delta\epsilon \text{ M}^{-1} \text{ cm}^{-1}$ , as it disrupts the three tetrad G4 at the 5' end. Coincidentally, this hairpin sequence is composed of contiguous GC base pairs that may exhibit a CD signature consistent with A-form DNA (peak at  $\sim 264 \text{ nm}$ , trough at  $\sim 240 \text{ nm}$ ) (84) and resemble a parallel G4 CD signature, making it difficult to quantify its contribution (other than by using the NMR spectra in Supplementary Figure S28). However, the post-DNaseI SEC purified particle is much more globular in shape and has improved agreement with the all-parallel stacked model (Figure 9C and D). Taken with the increase in  $S_{20,w}$  and lower  $f/f_0$ , this suggests that the major form is compact and all parallel, with a minor extended hairpin species. We note that although the parallel stacked model fits well *visually* within its *ab initio* model, the reconstruction itself is potentially ambiguous and, further, both c-Myc-12 reconstructions had calculated  $S_{20,w}$  values

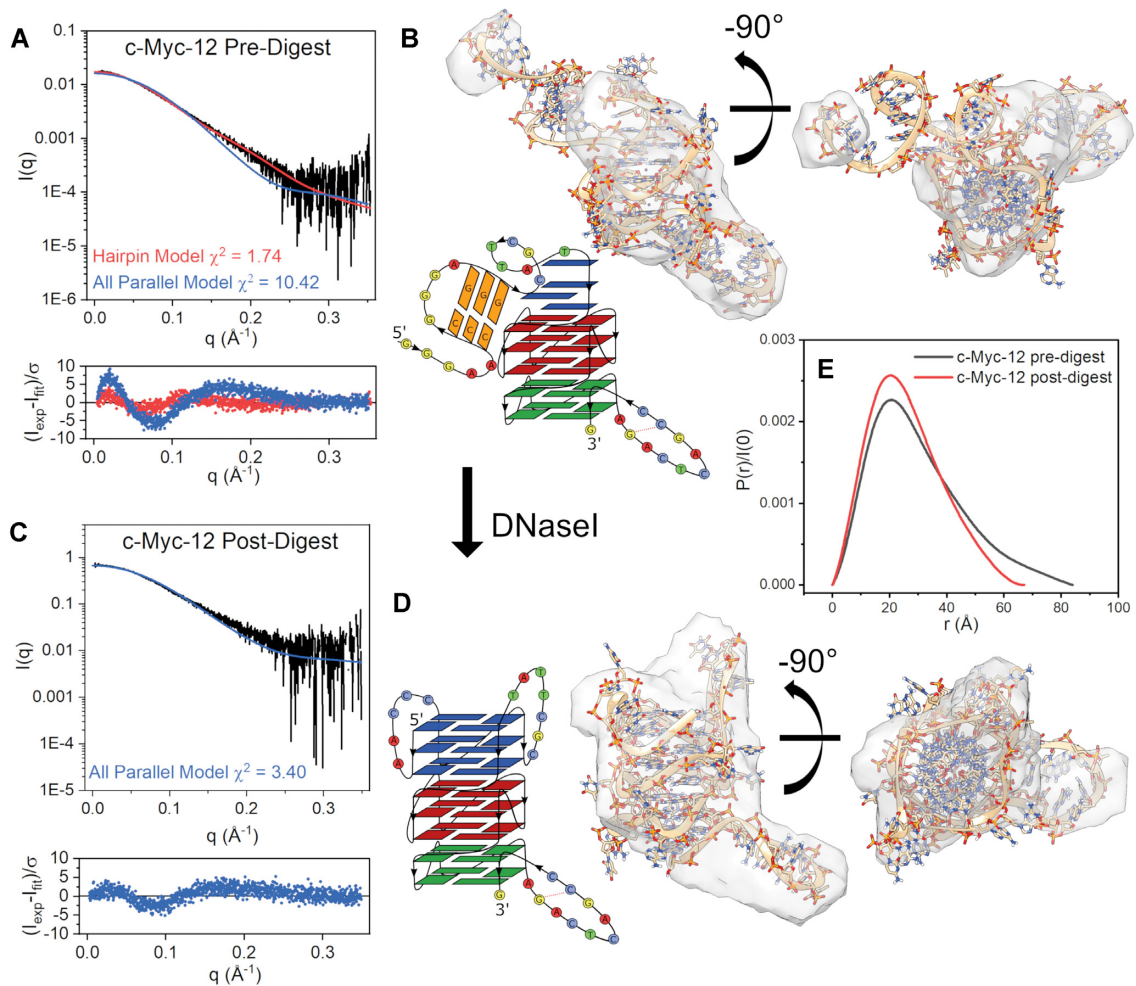


**Figure 8.** DNaseI-SEC-SAXS analysis of c-Kit-12 sequence. (A) SAXS scattering (Log-linear scale) of non-digested c-Kit-12 with overlaid fits of the all-parallel stacked or hairpin models and their weighted residuals. (B) Best-fitting hairpin model fit into *DAMMIF/N* refined scattering envelope along with its schematic representation. (C) SAXS scattering (Log-linear) of DNaseI-digested c-Kit-12 with overlaid fit of the c-Kit-8 four tetrad model (from Fig. 6) and its weighted residuals. (D) c-Kit-8 four stack model fit into the c-Kit-12 post-digestion *DAMMIF/N* refined scattering envelope along with its schematic representation. (E) Pair-distance distribution functions of c-Kit-12 pre-digestion (gray solid), c-Kit-12 post-digestion (gray dashed), and c-Kit-8 (red).

much lower than what was measured (Table 3). Also, because we don't know with certainty which loop or hairpin nucleotides are cleaved by DNaseI, the *CRY SOL* calculation was performed using the intact parallel model, which could be one reason for the non-uniformly distributed residuals. Figure 9E shows the pair distance distribution function of pre- and post-digested c-Myc-12. The decrease in  $D_{max}$  of and shift in  $R_g$  from 23.7 down to 20.2 Å yields an improvement in c-Myc-12's all-parallel model regression fit, (Figure 5 red arrow to blue). These collective results are consistent with a situation where coexistence of competing secondary structures can lead to convolution of CD, AUC, NMR, and scattering data, and so require an iterative integrative approach to parse out complex structural details.

## DISCUSSION

Our results show that an integrative structural biology approach can be used to effectively model the solution structures of higher-order G-quadruplexes, with a formal resolution of about 18 Å (107), but with derived atomistic models with greater detail. Structural models obtained for long *c-Myc*, *c-Kit* and *k-Ras* promoter sequences provide topological information not currently available from NMR or X-ray crystallography. While each of these promoter sequences fold into a unique structure, a common feature of these structures is the presence of contiguous parallel G4 units that interact to form unique interfacial pockets that might be targeted or recognized. Results from the longest

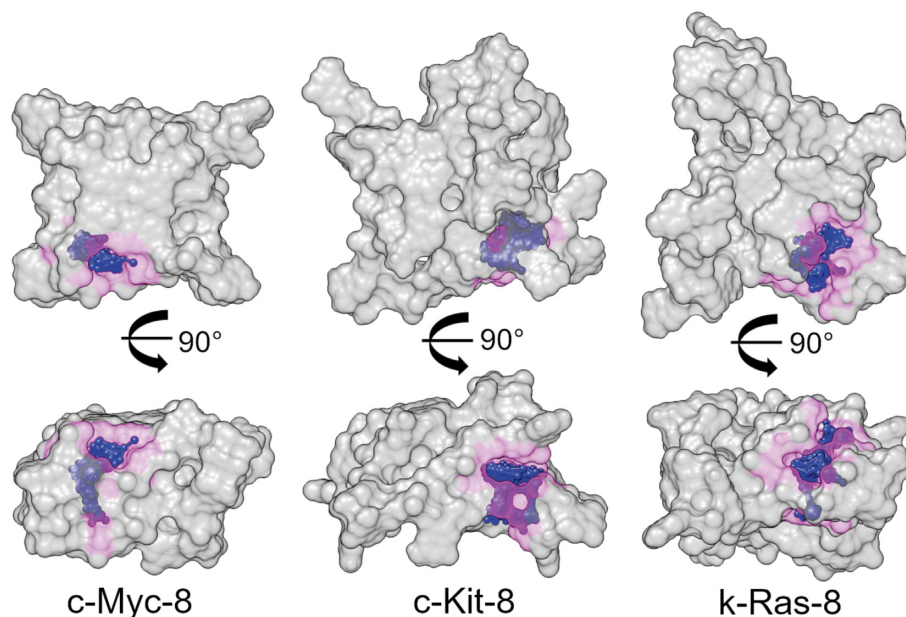


**Figure 9.** DNaseI-SEC-SAXS analysis of c-Myc-12 sequence. (A) SAXS scattering (Log-linear scale) of non-digested c-Myc-12 with overlaid fits of the all-parallel stacked or hairpin models and their weighted residuals. (B) Best-fitting hairpin model fit into the *DAMMIF/N* refined scattering envelope along with its schematic representation. (C) SAXS scattering (Log-linear) of DNaseI-digested c-Myc-12 sequence with overlaid fit of the all-parallel model (fully intact) and its weighted residuals. (D) c-Myc-12 all-parallel eight stack model fit into the post-digest *DAMMIF/N* refined scattering envelope and its schematic representation. (E) Pair-distance distribution functions of pre-digestion (gray) and post-digestion (red) particles.

*c-Myc* and *c-Kit* promoter sequences studied indicate that in addition to the G4 units, specific hairpin structures may also decorate the folded sequences to provide an even richer molecular terrain. These higher-ordered G4 structures all provide a rich array of potential drug binding sites.

Figure 10 shows surface renderings of the structures of c-Myc-8, c-Kit-8 and k-Ras-8. In this view, numerous unique topological features are evident, and these molecules at first glance look more like typical protein surfaces than canonical DNA. These folded G4 tertiary structures might be considered as DNA acting like a protein with respect to topological diversity. In order to illustrate the unique features of these higher-order G4 structures we used the program SiteMap (119) to generate information on the character and diversity of potential binding sites in the G4 structures we have determined (purple sites in Figure 10). SiteMap determines a ‘druggability score’ of a region of a protein or nucleic acid thereby providing a way to analyze potential binding sites and to predict target druggability. The calculated score characterizes a potential binding site with re-

spect to: (i) the size of the site; (ii) the degrees of enclosure by the receptor and exposure to solvent; (iii) the tightness with which the site points interact with the receptor; (iv) the hydrophobic and hydrophilic character of the site and the balance between them and (v) the degree to which a ligand might donate or accept hydrogen bonds. For reference, the average scores for undruggable, difficult, and druggable sites are 0.63, 0.87 and 1.1, respectively (119). For the top binding sites in k-Ras-8, c-Myc-8, and c-Kit-8 the SiteMap druggability scores were 0.96, 0.87 and 1.00, respectively. For comparison, the binding sites of the c-Myc-derived monomeric, parallel, three tetrad-quadruplex (1XAV.pdb) four low affinity binding sites were found, with druggability scores ranging from 0.55 to 0.69. For the c-Myc-12 8-stack, c-Myc-12 hairpin, and c-Kit-12 hairpin structures, the top druggability scores were 0.94, 0.95, and 0.95, respectively, with multiple binding sites predicted (not shown). For further comparison, the known B-form minor groove binding site (289D.pdb), mithromycin A-form minor groove binding site (146D.pdb), and the dauno-



**Figure 10.** Space-filling representations of the c-Myc-8, c-Kit-8, and k-Ras-8 higher-order G-quadruplex models. The top row is looking down the central G4 stem and the bottom row is a side view. Magenta highlights a zone 4 Å from the ‘SITE’ ball and stick model (in blue) to emphasize the size of the predicted top-scoring binding sites.

mycin intercalation binding site (1d11.pdb) gave druggability scores of 0.99, 1.01, 0.85, respectively. Overall, this indicates that the higher-order 8- and 12-track structures have better and more druggable binding sites than simpler monomeric G4s. Figure 10 shows the top ranked (of several) sites in c-Myc-8, c-Kit-8 and k-Ras-8. These sites are unique to each sequence and structure, suggesting the possibility of targeting specific G4 promoters. However, targeting these sites experimentally is an effort beyond the scope of this work, but has begun and is an ongoing focus in our laboratory using the drug discovery platform we have described previously (120,121).

The coexistence of G-quadruplex isomers and competing secondary structures is often the rule of long G-rich sequences, and not the exception (20,22,122). The ISB approach (75), as demonstrated here, can satisfactorily characterize such complexities. Models of the 8-track promoter sequences c-Myc-8 and c-Kit-8 (Figure 6) provide an excellent fit to all available biophysical measurements, although we caution that these models are only one solution to what could comprise an ensemble of G-register isomers. The k-Ras-8 model exemplifies this point (Figure 6, right panel). The parallel 4-stack model is self-consistent with spectroscopic and hydrodynamic measurements, yet its poor fit to its scattering data (e.g. skewing in its P(r) and Kratky profiles and poor  $\chi^2$  fit) indicates that a population of structurally distinct isomers likely co-exist. Marquieville et al. (123) recently reported on a 32-nt truncated version of the k-Ras-8 sequence. The authors show that their ‘K-RAS32R’ exists in a dynamic ensemble of two parallel G-register isomers that interconvert on a ms-timescale (123). This highlights the discriminatory power of combining hydrodynamics and spectroscopy with SAXS modeling.

The structural complexity of c-Myc-12 and c-Kit-12 are excellent examples of how powerful the ISB approach can be. In the former case, CD 264 nm values (Figure 2) indicate that c-Myc-12 is a fully stacked 8-tetrad parallel structure. However, its calculated radius of gyration was much smaller than measured (Figure 5). Integration of information from the Mfold DNA fold prediction server, physical theory, NMR (Supplemental Figure S28), and a DNaseI cleavage assay revealed that competing hairpin motifs (22) may be giving rise to a more extended prolate structure, as evidenced by the slightly higher frictional ratio, high  $R_g$ , and a prolate space-filling envelope. Based on the low ratio of Watson-Crick imino peaks relative to the Hoogsteen peaks by proton NMR (1:18), the much lower than expected  $S_{20,w}$  calculated for the extended *ab initio* models, and the nearly unchanged 264 nm CD magnitude after DNaseI digestion, it is reasonable to conclude that the major topology of c-Myc-12 is that of an all-parallel 8-tetrad system. Indeed, the all-parallel stacked atomistic model agrees well with the AUC sedimentation and SAXS scattering of the post-digestion sample (Figure 9 and Table 3). Conversely, c-Kit-12 was predicted to form extensive GC duplex features by Mfold, which was confirmed by DNaseI cleavage (Supplemental Figure S27). Hydrodynamic and scattering investigations of the post-digestion product reveals that c-Kit-12 likely forms the same parallel G4 as its truncated counterpart, c-Kit-8, with an extended hairpin loop as its major form, rather than a contiguous parallel stacked system. The biological importance of such a structure remains to be determined.

The sequence context in which we study extended sequences is also very important. We initially observed that promoter G4s preferentially adopt a parallel conformation, based on deposited structures and our prior investigations



(52). Chen *et al.* recently reported that the addition of 5'-flanking non-guanine nucleotides induced conformational shifts from antiparallel or hybrid to parallel in ~80% of the >300 sequences tested (70). We note that the sequences studied here do not have non-guanine 5' flanking residues, but we subsequently tested the effects of adding 5' nucleotides and found no substantial differences by CD (see Supplementary Figure S29). Recent studies also revealed that G4 motifs that are adjacent to one another tend to interact or stack rather than exist as a 'beads-on-a-string'. We (52) and others (51) have shown that the hTERT core promoter region forms a three parallel stacked assembly and, importantly, that the internal G4 region only forms in the presence of one or both of the outer G4s. Two independent studies of the c-Kit proximal promoter region (non-overlapping with the c-Kit-8 or -12 sequences studied here) have shown a similar phenomenon, although no atomistic models were proposed. In the first study Rigo and Sissi (53) used CD and melting studies to show that the kit2-kit\* higher-order quadruplex exhibits a thermodynamic and structural cross-talk between the two G4 subunits (53). A later study Ducani *et al.* (54) investigated a sequence containing all three of the previously reported monomeric G4s, kit2-kit\*-kit1, as well as mutated combinations thereof. They report that the kit\* sequence does not form an antiparallel G4 in the presence of kit2, but rather is stabilized as a parallel stacked higher-order complex (54). Importantly, two-tetrad parallel DNA G4s are unstable without external stabilizing forces, such as hairpin loops or stacking interactions (124), highlighting the importance of developing integrative approaches to tackle extended sequences.

As indicated in the introduction, there may be biological advantages in forming higher-order promoter G4s. Promoter G4s have long been suspected to exert regulatory functions on gene transcription based on genetic conservation, prevalence in nucleosome depleted regions, and their non-random distribution in gene promoters (4,125). Recently, state-of-the-art sequencing has revealed that promoter G4s act as transcription factor (TF) hubs, and that small molecules can effectively displace TF binding (9). However, biological studies of higher-order G4s are sparse. We have previously shown that the hTERT core promoter sequence adopts a higher-order three-stack parallel G4 (52). The Costello lab (126) has shown that the two frequent G > A hTERT core promoter mutations, both of which reside in the middle G4 G-tracts, lead to a TF profile switch that significantly increases promoter activity. Although they show that these specific mutations create new TF binding motifs, it is very difficult to parse out contributions from putative G4 secondary structure for this very reason. A more direct dissection of a higher-order G4 was recently conducted by the Terenzi lab (54) using the c-Kit promoter. By systematically mutating out each G4 motif with poly adenines, the authors show that promoter activity is directly modulated by the status of higher-order G4 formation. Moreover, their study reveals that the c-Kit G4s do not 'titrate' the promoter response or behave as simple switches, rather, they may instead 'code' for a particular promoter response. Collectively, these studies suggest that higher-order

G4 promoters offer an unrivaled 3-dimensional landscape that is critically linked to promoter activity.

## CONCLUSION

The promoter regions of oncogenes have long sequences with the potential to form multiple quadruplexes, yet this complexity has largely been intractable to modeling. Structural biologists have instead, for expediency, focused primarily on shorter sequences that fold into a single monomeric G4. The integrated structural biology approach provides the means for studying the structures of the more relevant extended sequences. Every such sequence studied to date shows that more complex higher-order G4 structures readily form, featuring contiguous G4 units, longer loop structures and in some cases coexisting duplex hairpins. We suggest that these more complex assemblies may be the more important regulators of promoter function in ways that are yet to be defined.

## DATA AVAILABILITY

Small-angle X-ray scattering data and models, where applicable, have been deposited in the publicly accessible Small Angle Scattering Biological Data Bank (<https://www.sasbdb.org/>) under the names (IDs): c-Myc-8 (SASDMJ6), c-Myc-12 (SASDMM6), c-Myc-12 post-DNaseI (SASDM36), c-Kit-8 (SASDMK6), c-Kit-12 (SASDMN6), c-Kit-12 post-DNaseI (SASDM46), k-Ras-8 (SASDML6), 2KZD (SASDM56), 201D (SASDM76), 6GH0 (SASDM86), 2GKU (SASDM96), 2JSL (SASDKF3), 5I2V (SASDMA6), 6L92 (SASDMB6), 2KQG (SASDMC6), 2LBY (SASDMD6), 6NEB (SASDME6), 1XAV (SASDMF6), 2KZE (SASDM66), 5CMX (SASDMG6), 2M27 (SASDMH6).

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

This research used resources of the Advanced Photon Source, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH11357. This project was supported by grant 9 P41 GM103622 from the National Institute of General Medical Sciences of the National Institutes of Health. Use of the Pilatus 3 1M detector was provided by grant 1S10OD018090-01 from NIGMS.

The content is solely the responsibility of the authors and does not necessarily reflect the official views of the National Institute of general Medical Sciences or the National Institutes of Health.

Molecular graphics and analyses performed with UCSF Chimera, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from NIH P41-GM103311.

## FUNDING

National Institutes of Health (NIH) [GM077422]. Funding for open access charge: UofL Health Brown Cancer Center. *Conflict of interest statement*. None declared.

## REFERENCES

- Todd, A.K., Johnston, M. and Neidle, S. (2005) Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res.*, **33**, 2901–2907.
- Huppert, J.L. and Balasubramanian, S. (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.*, **33**, 2908–2916.
- Lane, A.N., Chaires, J.B., Gray, R.D. and Trent, J.O. (2008) Stability and kinetics of G-quadruplex structures. *Nucleic Acids Res.*, **36**, 5482–5515.
- Huppert, J.L. and Balasubramanian, S. (2007) G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.*, **35**, 406–413.
- Balasubramanian, S., Hurley, L.H. and Neidle, S. (2011) Targeting G-quadruplexes in gene promoters: a novel anticancer strategy? *Nat. Rev. Drug Discov.*, **10**, 261.
- Bartas, M., Brazda, V., Karlicky, V., Cerven, J. and Pecinka, P. (2018) Bioinformatics analyses and in vitro evidence for five and six stacked G-quadruplex forming sequences. *Biochimie*, **150**, 70–75.
- Chambers, V.S., Marsico, G., Boutell, J.M., Di Antonio, M., Smith, G.P. and Balasubramanian, S. (2015) High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat. Biotechnol.*, **33**, 877–881.
- Hansel-Hertsch, R., Spiegel, J., Marsico, G., Tannahill, D. and Balasubramanian, S. (2018) Genome-wide mapping of endogenous G-quadruplex DNA structures by chromatin immunoprecipitation and high-throughput sequencing. *Nat. Protoc.*, **13**, 551–564.
- Spiegel, J., Cuesta, S.M., Adhikari, S., Hansel-Hertsch, R., Tannahill, D. and Balasubramanian, S. (2021) G-quadruplexes are transcription factor binding hubs in human chromatin. *Genome Biol.*, **22**, 117.
- Zhang, X., Spiegel, J., Martinez Cuesta, S., Adhikari, S. and Balasubramanian, S. (2021) Chemical profiling of DNA G-quadruplex-interacting proteins in live cells. *Nat. Chem.*, **13**, 626–633.
- Siddiqui-Jain, A., Grand, C.L., Bearss, D.J. and Hurley, L.H. (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 11593–11598.
- Dexheimer, T.S., Sun, D. and Hurley, L.H. (2006) Deconvoluting the structural and drug-recognition complexity of the G-quadruplex-forming region upstream of the bcl-2 P1 promoter. *J. Am. Chem. Soc.*, **128**, 5404–5415.
- Guo, K., Gokhale, V., Hurley, L.H. and Sun, D. (2008) Intramolecularly folded G-quadruplex and i-motif structures in the proximal promoter of the vascular endothelial growth factor gene. *Nucleic Acids Res.*, **36**, 4598–4608.
- De Armond, R., Wood, S., Sun, D., Hurley, L.H. and Ebbinghaus, S.W. (2005) Evidence for the presence of a guanine quadruplex forming region within a polypurine tract of the hypoxia inducible factor 1 $\alpha$  promoter. *Biochemistry*, **44**, 16341–16350.
- Palumbo, S.L., Memmott, R.M., Uribe, D.J., Krotova-Khan, Y., Hurley, L.H. and Ebbinghaus, S.W. (2008) A novel G-quadruplex-forming GGA repeat region in the c-myc promoter is a critical regulator of promoter activity. *Nucleic Acids Res.*, **36**, 1755–1769.
- Qin, Y., Rezler, E.M., Gokhale, V., Sun, D. and Hurley, L.H. (2007) Characterization of the G-quadruplexes in the duplex nuclease hypersensitive element of the PDGF-A promoter and modulation of PDGF-A promoter activity by TMPyP4. *Nucleic Acids Res.*, **35**, 7698–7713.
- Mitchell, T., Ramos-Montoya, A., Di Antonio, M., Murat, P., Ohnmacht, S., Micco, M., Jurmeister, S., Fryer, L., Balasubramanian, S., Neidle, S. *et al.* (2013) Downregulation of androgen receptor transcription by promoter G-quadruplex stabilization as a potential alternative treatment for castrate-resistant prostate cancer. *Biochemistry*, **52**, 1429–1436.
- Tong, X., Lan, W., Zhang, X., Wu, H., Liu, M. and Cao, C. (2011) Solution structure of all parallel G-quadruplex formed by the oncogene RET promoter sequence. *Nucleic Acids Res.*, **39**, 6753–6763.
- Wei, D., Todd, A.K., Zloh, M., Gunaratnam, M., Parkinson, G.N. and Neidle, S. (2013) Crystal structure of a promoter sequence in the B-raf gene reveals an intertwined dimer quadruplex. *J. Am. Chem. Soc.*, **135**, 19319–19329.
- Greco, M.L., Kotar, A., Rigo, R., Cristofari, C., Plavec, J. and Sissi, C. (2017) Coexistence of two main folded G-quadruplexes within a single G-rich domain in the EGFR promoter. *Nucleic Acids Res.*, **45**, 10132–10142.
- Sengar, A., Vandana, J.J., Chambers, V.S., Di Antonio, M., Winnerdy, F.R., Balasubramanian, S. and Phan, A.T. (2019) Structure of a (3+1) hybrid G-quadruplex in the PARP1 promoter. *Nucleic Acids Res.*, **47**, 1564–1572.
- Kuo, M.H., Wang, Z.F., Tseng, T.Y., Li, M.H., Hsu, S.T., Lin, J.J. and Chang, T.C. (2015) Conformational transition of a hairpin structure to G-quadruplex within the WNT1 gene promoter. *J. Am. Chem. Soc.*, **137**, 210–218.
- Wang, J.M., Huang, F.C., Kuo, M.H., Wang, Z.F., Tseng, T.Y., Chang, L.C., Yen, S.J., Chang, T.C. and Lin, J.J. (2014) Inhibition of cancer cell migration and invasion through suppressing the Wnt1-mediating signal pathway by G-quadruplex structure stabilizers. *J. Biol. Chem.*, **289**, 14612–14623.
- Shklover, J., Weisman-Shomer, P., Yafe, A. and Fry, M. (2010) Quadruplex structures of muscle gene promoter sequences enhance in vivo myod-dependent gene expression. *Nucleic Acids Res.*, **38**, 2369–2377.
- Yafe, A., Etzioni, S., Weisman-Shomer, P. and Fry, M. (2005) Formation and properties of hairpin and tetraplex structures of guanine-rich regulatory sequences of muscle-specific genes. *Nucleic Acids Res.*, **33**, 2887–2900.
- Zhu, J., Fleming, A.M. and Burrows, C.J. (2018) The RAD17 promoter sequence contains a potential tail-dependent G-Quadruplex that downregulates gene expression upon oxidative modification. *ACS Chem. Biol.*, **13**, 2577–2584.
- Huang, M.C., Chu, I.T., Wang, Z.F., Lin, S., Chang, T.C. and Chen, C.T. (2018) A G-Quadruplex structure in the promoter region of CLIC4 functions as a regulatory element for gene expression. *Int. J. Mol. Sci.*, **19**, 2678.
- Redstone, S.C.J., Fleming, A.M. and Burrows, C.J. (2019) Oxidative modification of the potential G-Quadruplex sequence in the PCNA gene promoter can turn on transcription. *Chem. Res. Toxicol.*, **32**, 437–446.
- Huang, W., Smaldino, P.J., Zhang, Q., Miller, L.D., Cao, P., Stadelman, K., Wan, M., Giri, B., Lei, M., Nagamine, Y. *et al.* (2012) Yin yang 1 contains G-quadruplex structures in its promoter and 5'-UTR and its expression is modulated by G4 resolvase 1. *Nucleic Acids Res.*, **40**, 1033–1049.
- Ohnmacht, S.A., Micco, M., Petrucci, V., Todd, A.K., Reszka, A.P., Gunaratnam, M., Carvalho, M.A., Zloh, M. and Neidle, S. (2012) Sequences in the HSP90 promoter form G-quadruplex structures with selectivity for disubstituted phenyl bis-oxazole derivatives. *Bioorg. Med. Chem. Lett.*, **22**, 5930–5935.
- Wei, P.C., Wang, Z.F., Lo, W.T., Su, M.I., Shew, J.Y., Chang, T.C. and Lee, W.H. (2013) A cis-element with mixed G-quadruplex structure of NPGPx promoter is essential for nucleolin-mediated transactivation on non-targeting siRNA stress. *Nucleic Acids Res.*, **41**, 1533–1543.
- Stevens, A.J. and Kennedy, M.A. (2017) Structural analysis of G-quadruplex formation at the human MEST promoter. *PLoS One*, **12**, e0169433.
- Jana, S., Jana, J., Patra, K., Mondal, S., Bhat, J., Sarkar, A., Sengupta, P., Biswas, A., Mukherjee, M., Tripathi, S.P. *et al.* (2017) LINC RNA00273 promotes cancer metastasis and its G-Quadruplex promoter can serve as a novel target to inhibit cancer invasiveness. *Oncotarget*, **8**, 110234–110256.
- Purohit, G., Mukherjee, A.K., Sharma, S. and Chowdhury, S. (2018) Extratelomeric binding of the telomere binding protein TRF2 at the PCGF3 promoter is G-Quadruplex motif-dependent. *Biochemistry*, **57**, 2317–2324.

35. Farhath, M.M., Thompson, M., Ray, S., Sewell, A., Balci, H. and Basu, S. (2015) G-Quadruplex-enabling sequence within the human tyrosine hydroxylase promoter differentially regulates transcription. *Biochemistry*, **54**, 5533–5545.
36. Schlag, K., Steinhilber, D., Karas, M. and Sorg, B.L. (2020) Analysis of proximal ALOX5 promoter binding proteins by quantitative proteomics. *FEBS J.*, **287**, 4481–4499.
37. Salvati, E., Zizza, P., Rizzo, A., Iachettini, S., Cingolani, C., D'Angelo, C., Porru, M., Randazzo, A., Pagano, B., Novellino, E. et al. (2014) Evidence for G-quadruplex in the promoter of vegfr-2 and its targeting to inhibit tumor angiogenesis. *Nucleic Acids Res.*, **42**, 2945–2957.
38. Basundra, R., Kumar, A., Amrane, S., Verma, A., Phan, A.T. and Chowdhury, S. (2010) A novel G-quadruplex motif modulates promoter activity of human thymidine kinase 1. *FEBS J.*, **277**, 4254–4264.
39. Waller, Z.A., Howell, L.A., Macdonald, C.J., O'Connell, M.A. and Searcey, M. (2014) Identification and characterisation of a G-quadruplex forming sequence in the promoter region of nuclear factor (erythroid-derived 2)-like 2 (Nrf2). *Biochem. Biophys. Res. Commun.*, **447**, 128–132.
40. Yan, J., Zhao, D., Dong, L., Pan, S., Hao, F. and Guan, Y. (2017) A novel G-quadruplex motif in the human MET promoter region. *Biosci. Rep.*, **37**, BSR20171128.
41. Yang, D. and Hurley, L.H. (2006) Structure of the biologically relevant G-quadruplex in the c-MYC promoter. *Nucleosides. Nucleotides. Nucleic Acids.*, **25**, 951–968.
42. Fernando, H., Reszka, A.P., Huppert, J., Ladame, S., Rankin, S., Venkitaraman, A.R., Neidle, S. and Balasubramanian, S. (2006) A conserved quadruplex motif located in a transcription activation site of the human c-kit oncogene. *Biochemistry*, **45**, 7854–7860.
43. Zhang, L., Tan, W., Zhou, J., Xu, M. and Yuan, G. (2017) Investigation of G-quadruplex formation in the FGFR2 promoter region and its transcriptional regulation by liensinine. *Biochim. Biophys. Acta Gen. Subj.*, **1861**, 884–891.
44. Ambrus, A., Chen, D., Dai, J., Jones, R.A. and Yang, D. (2005) Solution structure of the biologically relevant G-quadruplex element in the human c-MYC promoter. Implications for G-quadruplex stabilization. *Biochemistry*, **44**, 2048–2058.
45. Hsu, S.T., Varnai, P., Bugaut, A., Reszka, A.P., Neidle, S. and Balasubramanian, S. (2009) A G-rich sequence within the c-kit oncogene promoter forms a parallel G-quadruplex having asymmetric G-tetrad dynamics. *J. Am. Chem. Soc.*, **131**, 13399–13409.
46. Kerkour, A., Marquieville, J., Ivashchenko, S., Yatsunyk, L.A., Mergny, J.L. and Salgado, G.F. (2017) High-resolution three-dimensional NMR structure of the KRAS proto-oncogene promoter reveals key features of a G-quadruplex involved in transcriptional regulation. *J. Biol. Chem.*, **292**, 8082–8091.
47. Lim, K.W., Lacroix, L., Yue, D.J., Lim, J.K., Lim, J.M. and Phan, A.T. (2010) Coexistence of two distinct G-quadruplex conformations in the hTERT promoter. *J. Am. Chem. Soc.*, **132**, 12331–12342.
48. Neidle, S. (2020) In: Neidle, S. (ed). *Annual Reports in Medicinal Chemistry*. Academic Press, Vol. **54**, pp. 517–546.
49. Monsen, R.C. and Trent, J.O. (2018) G-quadruplex virtual drug screening: a review. *Biochimie*, **152**, 134–148.
50. Neidle, S. (2016) Quadruplex nucleic acids as novel therapeutic targets. *J. Med. Chem.*, **59**, 5987–6011.
51. Micheli, E., Martufi, M., Cacchione, S., De Santis, P. and Savino, M. (2010) Self-organization of G-quadruplex structures in the hTERT core promoter stabilized by polyaminic side chain perylene derivatives. *Biophys. Chem.*, **153**, 43–53.
52. Monsen, R.C., DeLeeuw, L., Dean, W.L., Gray, R.D., Sabo, T.M., Chakravarthy, S., Chaires, J.B. and Trent, J.O. (2020) The hTERT core promoter forms three parallel G-quadruplexes. *Nucleic Acids Res.*, **48**, 5720–5734.
53. Rigo, R. and Sissi, C. (2017) Characterization of G4-G4 crosstalk in the c-KIT promoter region. *Biochemistry*, **56**, 4309–4312.
54. Ducani, C., Bernardinelli, G., Hogberg, B., Keppler, B.K. and Terenzi, A. (2019) Interplay of three G-quadruplex units in the KIT promoter. *J. Am. Chem. Soc.*, **141**, 10205–10213.
55. Navarro, A., Benabou, S., Eritja, R. and Gargallo, R. (2020) Influence of pH and a porphyrin ligand on the stability of a G-quadruplex structure within a duplex segment near the promoter region of the SMARCA4 gene. *Int. J. Biol. Macromol.*, **159**, 383–393.
56. Do, N.Q., Lim, K.W., Teo, M.H., Heddi, B. and Phan, A.T. (2011) Stacking of G-quadruplexes: NMR structure of a G-rich oligonucleotide with potential anti-HIV and anticancer activity. *Nucleic Acids Res.*, **39**, 9448–9457.
57. Kogut, M., Kleist, C. and Czub, J. (2019) Why do G-quadruplexes dimerize through the 5'-ends? Driving forces for G4 DNA dimerization examined in atomic detail. *PLoS Comput. Biol.*, **15**, e1007383.
58. Kolesnikova, S. and Curtis, E.A. (2019) Structure and function of multimeric G-quadruplexes. *Molecules*, **24**, 3074.
59. Lago, S., Nadai, M., Cernilogar, F.M., Kazerani, M., Dominiguez Moreno, H., Schotta, G. and Richter, S.N. (2021) Promoter G-quadruplexes and transcription factors cooperate to shape the cell type-specific transcriptome. *Nat. Commun.*, **12**, 3885.
60. Lightfoot, H.L., Hagen, T., Tatum, N.J. and Hall, J. (2019) The diverse structural landscape of quadruplexes. *FEBS Lett.*, **593**, 2083–2102.
61. Harkness, R.W. and Mittermaier, A.K. (2016) G-register exchange dynamics in guanine quadruplexes. *Nucleic Acids Res.*, **44**, 3481–3494.
62. Fleming, A.M., Zhou, J., Wallace, S.S. and Burrows, C.J. (2015) A role for the fifth G-Track in G-Quadruplex forming oncogene promoter sequences during oxidative stress: do these "Spare tires" have an evolved function? *ACS Cent. Sci.*, **1**, 226–233.
63. Le, H.T., Miller, M.C., Buscaglia, R., Dean, W.L., Holt, P.A., Chaires, J.B. and Trent, J.O. (2012) Not all G-quadruplexes are created equally: an investigation of the structural polymorphism of the c-Myc G-quadruplex-forming sequence and its interaction with the porphyrin TMPyP4. *Org. Biomol. Chem.*, **10**, 9393–9404.
64. Le, H.T., Buscaglia, R., Dean, W.L., Chaires, J.B. and Trent, J.O. (2013) Calculation of hydrodynamic properties for G-quadruplex nucleic acid structures from in silico bead models. *Top. Curr. Chem.*, **330**, 179–210.
65. Chaires, J.B., Dean, W.L., Le, H.T. and Trent, J.O. (2015) Hydrodynamic models of G-Quadruplex structures. *Methods Enzymol.*, **562**, 287–304.
66. Do, N.Q. and Phan, A.T. (2012) Monomer-dimer equilibrium for the 5'-5' stacking of propeller-type parallel-stranded G-quadruplexes: NMR structural study. *Chemistry*, **18**, 14752–14759.
67. Onel, B., Carver, M., Wu, G., Timonina, D., Kalarn, S., Larriva, M. and Yang, D. (2016) A new G-quadruplex with hairpin loop immediately upstream of the human BCL2 P1 promoter modulates transcription. *J. Am. Chem. Soc.*, **138**, 2563–2570.
68. Ngoc Nguyen, T.Q., Lim, K.W. and Phan, A.T. (2020) Duplex formation in a G-quadruplex bulge. *Nucleic Acids Res.*, **48**, 10567–10575.
69. Ravichandran, S., Razzaq, M., Parveen, N., Ghosh, A. and Kim, K.K. (2021) The effect of hairpin loop on the structure and gene expression activity of the long-loop G-quadruplex. *Nucleic Acids Res.*, **49**, 10689–10706.
70. Chen, J., Cheng, M., Salgado, G.F., Stadlbauer, P., Zhang, X., Amrane, S., Guedin, A., He, F., Sponer, J., Ju, H. et al. (2021) The beginning and the end: flanking nucleotides induce a parallel G-quadruplex topology. *Nucleic Acids Res.*, **49**, 9548–9559.
71. 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.
72. Dickerhoff, J., Onel, B., Chen, L., Chen, Y. and Yang, D. (2019) Solution structure of a MYC promoter G-Quadruplex with 1:6:1 loop length. *ACS Omega*, **4**, 2533–2539.
73. Agrawal, P., Hatzakis, E., Guo, K., Carver, M. and Yang, D. (2013) Solution structure of the major G-quadruplex formed in the human VEGF promoter in K<sup>+</sup>: insights into loop interactions of the parallel G-quadruplexes. *Nucleic Acids Res.*, **41**, 10584–10592.
74. Liu, H.Y., Zhao, Q., Zhang, T.P., Wu, Y., Xiong, Y.X., Wang, S.K., Ge, Y.L., He, J.H., Lv, P., Ou, T.M. et al. (2016) Conformation selective antibody enables genome profiling and leads to discovery of parallel G-quadruplex in human telomeres. *Cell Chem Biol*, **23**, 1261–1270.
75. Rout, M.P. and Sali, A. (2019) Principles for integrative structural biology studies. *Cell*, **177**, 1384–1403.

76. Karsisiotis, A.I. and Webba da Silva, M. (2012) Structural probes in quadruplex nucleic acid structure determination by NMR. *Molecules*, **17**, 13073–13086.
77. Hansel, R., Foldynova-Trantirkova, S., Dotsch, V. and Trantirek, L. (2013) Investigation of quadruplex structure under physiological conditions using in-cell NMR. *Top. Curr. Chem.*, **330**, 47–65.
78. Del Villar-Guerra, R., Gray, R.D. and Chaires, J.B. (2017) Characterization of quadruplex DNA structure by circular dichroism. *Curr. Protoc. Nucleic Acid Chem.*, **68**, 17.8.1–17.8.16.
79. Del Villar-Guerra, R., Trent, J.O. and Chaires, J.B. (2018) G-Quadruplex secondary structure obtained from circular dichroism spectroscopy. *Angew. Chem. Int. Ed. Engl.*, **57**, 7171–7175.
80. Dean, W.L., Gray, R.D., DeLeeuw, L., Monsen, R.C. and Chaires, J.B. (2019) Putting a new spin of G-Quadruplex structure and binding by analytical ultracentrifugation. *Methods Mol. Biol.*, **2035**, 87–103.
81. Monsen, R.C., Chakravarthy, S., Dean, W.L., Chaires, J.B. and Trent, J.O. (2021) The solution structures of higher-order human telomere G-quadruplex multimers. *Nucleic Acids Res.*, **49**, 1749–1768.
82. Chaires, J.B., Trent, J.O., Gray, R.D., Dean, W.L., Buscaglia, R., Thomas, S.D. and Miller, D.M. (2014) An improved model for the hTERT promoter quadruplex. *PLoS One*, **9**, e115580.
83. 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.
84. Kyrp, J., Kejnovska, I., Renciuik, D. and Vorlickova, M. (2009) Circular dichroism and conformational polymorphism of DNA. *Nucleic Acids Res.*, **37**, 1713–1725.
85. Karsisiotis, A.I., Hessari, N.M., Novellino, E., Spada, G.P., Randazzo, A. and Webba da Silva, M. (2011) Topological characterization of nucleic acid G-quadruplexes by UV absorption and circular dichroism. *Angew. Chem. Int. Ed. Engl.*, **50**, 10645–10648.
86. Aboul-ela, F., Murchie, A.I. and Lilley, D.M. (1992) NMR study of parallel-stranded tetraplex formation by the hexadeoxynucleotide d(TG4T). *Nature*, **360**, 280–282.
87. Wang, Y. and Patel, D.J. (1993) Solution structure of a parallel-stranded G-quadruplex DNA. *J. Mol. Biol.*, **234**, 1171–1183.
88. Holm, A.I., Kohler, B., Hoffmann, S.V. and Brondsted Nielsen, S. (2010) Synchrotron radiation circular dichroism of various G-quadruplex structures. *Biopolymers*, **93**, 429–433.
89. Maruyama, R., Makino, K., Yoshitomi, T., Yui, H., Furusho, H. and Yoshimoto, K. (2018) Estimation of G-quartet-forming guanines in parallel-type G-quadruplexes by optical spectroscopy measurements of their single-nucleobase substitution sequences. *Analyst*, **143**, 4022–4026.
90. Mergny, J.L., De Cian, A., Ghelab, A., Sacca, B. and Lacroix, L. (2005) Kinetics of tetramolecular quadruplexes. *Nucleic Acids Res.*, **33**, 81–94.
91. Kirby, N., Cowieson, N., Hawley, A.M., Mudie, S.T., McGillivray, D.J., Kusel, M., Samardzic-Boban, V. and Ryan, T.M. (2016) Improved radiation dose efficiency in solution SAXS using a sheath flow sample environment. *Acta Crystallogr. D Struct. Biol.*, **72**, 1254–1266.
92. Hopkins, J.B., Gillilan, R.E. and Skou, S. (2017) BioXTAS RAW: improvements to a free open-source program for small-angle X-ray scattering data reduction and analysis. *J. Appl. Crystallogr.*, **50**, 1545–1553.
93. Meisburger, S.P., Taylor, A.B., Khan, C.A., Zhang, S., Fitzpatrick, P.F. and Ando, N. (2016) Domain movements upon activation of phenylalanine hydroxylase characterized by crystallography and chromatography-coupled small-angle X-ray scattering. *J. Am. Chem. Soc.*, **138**, 6506–6516.
94. Maeder, M. (1987) Evolving factor analysis for the resolution of overlapping chromatographic peaks. *Anal. Chem.*, **59**, 527–530.
95. Trehwella, J., Duff, A.P., Durand, D., Gabel, F., Guss, J.M., Hendrickson, W.A., Hura, G.L., Jacques, D.A., Kirby, N.M., Kwan, A.H. *et al.* (2017) 2017 publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution: an update. *Acta Crystallogr. D Struct. Biol.*, **73**, 710–728.
96. Franke, D. and Svergun, D.I. (2009) DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. *J. Appl. Crystallogr.*, **42**, 342–346.
97. Volkov, V.V. and Svergun, D.I. (2003) Uniqueness of ab initio shape determination in small-angle scattering. *J. Appl. Crystallogr.*, **36**, 860–864.
98. Petoukhov, M.V., Franke, D., Shkumatov, A.V., Tria, G., Kikhney, A.G., Gajda, M., Gorba, C., Mertens, H.D., Konarev, P.V. and Svergun, D.I. (2012) New developments in the ATSAS program package for small-angle scattering data analysis. *J. Appl. Crystallogr.*, **45**, 342–350.
99. Svergun, D.I. (1999) Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biophys. J.*, **76**, 2879–2886.
100. Svergun, D. (1992) Determination of the regularization parameter in indirect-transform methods using perceptual criteria. *Appl. Crystallogr.*, **25**, 495–503.
101. Tuukkanen, A.T., Kleywegt, G.J. and Svergun, D.I. (2016) Resolution of ab initio shapes determined from small-angle scattering. *IUCrJ*, **3**, 440–447.
102. Svergun, D., Barberato, C. and Koch, M.H.J. (1995) CRYSOLE - a Program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.*, **28**, 768–773.
103. Kozin, M.B. and Svergun, D.I. (2001) Automated matching of high- and low-resolution structural models. *J. Appl. Crystallogr.*, **34**, 33–41.
104. Ortega, A., Amoros, D. and Garcia de la Torre, J. (2011) Prediction of hydrodynamic and other solution properties of rigid proteins from atomic- and residue-level models. *Biophys. J.*, **101**, 892–898.
105. Garbett, N.C., Mekmaysy, C.S. and Chaires, J.B. (2010) Sedimentation velocity ultracentrifugation analysis for hydrodynamic characterization of G-quadruplex structures. *Methods Mol. Biol.*, **608**, 97–120.
106. Kikhney, A.G. and Svergun, D.I. (2015) A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Lett.*, **589**, 2570–2577.
107. Svergun, D.I. and Koch, M.H.J. (2003) Small-angle scattering studies of biological macromolecules in solution. *Rep. Prog. Phys.*, **66**, 1735–1782.
108. Fang, X., Stagno, J.R., Bhandari, Y.R., Zuo, X. and Wang, Y.X. (2015) Small-angle X-ray scattering: a bridge between RNA secondary structures and three-dimensional topological structures. *Curr. Opin. Struct. Biol.*, **30**, 147–160.
109. Gräwert, T.W. and Svergun, D.I. (2020) Structural modeling using solution small-angle X-ray scattering (SAXS). *J. Mol. Biol.*, **432**, 3078–3092.
110. Durand, D., Vivès, C., Cannella, D., Pérez, J., Pebay-Peyroula, E., Vachette, P. and Fieschi, F. (2010) NADPH oxidase activator p67phox behaves in solution as a multidomain protein with semi-flexible linkers. *J. Struct. Biol.*, **169**, 45–53.
111. Schneidman-Duhovny, D., Hammel, M., Tainer, J.A. and Sali, A. (2013) Accurate SAXS profile computation and its assessment by contrast variation experiments. *Biophys. J.*, **105**, 962–974.
112. Meier, M., Moya-Torres, A., Krahn, N.J., McDougall, M.D., Orriss, G.L., McRae, E.K.S., Booy, E.P., McEleney, K., Patel, T.R., McKenna, S.A. *et al.* (2018) Structure and hydrodynamics of a DNA G-quadruplex with a cytosine bulge. *Nucleic Acids Res.*, **46**, 5319–5331.
113. Mazzanti, L., Alferkh, L., Frezza, E. and Pasquali, S. (2021) Biasing RNA coarse-grained folding simulations with small-angle X-ray scattering data. *J. Chem. Theory Comput.*, **17**, 6509–6521.
114. Svergun, D.I., Richard, S., Koch, M.H., Sayers, Z., Kuprin, S. and Zaccai, G. (1998) Protein hydration in solution: experimental observation by x-ray and neutron scattering. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 2267–2272.
115. Laage, D., Elsaesser, T. and Hynes, J.T. (2017) Water dynamics in the hydration shells of biomolecules. *Chem. Rev.*, **117**, 10694–10725.
116. Petoukhov, M.V. and Svergun, D.I. (2015) Ambiguity assessment of small-angle scattering curves from monodisperse systems. *Acta Crystallogr. D Biol. Crystallogr.*, **71**, 1051–1058.
117. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.

118. Bernadó,P., Mylonas,E., Petoukhov,M.V., Blackledge,M. and Svergun,D.I. (2007) Structural characterization of flexible proteins using small-angle X-ray scattering. *J. Am. Chem. Soc.*, **129**, 5656–5664.
119. Halgren,T.A. (2009) Identifying and characterizing binding sites and assessing druggability. *J. Chem. Inf. Model.*, **49**, 377–389.
120. Holt,P.A., Ragazzon,P., Strekowski,L., Chaires,J.B. and Trent,J.O. (2009) Discovery of novel triple helical DNA intercalators by an integrated virtual and actual screening platform. *Nucleic Acids Res.*, **37**, 1280–1287.
121. Holt,P.A., Buscaglia,R., Trent,J.O. and Chaires,J.B. (2011) A discovery funnel for nucleic acid binding drug candidates. *Drug Dev. Res.*, **72**, 178–186.
122. Grun,J.T., Hennecker,C., Klotzner,D.P., Harkness,R.W., Bessi,I., Heckel,A., Mittermaier,A.K. and Schwalbe,H. (2020) Conformational dynamics of strand register shifts in DNA G-Quadruplexes. *J. Am. Chem. Soc.*, **142**, 264–273.
123. Marquevielle,J., Robert,C., Lagrabette,O., Wahid,M., Bourdoncle,A., Xodo,L.E., Mergny,J.L. and Salgado,G.F. (2020) Structure of two G-quadruplexes in equilibrium in the KRAS promoter. *Nucleic Acids Res.*, **48**, 9336–9345.
124. Kejnovská,I., Stadlbauer,P., Trantírek,L., Renčiuk,D., Gajarský,M., Krafčík,D., Palacký,J., Bednářová,K., Šponer,J., Mergny,J.L. *et al.* (2021) G-Quadruplex formation by DNA sequences deficient in guanines: two tetrad parallel quadruplexes do not fold intramolecularly. *Chemistry*, **27**, 12115–12125.
125. Hansel-Hertsch,R., Beraldi,D., Lensing,S.V., Marsico,G., Zyner,K., Parry,A., Di Antonio,M., Pike,J., Kimura,H., Narita,M. *et al.* (2016) G-quadruplex structures mark human regulatory chromatin. *Nat. Genet.*, **48**, 1267–1272.
126. Bell,R.J., Rube,H.T., Kreig,A., Mancini,A., Fouse,S.D., Nagarajan,R.P., Choi,S., Hong,C., He,D., Pekmezci,M. *et al.* (2015) Cancer. The transcription factor GABP selectively binds and activates the mutant TERT promoter in cancer. *Science*, **348**, 1036–1039.
127. Miller,M.C. and Trent,J.O. (2011) Resolution of quadruplex polymorphism by size-exclusion chromatography. *Curr. Protoc. Nucleic Acid Chem.*, **Chapter 17**, Unit17 13.
128. Adrian,M., Heddi,B. and Phan,A.T. (2012) NMR spectroscopy of G-quadruplexes. *Methods*, **57**, 11–24.
129. Lin,C., Dickerhoff,J. and Yang,D. (2019) NMR studies of G-Quadruplex structures and G-Quadruplex-Interactive compounds. *Methods Mol. Biol.*, **2035**, 157–176.
130. Gray,R.D., Petraccone,L., Buscaglia,R. and Chaires,J.B. (2010) 2-aminopurine as a probe for quadruplex loop structures. *Methods Mol. Biol.*, **608**, 121–136.
131. Maleki,P., Budhathoki,J.B., Roy,W.A. and Balci,H. (2017) A practical guide to studying G-quadruplex structures using single-molecule FRET. *Mol. Genet. Genomics*, **292**, 483–498.