Structural Studies of DNA Repeats Implicated in Cancer

Presented as a Senior Thesis in Chemistry

Deondre Jordan
May 17, 2019
Swarthmore College

Advisor: Liliya A. Yatsunyk
Acknowledgements

Joining the Yatsunyk lab in the spring of my freshmen year at Swarthmore College, I could not imagine all the experiences I would gain and the impact research would have on me. The supportive efforts of those both in and outside of the laboratory helped make this work possible. I have to start by thanking Professor Yatsunyk for mentoring me through my entire time at Swarthmore. With unwavering enthusiasm and care, she has helped me grow academically, professionally, and, personally. I am also thankful for my fellow Yatlab members: Linda Lin, Hyun Kyung Lee, Ariana Yett, Yanti Manurung, Samantha Nyovanie Joanne Miao, Dana Beseiso, Allan Gao, Letitia Ho, Amber Sheth, Sayed Malawi, Irene Xiang, Alice Liu, Jessica Chen, and Barrett Powell, for fostering a supportive and enjoyable lab experience. I want to give a special thanks to Barrett for being my partner-in-crime and one of my biggest inspirations for our two and half years together.

I want to extend my thanks to our collaborators at the University of Pennsylvania, the University of Louisville, and the NIH. Thank you Eric Brown and Nishita Shastri for consistently bringing excitement to our CAGAGG project and allowing me to conduct new experiments with you. Thank you Brad Chaires for being one of my science heroes. Many of your projects inspired the work I present here and significantly shaped me as an investigator of nucleic acids. Thank you Brad, John Trent, Bill Dean, and Robert Gray for hosting Yatlab (for Biophys Bootcamp) and helping me grow as a scientist. I want to also thank Jay Schneekloth and David Calabrese for being fun and inviting collaborators.

To the Swarthmore chemistry and biochemistry communities, thank you for supporting me through my undergraduate career. Special thanks to some my fellow classmates who made learning chemistry enjoyable: Jeffrey Zhou, Colin Howell, Audra Woodside, Laela Ezra, and Diep Nyugen. I want to share equal thanks with my professors and lab instructors for the wonderful instruction you provided: Bob Paley, Paul Rablen, Kathryn Riley, Chris Graves, Kathleen Howard, Daniela Fera, Stephen Miller, Kyle Wagner, Silvia Porello, Jessica Sarver, Donna Halley, Caroline Burkhard, Maria Newport, Ginger Heck, Lori Sonntag. Thank you for Ian McGarvey for your flexibility and industriousness, keeping the instruments running. Thank you Cathy Cinquina for helping me and many other lab members with our research that required shipment.

Finally, I want to thank my family for supporting me through this journey. Special thank you to the Chester Children’s Chorus for exposing me to Swarthmore and the academic opportunities it offered. I am especially grateful for John Alston because he always pushed me to achieve more both in and outside of the laboratory. To my mother, grandmother, and grandfather, thank you so much for always encouraging me to succeed at every step of my journey here.
Abstract

DNA replication is an essential process for the survival of all organisms. Errors can occur during the process that leads to replication stress, which can promote genomic instability, a hallmark of cancer. The Brown lab conducted a genome wide study to identify sites and sequences linked to replication stress. Tandem DNA repeats comprised the sequences identified in this study. One particular DNA repeat \((CAGAGG)\_n\) stood out due to being highly associated with replication stress. Repeats such as \((CAGAGG)\_n\) are believed to fold into intramolecular structures that physically block DNA polymerase during replication. Understanding the structures of these problematic DNA repeats may shed light on the mechanism behind replication stress and encourage the design of therapeutics to treat them.

This work starts with biophysical characterization of identified DNA repeats associated with replication stress. This study highlighted \((CAGAGG)\_n\) as one of the most stable, well-folded, and spectroscopically unique repeats. Given the biological and biophysical data, we focused our efforts on elucidating the secondary structure of \((CAGAGG)\_n\). Based on previous biophysical characterization in the Yatsunyk laboratory, a hypothetical model was designed for the smallest stable structure formed by the DNA repeat. The model is a tetrastranded monomer comprised of two-stacked GCGC tetrads connected by three lateral AGAG loops in an antiparallel fashion. The rest of the work presented here focuses on testing and refining the model for \((CAGAGG)\_n\).

To shed light on the secondary structure of \((CAGAGG)\_n\), we conducted a ligand screen with over 20 known nucleic acid binders. If the DNA repeat interacts with a class of ligands specific for particular DNA structures (such as B-DNA), we can gain further structural information. Secondly, if a binder stabilizes the unique fold of \((CAGAGG)\_n\), it can be used as a co-crystallization agent to produce high-quality crystals for structure determination via X-ray crystallography. This study showed \((CAGAGG)\_n\) interacts mostly with ligands specific for G-Quadruplex DNA. We also discovered Methylene Blue provided moderate stability to the DNA repeat. This ligand screen was extended using a Small Molecule Microarray assay, where 7042 ligands were simultaneously screened with \((CAGAGG)\_n\). Biophysical characterization of the interactions of \((CAGAGG)\_n\) with some of the best hits from this assay showed one binder worth further study.

To test our model for \((CAGAGG)\_n\), we designed a variety of structural studies based on mutations. Some of these studies include: single point mutations, double point mutations, addition and removal of GCGC tetrads, loop length variations, and substitution of adenine with 2-amino purine. These studies showed the core was crucial for the stability, while the core and the loops together maintain the structural integrity. The loops proved to be crucial for the structure formation where the minimum loop length for the unique \((CAGAGG)\_n\) fold is three nucleotides. Overall, our biophysical studies corroborated our hypothetical model and shed light on the importance of the loops for the unique structure.
# Table of Contents

Acknowledgements ........................................................................................................... ii

Abstract ................................................................................................................................ iii

Table of Contents ................................................................................................................ iv

**Chapter 1: DNA replication, DNA repeats, and non-canonical DNA structures**
- DNA replication and replication stress .................................................................................. 1
- DNA repeats and non-canonical DNA structures ................................................................. 5

**Chapter 2: Materials and Methods** .................................................................................. 7
- DNA and buffers .................................................................................................................. 7
- Ligands .................................................................................................................................... 10
- UV-visible spectroscopy (UV) ............................................................................................. 13
- Circular dichroism spectroscopy (CD) ................................................................................ 15
- Thermal difference spectra (TDS) ...................................................................................... 17
- Polyacrylamide gel electrophoresis (PAGE) ...................................................................... 18
- Fluorescent spectroscopy studies (FL) ................................................................................ 19
- Small molecular microarray (SMM) ................................................................................... 19

**Chapter 3: DNA repeats linked to replication stress** ........................................................... 21
- Biophysical screen of DNA repeats ...................................................................................... 21
- Building the model of (CAGAGG)_n ..................................................................................... 29

**Chapter 4: Ligand screen of (CAGAGG)_n with DNA binders** ............................................. 34
- Interactions of 1Core with small molecule ligands ............................................................. 34
- Small molecule microarray screen of 1Core+2 ................................................................... 42

**Chapter 5: Mutagenesis study of 1Core+2** ........................................................................ 49
- Point mutation study ......................................................................................................... 49
- Loop variations ................................................................................................................. 55
- Fluorescence studies of 1Core+2 2-amino purine oligos ................................................. 65

**Chapter 6: Conclusion** .................................................................................................... 72

References ............................................................................................................................. 74

Appendix ................................................................................................................................ 80
- Chapter 3: Ion dependence study of (CAGAGG)_n ............................................................. 80
- Chapter 4: Interactions of 1Core with small molecule ligands ........................................... 82
- Chapter 4: Small molecule microarray screen of 1Core+2 .............................................. 94
- Chapter 5: Point mutant study of 1Core+2 ...................................................................... 103
- Chapter 5: Double mutant study of 1Core+2 ................................................................... 105
- Chapter 5: Three-dimensional melting studies of 1Core+2 ............................................ 109
- Longer more biologically relevant repeats of (CAGAGG)_n: 2Core+2 ............................. 112
  - Point mutation study of 2Core+2 .................................................................................. 112
  - Fluorescence studies of 2Core+2 2-AP oligos ............................................................... 119
  - Three-dimensional melting studies of 2Core ............................................................... 123
- Next Steps/To-Do: ............................................................................................................. 126
- Copy of published work ................................................................................................... 127
Chapter 1: DNA replication, DNA repeats, and non-canonical DNA structures

DNA replication and replication stress

The cell-division cycle is a critical process for the growth and maintenance of all organisms, requiring accurate replication of over 6 billion DNA base pairs. The cell cycle consists of four distinct phases: Gap 1 (G1), Synthesis (S), Gap 2 (G2), and Mitosis (M). This series of events produces duplicate DNA and give rise to two new daughter cells. This intricate process contains many steps and depends on a large network of enzymes for proper cell division. Despite self-regulating cell cycle checkpoints, problems still occur through cell division and could lead to diseases like cancer. Cancer is typically associated with abnormal cell growth and uncontrolled cell proliferation, both results of complications during the G1, G2, and M phases. Like the other phases of the cell cycle, DNA replication (Figure 1) during the S phase contributes to cancer. This phase is particularly important for the survival of the cell because its success determines the integrity of the genome. If DNA replication occurs with a high degree of fidelity, essential genes may carry mutations deleterious to the fate of the cell.

DNA replication typically begins at specific sites within the genome called origins where DNA gyrase (a topoisomerase) will diminish the topological strain of canonical duplex DNA (B-DNA). Helicase then unwinds the unstrained DNA and creates a replication fork site, making two single stranded DNA (ssDNA) branches known as the
leading strand and lagging strand. These ssDNA regions are coated by an ensemble of proteins including the well-studied Replication Protein A (RPA), preventing the ssDNA from folding back on itself. DNA polymerase binds to either of the stands and synthesizes B-DNA from 5’ to 3’ by replicating the complementary strands of the leading and lagging strands. DNA polymerase on the leading strand makes B-DNA in the same direction of the replication fork formation. Synthesis of B-DNA on the lagging strand occurs in the opposite direction of the forming replication fork. An RNA primase is required to synthesize a short RNA primer to initiate the synthesis on the both strands. On the lagging strand, these primers allow DNA polymerase to make small, separated fragments known as Okazaki fragments that will finally be connected by DNA ligase.

Figure 1. Summary of DNA replication. Simplified model of DNA replication, highlighting the function of many key enzymes.

This complex replicative process is prone to errors that lead to inefficient DNA replication known as replication stress. Replication stress results in slowed or stalled DNA replication forks; many factors contribute to replication stress such as decreased origin activity, hyper-replication of DNA, DNA regions difficult to unwind, and non-canonical DNA secondary structures. DNA damage response (DDR) systems have been
evolved to account for replication stress.\textsuperscript{2} DDR is triggered by persistent ssDNA sites coated with Replication protein A (RPA) that signify replication fork stalling. RPA is a protein that binds ssDNA in eukaryotic cells. The RPA proteins saturating the ssDNA sites recruit a complex of proteins including ataxia telangiectasia and Rad3-related protein (ATR)-interacting protein (ATRIP), recruiting ATR to sites of replication stress.\textsuperscript{3} ATR is considered one of the guardians of the genome because of its ability to respond to replication stress by stopping the cell cycle and stabilizing and repairing stalled replication forks.\textsuperscript{4,5} Without ATR, replicative problems such as double-strand breaks (DSB) and replication fork collapse can go unchecked, leading to genomic instability—one of the hallmarks of cancer.\textsuperscript{6,7}

Identifying sites throughout the genome that elicit DDR may provide mechanistic insight into replication stress and genomic instability. Glover et al. initiated this search and observed common fragile sites in the genome by inducing chromosomal rearrangements with aphidicolin, an inhibitor of DNA polymerase α.\textsuperscript{8} Similar studies have identified specific sequences that are linked to replication stress such as telomeric repeats and trinucleotide expansion repeats.\textsuperscript{9,10} Many studies which report intrinsically difficult-to-replicate sequences use a candidate approach which is based on pre-specified sites of interest. To take a less biased approach towards identifying difficult-to-replicate sequences, the Brown laboratory used ATR inhibition (ATRi) and RPA-ChIP sequencing (RPA-ChIP-seq) as shown in Figure 2.\textsuperscript{11} ATRi promotes sites of replication fork collapse, which will be saturated with RPA. RPA-ChIP-seq then identifies those sites called replication perturbed loci (RPLs). This study revealed over 500 RPLs; many of these sites had strong correlations with microsatellite (di-,tri-,or
tetra-nucleotide) repeats. While some of these repeats were familiar like the telomeric and trinucleotide repeats, less familiar repeats where those that had significantly higher RPA accumulation. These DNA repeats are suspected to fold into stable intramolecular structures at stalled fork sites, physically blocking DNA polymerase. The functional uncoupling between DNA polymerase and helicase will trigger DSB or replication fork collapse. The ATR dependent DNA repeat most commonly observed in RPLs was purine-rich (CAGAGG)$_n$. Because the hexanucleotide repeat is heavily linked to replication stress, the Yatsunyk laboratory focused specifically on solving its structure.

**Figure 2. ATRi induced fork collapse or DSB.** Replication stress naturally occurs, triggering recruitments of ATR at stalled sites. Without ATR, stalled forks become destabilized and fail to restart, resulting in fork collapse. In particular cases, these destabilized forks cause DSB formation.
DNA repeats and non-canonical DNA structures

Interestingly, this genome-wide investigation by the Brown lab found DNA repeats to be intrinsically difficult-to-replicate sequences. Most of the genome is comprised of simple repetitive sequences with yet unknown functions. When the highly repetitive nature of the genome was first discovered, repetitive DNA were considered to be junk or selfish DNA that contained no particular function. Various studies demonstrated the diversity and function of tandem DNA repeats. One classic example is telomeric DNA (TTAGGG)$_n$, which exists at the end of chromosomes and can lengthen indefinitely in the presence of telomerase, extending the life of a cell. DNA repeats have gained attention for their involvement in the so-called expansion diseases such as Mytonic dystrophy (type I and II) and fragile X syndrome. The most notorious expansion repeat is (CAG)$_n$, causing neurodegenerative complications (such as Huntington’s disease) by making proteins with abnormal polyglutamine stretches. Over 30 hereditary disorders are due to these simple expansion repeats.

Tandem DNA repeat are capable of folding into structures beyond canonical B-DNA. In general, DNA repeats may adopt various folds such as B-DNA, stem-loop hairpins, triplex DNA (H-DNA), G-Quadruplex (GQ), and i-motif (Figure 3). Telomeric DNA folds into GQ which are tetrastranded DNA structures that have been well characterized over the last several decades. Other DNA repeats such as (CAG)$_n$, (CGG)$_n$, and (CAGG)$_n$ have less characterized folds. Most of the expansion repeats including (CAG)$_n$ and (CAGG)$_n$ are suspected to fold into simple hairpins which are common in biology. Interestingly, previous findings has shown (CCG)$_n$ folds into a
stable tetrahelical structure (similar to GQ).\textsuperscript{20,21} Much less is known about the relationship the DNA repeats structure and their biological activity.

\textbf{Figure 3. Various non-canonical structures of DNA.} Besides B-DNA, DNA can fold into a non-canonical duplex, triplex, and quadruplex (GQ and i-motif) DNA structures.\textsuperscript{22,23}

Structural diversity of DNA repeats provide many possibilities for the fold of the tandem repeats linked to replication stress like (CAGAGG)\textsubscript{n}. Using biophysical methods, the Yatsunyk laboratory has characterized the folds of tandem repeats with the goal of elucidating their secondary structure. Knowledge of the tandem DNA repeat structure may inform the mechanism underlying the role of repeats in genomic instability. This information could prove indispensable for the rational design of novel anti-cancer therapeutics.
Chapter 2: Materials and Methods

DNA and buffers

DNA was purchased from Integrated DNA Technologies (IDT, Coralville, IA) and purified by standard desalting or by HPLC. A full list of the oligonucleotides used is included in Table 1. DNA was hydrated with doubly deionized H₂O (ddH₂O) at 0.25 - 2 mM and stored at -80°C. All experiments were conducted in a buffer composed of 100 mM KCl, 2 mM MgCl₂, and 10 mM lithium cacodylate at pH 7.2 (100K2Mg buffer).

Prior to all experiments, DNA was diluted in 100K2Mg buffer to the desired concentration and annealed by heating samples to 95 °C for 5 min followed by cooling for 4 hrs to room temperature. Samples were then equilibrated at 4°C overnight (for a minimum of 12 hrs). The concentrations for all samples were determined at 95 °C (when DNA is completely unfolded) using extinction coefficients calculated based on the nearest neighbor method from IDT’s calculator and shown in Table 1.

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
<th>Extinction Coefficient (M⁻¹cm⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1Core</td>
<td>G(CAGAGG)₃C</td>
<td>206300</td>
</tr>
<tr>
<td>1Core+2</td>
<td>AGG(CAGAGG)₂CAG</td>
<td>253300</td>
</tr>
<tr>
<td>1Core+2 Cy5</td>
<td>Cy5–AGG(CAGAGG)₂CAG</td>
<td>263300</td>
</tr>
<tr>
<td>2Core</td>
<td>G(CAGAGG)₂C</td>
<td>457900</td>
</tr>
<tr>
<td>2Core+2</td>
<td>AGG(CAGAGG)₂CAG</td>
<td>504900</td>
</tr>
<tr>
<td>2RSS-S</td>
<td>TAG(CAGAGG)₂CAT</td>
<td>500600</td>
</tr>
<tr>
<td>VK1</td>
<td>GGGAGCGAGGGAGGCGAGCGAGCG</td>
<td>155200</td>
</tr>
<tr>
<td>VK2</td>
<td>GGGAGCGAGGGAGGCGAGCGAGGGAGG</td>
<td>322200</td>
</tr>
<tr>
<td>GCATGCT</td>
<td>GCATGCT</td>
<td>63000</td>
</tr>
</tbody>
</table>
L0    AGG CG CG CG CAG       114700
L1    AGG CAG CAG CGG CAG       250400
L2    AGG CAGG CAGG CAGG CAG       183400
1Core+2 424 AGG CAGAGG CAGAG CAGAG CAG       230000
L3-GAG AGG CAGAGG CAGAG CAGAG CAG       230000
L3-AAG AGG CAGGG CAGGG CAGGG CAG       219400
L3-AGA AGG CAGAG CAGAG CAGAG CAG       223000
L6    AGG CAGAGAGG CAGAGAGG CAGAGAGG CAG       323200
L8    AGG CAGAGAGAGG CAGAGAGAGG CAGAGAGAGG CAG       231400
TGAT  AGG CTGATG CTGATG CTGATG CAG       301300
TGAGAT AGG CTGAGATG CTGAGATG CTGAGATG CAG       301300

CAGG-22 (CAGG)_{22}       872500
CCTG-22 (CCTG)_{22}       692100
c-CT18 (CACAG)_{18}       885100
CT18 (CTGTG)_{18}       775300
CAGAGTG-15 (CAGAGT)_{15}       940400
ACTCTG-15 (ACTCTG)_{15}       795200
CACAGG-15 (CACAGG)_{15}       889300
CCTGTG-15 (CCTGTG)_{15}       754300
CAGAGAGG-11 (CAGAGAGG)_{11}       949500
CCTCTCTG-11 (CCTCTCTG)_{11}       683300
CA15 (CAGAGG)_{15}       944800
CCTCTG-15 (CCTCTG)_{15}       701800
CAGAGGG-12 (CAGAGGG)_{12}       877300
CCCTCTG-12 (CCCTCTG)_{12}       648100
Endo CTGTG GT CTGTG CTGTG CTGTG CTGTG CTGTG CTGTG CTGTG CTGTG CTGTG CTGTG CTGTG CTGTG CT
C rec CTGTG

[c-Endo CTGTG AG CACAG ACAG CACAG ACACG CACAG TACAG CACAG CACAG CACAG CACAG CACAG CACAG CACAG CACAC AG AC

1-5 AGGCACAGGCAGGGCAGGAGGCAGAGGCAG       241500
1-7 AGGCACAGGCAGGGCAGGAGGCAGAGGCAG       241100
1-13 AGGCAGAGGCAGGGCAGGAGGCAGAGGCAG       241100
1-19 AGGCAGAGGCAGGGCAGGAGGCAGAGGCAG       241100
1-23 AGGCAGAGGCAGGGCAGGAGGCAGAGGCAG       241500
2-5 AGGCAGAGGCAGGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGGAGGCAGG

8
<table>
<thead>
<tr>
<th></th>
<th>Sequence</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1T</td>
<td>TGG CAGAGG CAGAGG CAGAGG CAG</td>
<td>251300</td>
</tr>
<tr>
<td>G2T</td>
<td>ATG CAGAGG CAGAGG CAGAGG CAG</td>
<td>253100</td>
</tr>
<tr>
<td>G3T</td>
<td>AGT CAGAGG CAGAGG CAGAGG CAG</td>
<td>256600</td>
</tr>
<tr>
<td>C4T</td>
<td>AGG TAGAGG CAGAGG CAGAGG CAG</td>
<td>248000</td>
</tr>
<tr>
<td>A5T</td>
<td>AGG CTGAGG CAGAGG CAGAGG CAG</td>
<td>252100</td>
</tr>
<tr>
<td>G6T</td>
<td>AGG CATAGG CAGAGG CAGAGG CAG</td>
<td>248800</td>
</tr>
<tr>
<td>A7T</td>
<td>AGG CAGTGG CAGAGG CAGAGG CAG</td>
<td>251300</td>
</tr>
<tr>
<td>G8T</td>
<td>AGG CAGATT CAGAGG CAGAGG CAG</td>
<td>253100</td>
</tr>
<tr>
<td>G9T</td>
<td>AGG CAGAGT CAGAGG CAGAGG CAG</td>
<td>256600</td>
</tr>
<tr>
<td>C10T</td>
<td>AGG CAGAGG TAGAGG CAGAGG CAG</td>
<td>248000</td>
</tr>
<tr>
<td>A11T</td>
<td>AGG CAGAGG CTGAGG CAGAGG CAG</td>
<td>252100</td>
</tr>
<tr>
<td>G12T</td>
<td>AGG CAGAGG CATGGG CAGAGG CAG</td>
<td>248800</td>
</tr>
<tr>
<td>A13T</td>
<td>AGG CAGAGG CAGTGG CAGAGG CAG</td>
<td>251300</td>
</tr>
<tr>
<td>G14T</td>
<td>AGG CAGAGG CAGATG CAGAGG CAG</td>
<td>253100</td>
</tr>
<tr>
<td>G15T</td>
<td>AGG CAGAGG CAGAGT CAGAGG CAG</td>
<td>256600</td>
</tr>
<tr>
<td>C16T</td>
<td>AGG CAGAGG CAGAGG TAGAGG CAG</td>
<td>248000</td>
</tr>
<tr>
<td>A17T</td>
<td>AGG CAGAGG CAGAGG CTGAGG CAG</td>
<td>252100</td>
</tr>
<tr>
<td>G18T</td>
<td>AGG CAGAGG CAGAGG CATAGG CAG</td>
<td>248800</td>
</tr>
<tr>
<td>A19T</td>
<td>AGG CAGAGG CAGAGG CATGGG CAG</td>
<td>251300</td>
</tr>
<tr>
<td>G20T</td>
<td>AGG CAGAGG CAGAGG CAGATG CAG</td>
<td>253100</td>
</tr>
<tr>
<td>G21T</td>
<td>AGG CAGAGG CAGAGG CAGAGT CAG</td>
<td>256600</td>
</tr>
<tr>
<td>C22T</td>
<td>AGG CAGAGG CAGAGG CAGAGG TAG</td>
<td>248000</td>
</tr>
<tr>
<td>A23T</td>
<td>AGG CAGAGG CAGAGG CAGAGG CTG</td>
<td>253100</td>
</tr>
<tr>
<td>G24T</td>
<td>AGG CAGAGG CAGAGG CAGAGG CAT</td>
<td>247300</td>
</tr>
<tr>
<td>G2C-G3T</td>
<td>AGG TAGAGG CAGAGG CAGAGG CAG</td>
<td>248600</td>
</tr>
<tr>
<td>G8C-G9T</td>
<td>AGG CAGACT CAGAGG CAGAGG CAG</td>
<td>248600</td>
</tr>
<tr>
<td>G14C-G15T</td>
<td>AGG CAGAGG CAGACT CAGAGG CAG</td>
<td>248600</td>
</tr>
<tr>
<td>G20C-G21T</td>
<td>AGG CAGAGG CAGAGG CAGACT CAG</td>
<td>248600</td>
</tr>
<tr>
<td>1GC1</td>
<td>AG CAGAGG AGAG CAGAGG AG</td>
<td>220700</td>
</tr>
<tr>
<td>1GC2</td>
<td>AGG AGAG CAGAGG AGAG CAG</td>
<td>220700</td>
</tr>
<tr>
<td>3GCa</td>
<td>AGG CGAGACG CAGAGG CGAGACG CAG</td>
<td>287500</td>
</tr>
<tr>
<td>3GCb</td>
<td>AGCG CAGAGG CGAGACG CAGAGG CGAG</td>
<td>287500</td>
</tr>
</tbody>
</table>
Ligands

The collection of ligands investigated here came from Prof. Jonathan B. Chaires’ list of nucleic acid binders. A list of all the ligands including their extinction coefficients can be found in Table 2. Berberine Chloride Form, Berenil, Coralyne Chloride Form, Hoechst 33258, and Pentamidine isethionate salt were all gifts from Rafael Del Virra Guerra (University of Louisville). Neutral Red, Methylene Blue, Malachite Green, Methyl Green, Ethidium Bromide, Coomassie Brilliant Blue, and Crystal Violet were gifts from the Biology Department at Swarthmore College. Methylene Blue, Crystal Violet, and Coomassie Brilliant Blue were stored as dry stocks at 4 °C and diluted to 2 mM in ddH2O before experiments. CuTMPyP4, Phen-DC3, Pentamidine Isethionate Salt, Berberine Chloride, and Coralyne Chloride were prepared in DMSO at 5-10 mM concentration and stored at -80 °C. Stocks of all other ligands were prepared in ddH2O and were stored at -80 °C or once diluted for the experiment, kept at 4 °C. Ligand aliquots were diluted to the desired concentrations right before the experiment in 100K2Mg buffer.

Table 2. *A full list of the ligands studied, including abbreviations, solvents, and extinction coefficients.*

<table>
<thead>
<tr>
<th>Name</th>
<th>Abbreviation</th>
<th>Solvent</th>
<th>Extinction Coefficient (M⁻¹cm⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-methyl mesoporphyrin IX</td>
<td>NMM</td>
<td>H₂O</td>
<td>1.45 × 10⁵ at 379 nm</td>
</tr>
<tr>
<td>5,10,15,20-tetrakis (N-methyl-4-pyridyl) poryphyrin</td>
<td>TMPyP4</td>
<td>H₂O</td>
<td>2.26 × 10⁵ at 424 nm</td>
</tr>
<tr>
<td>5,10,15,20-tetrakis (N-methyl-3-pyridyl) poryphyrin</td>
<td>TMPyP3</td>
<td>H₂O</td>
<td>2.5 × 10⁵ at 417 nm</td>
</tr>
<tr>
<td>5,10,15,20-tetrakis (N-methyl-2-pyridyl) poryphyrin</td>
<td>TMPyP2</td>
<td>H₂O</td>
<td>1.82 × 10⁵ at 414 nm</td>
</tr>
<tr>
<td>Zn(II)-5,10,15,20-tetrakis (N-methyl-4-pyridyl) poryphyrin</td>
<td>ZnTMPyP4</td>
<td>H₂O</td>
<td>2.04 × 10⁵ at 437 nm</td>
</tr>
<tr>
<td>Ligand</td>
<td>Formula</td>
<td>Solvent</td>
<td>IC50</td>
</tr>
<tr>
<td>--------------------------------------------</td>
<td>-----------</td>
<td>---------</td>
<td>----------</td>
</tr>
<tr>
<td>Cu(II)-5,10,15,20-tetrakis (N-methyl-4-pyridyl) porphyrin</td>
<td>CuTMPyP4</td>
<td>H2O</td>
<td>2.31 x 10^5</td>
</tr>
<tr>
<td>Pentacationic Mangenese (III) porphyrin</td>
<td>PM3P</td>
<td>H2O</td>
<td>5.93 x 10^4</td>
</tr>
<tr>
<td>RHSP4</td>
<td>DMSO</td>
<td></td>
<td>1.03 x 10^5</td>
</tr>
<tr>
<td>Braco 19</td>
<td>DMSO</td>
<td></td>
<td>2.35 x 10^4</td>
</tr>
<tr>
<td>Phen-DC3</td>
<td>DMSO</td>
<td></td>
<td>1.01 x 10^4</td>
</tr>
<tr>
<td>Pyridostatin</td>
<td>PDS</td>
<td>H2O</td>
<td>6.15 x 10^4</td>
</tr>
<tr>
<td>Crystal Violet</td>
<td></td>
<td>H2O</td>
<td>3.63 x 10^5</td>
</tr>
<tr>
<td>Coomassie Brilliant Blue r-250</td>
<td></td>
<td>H2O</td>
<td>4.3 x 10^4</td>
</tr>
<tr>
<td>Methylene Blue</td>
<td></td>
<td>H2O</td>
<td>7.40 x 10^4</td>
</tr>
<tr>
<td>Methyl Green</td>
<td></td>
<td>H2O</td>
<td>8.53 x 10^4</td>
</tr>
<tr>
<td>Malachite Green</td>
<td></td>
<td>H2O</td>
<td>1.49 x 10^5</td>
</tr>
<tr>
<td>Neutral Red</td>
<td>NR</td>
<td>H2O</td>
<td>2.93 x 10^3</td>
</tr>
<tr>
<td>Ethidium Bromide</td>
<td>EB</td>
<td>H2O</td>
<td>3.39 x 10^3</td>
</tr>
<tr>
<td>Coralyne Chloride Form</td>
<td></td>
<td>DMSO</td>
<td>1.45 x 10^4</td>
</tr>
<tr>
<td>Berberine Chloride</td>
<td></td>
<td></td>
<td>2.25 x 10^4</td>
</tr>
<tr>
<td>Diminizaene Aceturate</td>
<td>Berenil</td>
<td>DMSO</td>
<td>3.33 x 10^4</td>
</tr>
<tr>
<td>Pentamidine Isethionate Salt</td>
<td></td>
<td>H2O</td>
<td>2.89 x 10^4</td>
</tr>
<tr>
<td>Bisbenzimidie</td>
<td></td>
<td>H2O</td>
<td>4.6 x 10^4</td>
</tr>
</tbody>
</table>

Most ligands found in the Small molecule microarray study were purchased from Chemical Diversity (San Diego, CA). Dr. John Schneekloth and Dr. David Calabrese of the NIH provided other ligands used in this study. A full list of the ligands from this study is included in Table 3. All ligand stock solutions were prepared by dissolving ligand powder in DMSO to 10 mM using the mass of powder stocks provided by the manufacturer. All stock solutions were stored at -80°C. Ligand aliquots were diluted to the desired concentrations right before the experiment in 100K2Mg buffer.
Table 3. List of ligands from SMM study with 1Core, showing their code name and molecular structures. These were the best hits according to SMM studies in both PBS and 100K2Mg buffers.

<table>
<thead>
<tr>
<th>Code</th>
<th>Structure</th>
<th>Code</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>F025-0042</td>
<td><img src="image1.png" alt="Structure" /></td>
<td>V020-6722</td>
<td><img src="image2.png" alt="Structure" /></td>
</tr>
<tr>
<td>F025-0098</td>
<td><img src="image3.png" alt="Structure" /></td>
<td>V022-1170</td>
<td><img src="image4.png" alt="Structure" /></td>
</tr>
<tr>
<td>G702-2203</td>
<td><img src="image5.png" alt="Structure" /></td>
<td>V022-2346</td>
<td><img src="image6.png" alt="Structure" /></td>
</tr>
<tr>
<td>G702-2334</td>
<td><img src="image7.png" alt="Structure" /></td>
<td>V022-1170</td>
<td><img src="image8.png" alt="Structure" /></td>
</tr>
<tr>
<td>G725-1560</td>
<td><img src="image9.png" alt="Structure" /></td>
<td>V022-2346</td>
<td><img src="image10.png" alt="Structure" /></td>
</tr>
<tr>
<td>V008-2251</td>
<td><img src="image11.png" alt="Structure" /></td>
<td>L827-1823</td>
<td><img src="image12.png" alt="Structure" /></td>
</tr>
</tbody>
</table>
UV-visible spectroscopy (UV)

UV titrations

UV-vis titrations were performed with 1Core and 1Core+2 for the following ligands: TMPyP4, TMPyP3, PM3P, RHPS4, Methylene Blue, ZnTMPyP4, and CuTMPyP4. UV-vis titrations were performed on the Cary 300 Bio UV Spectrophotometer (Agilen). Ligands absorption spectra change upon interaction with DNA molecules, allowing determination of binding affinity and stoichiometry between ligands and DNA. The selected ligand was diluted to 1 mL in a 1-cm methamethacrylate cuvette to concentrations ranging from 5 to 10 µM in 100K2Mg buffer. DNA was in the same buffer at ~25 to 50 µM so that a 5:1 ratio of DNA: Ligand would be attained by the end of the titration. No more than ~100 µL DNA was added to the ligand to keep change in volume below 10%. The concentration of ligand in the cuvette was matched with the ligand concentration in the DNA sample to keep ligand concentration constant through the experiment. DNA sample was kept on ice throughout the titration. Titration was deemed complete when the absorption spectrum remained unchanged after addition of DNA for at least three additions. Data were collected from 352 - 670 nm with 0.1 s averaging time and 0.5 nm intervals at 25 °C. Titrations were repeated at least three times. The binding constant (Kₐ) was determined using the Direct Fit method details of which can be found in Bhattachrjee et. al.²⁴ In short, the percent hypochromicity (%H) and the red shift (Δλ) were determined to provide information about the binding mode. Percent Hypochromicity was determined using the following formula:

\[
%H = \frac{\varepsilon_{\text{free}} - \varepsilon_{\text{bound}}}{\varepsilon_{\text{free}}}
\]
where $\varepsilon_{\text{free}}$ is the extinction coefficient of free ligand and $\varepsilon_{\text{bound}}$ is the maximum absorbance of the final spectrum divided by total ligand concentration. The red shift was calculated using the following formula:

$$\Delta \lambda = \lambda_{\text{max}}(\text{complex}) - \lambda_{\text{max}}(\text{ligand})$$

where the maximum wavelength of the complex was subtracted from the maximum wavelength of the ligand alone.

**UV-vis stability studies**

UV-vis melting studies were performed at wavelengths 260, 295, and 330 nm for all UV melts. To detect duplex DNA, 260 nm was monitored, while 295 nm was monitored to detect quadruplex DNA structures. The 330 nm wavelength was monitored as reference and baseline; the absorbance at 330 nm was subtracted from both the melting and cooling curves monitored at 260 and 295 nm. Melting studies were completed in 1-cm pathlength quartz cuvettes. Data is typically collected from 4 – 95 °C, 2 nm bandwidth, 1 °C temperature step, 2 sec averaging time. Both melting and cooling curves of all samples were collected to determine the reversibility of the melting process. UV-vis melts were analyzed with two methods. In the first method, the data was smoothed using a 13-point Savitzky-Golay second order polynomial, followed by taking the first derivative. The melting temperature ($T_m$) was determined at the extrema of the first derivative curves. These values are associated with an error of ±0.5 °C since the $T_m$ were read by eye. The second method uses a nonlinear curve fit on the data, assuming a two-state equilibrium with a constant $\Delta H$ (heat capacity, $C_p = 0$).25 Fitting the data to this model provides both the $T_m$ and the $\Delta H$ values, assuming that the residuals are random. Melting and cooling data are processed separately and the hysteresis (difference in $T_m$
between the heating and cooling curves) was determined. When the melting transition was poorly defined in the melting curve, the first method was used as it can provide more reliable data. When neither method can be used to determine $T_m$, the sequence is deemed to be not analyzable (or NF for no fit).

**Circular dichroism spectroscopy (CD)**

*Structural studies*

All CD studies were performed on an Aviv model 410 which was recently updated to Avis 435 CD spectrophotometer. The CD is a powerful tool for investigating the secondary structure of biological molecules which are inherently chiral. CD takes advantage of the chirality and base stacking interactions of biomolecules. Here CD is mainly used as a finger-print method where different types of DNA secondary structure produce unique CD signatures. One-cm pathlength quartz cuvettes, 1.2 mL, were filled with annealed DNA in 100K2Mg buffer, targeting a concentration that would give an absorbance of 0.8-1 for optimal CD signal intensity. The CD wavelength scans were collected with the following parameters: 220-330 nm spectral window, 2 nm bandwidth, 1 nm step, 1 s averaging time, at 4 °C, and 5-7 scans. The raw CD signal measured in mdeg was converted to molar ellipticity using the formula shown below to allow comparison of samples regardless of concentration.

$$
\epsilon = \frac{\theta}{0.03298 \times c \times l}
$$

Where $\theta$ is CD signal in mdeg, $c$ is [DNA] in $\mu$M, and $l$ is the pathlength of the cell.
**Stability studies**

Melting experiments were conducted by monitoring the CD signal at a wavelength corresponding to a prominent feature (such as peak or troughs) of the CD signature of DNA or DNA-ligand complex. Melting was completed with the following parameters: 1-2 nm bandwidth, 1 °C temperature step, 0.33 temperature dead band, 0.083 min (5 s) equilibrium time, 15 - 20 s averaging time, and 4 - 95 °C temperature range. Both the heating and cooling curves were collected to determine reversibility of folding process. In addition, CD wavelength scans were taken both before and after melting study. CD melting data were analyzed in the same way as UV-vis data.

**Full wavelength melts**

Quartz cuvettes, 1.2 mL, were filled with 1Core or 2Core in 100K2Mg buffer, with a concentration giving 0.8-1.0 abs. The CD full wavelength melt were completed with the following parameters: 220-330 nm spectral window, 2 nm bandwidth, 1 nm step, 2 s averaging time, 20-80°C or 4-85 °C and 1 scan. The melting surface was corrected by subtracting the melting surface of buffer alone. Singular Value Decomposition (SVD) analysis was subsequently completed on the corrected melting surface of DNA. SVD decomposes the CD data matrix $M$ into the form:

$$ M = USV^T $$

where $U$ is the basis spectra, $S$ is a diagonal matrix comprised of the singular values (weight of each species), and $V$ is a matrix comprised of the amplitude vectors. Using a Globalworks program developed by the Chaires lab, the three-dimensional melting data was fit to various unfolding models (Figure 4) to determine kinetic rate constants thermal
and spectra of significant species. Autocorrelation of the U and V matrices was used to determine the number of significant species in each matrix.

![Diagram](image)

Figure 4. Models of the (A) single-state transition, (B) two-state transition, and (C) three-state transition, (D) parallel transition, and (E) parallel intermediate transition. F is folded species, I is intermediate species, and U is unfolded species.

The quality of fitting results for each model were evaluated based on statistics provided by the program, including $\chi^2$ values and the Akaike information criterion (AIC). Reasonable models with the lowest $\chi^2$ and AIC values were deemed best fits and likely model for the unfolding transition. The CD data matrix was reconstructed ($M^*$) using a new S matrix ($S^*$), which contained only the significant singular values. Taking the difference between the old ($M$) and the new ($M^*$) data matrices provided a surface of noise, showing only noise was removed from the $M^*$. 3D fitting was conducted on the $M^*$ for both 1Core and 2Core data.

**Thermal difference spectra (TDS)**

The TDS method is a wonderful complement to the CD in conducting a thorough biophysical characterization of nucleic acid secondary structures. A TDS signature is unique to various nucleic acid structures. Two scans are collected at a temperature much
lower than $T_m$ of the DNA (usually 4 °C) and at a temperature above the $T_m$ (usually 95 °C). The difference between the two spectra is taken and the resulting spectrum is normalized between 0 and 1 for the ease of comparison. As an ordered DNA structure is heated, disruption of base-stacking interactions leads to the changes in the UV absorbance, which is reflected in TDS signature. The laboratories of Prof. Mergny (University of Bordeaux, France) and Prof. Chaires (University of Louisville) collected TDS signatures of a variety of DNA structures to help identify the conformation of unknown DNA structures.28

**Polyacrylamide gel electrophoresis (PAGE)**

Native PAGE gels were prepared with 12% acrylamide, 1×TBE (Tris-borate-EDTA) and 3 mM MgCl₂. Gels were premigrated in 1× TBE with 3 mM MgCl₂ running buffer at 150 V for 30 minutes. Ten µL samples containing 3 µg of DNA annealed in 100K2Mg buffer were combined with 3 µL 30% w/v sucrose prior to loading to weigh them down. A mix of dT₁₅, dT₂₄, dT₃₀, and dT₆₀ polythymine sequences was used as a ladder. All samples and the ladder were loaded on the gel and run at 150 V for 130 minutes.

Denaturing PAGE gels were prepared with 12% acrylamide, 1× TBE. Gels were premigrated in 1× TBE running buffer at 300 V for 30 minutes. Ten µL samples containing 3 µg of DNA annealed in water were combined with 3 µL formamide prior to loading to weigh them down. A mix of dT₁₅, dT₂₄, dT₃₀, and dT₆₀ polythymine sequences was used as a ladder. All samples and the ladder were loaded on the gel and run at 300 V
for 120 minutes. All PAGE gels (native and denaturing) were visualized using Stains-All and photographed using conventional camera.

**Fluorescent spectroscopy studies (FL)**

All fluorescent experiments were conducted using the following parameters: excitation wavelength of 305 nm for 2AP, emission wavelength range of 325-500 nm, excitation and emission slit widths of 10 nm and 5 nm, respectively, gain of 250, and temperature of 4 °C. Fluorescence quartz cuvettes (3 mL) were filled to 2.0 mL with freshly filtered 100K2Mg buffer and scanned after 5 min of thermal equilibration. Part of buffer was removed, and an equal volume of DNA was added to the cuvette to keep a constant volume of 2.0 mL and targeting a final [DNA] corresponding to 0.05 absorbance on UV-vis at 4 °C. The final [DNA] for 1Core+2 and 2Core+2 2AP samples were 0.2 and 0.1 uM, respectively. After thorough mixing and 5 min of thermal equilibration, fluorescence data were collected. Baseline contribution to fluorescence was subtracted for each sample.

**Small molecular microarray (SMM)**

SMMs were generated by spatially arraying and covalently linking a library of small molecules to a glass surface as previously described by Bradner et. al. The Schneekloth lab provided appropriate microarrays that included duplicated small molecule libraries, dyes, and control samples. This experiment was run in two duplicates (one pair in Phosphate Saline Buffer (PBS) and one in 100K2Mg). Dust was removed from ligand-covered glass plates using an air blower before the incubating plates with a lRNA
(transfer ribonucleic acid) solution for 30 min on a nutator. After pouring off the 1RNA solution, the glass plates were incubated in sample solutions containing and 0.005% Tween-20, 400 µM 1RNA, 90 µM 1Core+2 Cy5 in either PBS or 100K2Mg buffers for 2 hrs. Three washes with the respective buffers (1x PBS or 1x 100K2Mg) and a final wash with ddH2O followed incubation with sample solutions.

Microarrays were dried by centriguation at 3000 rpm for 3 min then imaged immediately using the InnoScan 1100 AL which scanned 5 µm/pixel at excitation wavelengths 488, 532, and 635 nm. The latter wavelength is the most important in this experiment as it correlates to the Cy5 attached to 1Core+2; fluorescence of microarray at this wavelength (visually seen as red) indicate compound binding. Flourescenee quantification of the SMM scanned images were completed as previosuly descibred by the Shneekloth lab.31 We focused on three criterions determined a hit: 1)

\[ Z_{score_{DNA\, incubated}} > 3 \]

2) \[ \frac{Z_{score_{DNA\, incubated}} - Z_{score_{control}}}{Z_{score_{control}}} > 3 \]

3) well-defined spot morphologies consistent with compound binding (amongst duplicates). Promising hits from this study (Table 3) were individually studies using UV and CD spectroscopies to characterize ligand-DNA interactions.
Chapter 3: DNA repeats linked to replication stress

Biophysical screen of DNA repeats

Introduction

Biological assays and bioinformatics completed by Brown’s lab (University of Pennsylvania) identified a variety of short tandem DNA repeats linked to replication stress.\(^ {11}\) Brown’s lab hypothesized these problematic repeats fold into stable secondary structures capable of physically blocking replicative polymerase. We conducted thermal stability and structural studies on these tandem DNA repeats via CD and UV to test this hypothesis. This section is dedicated to a more thorough look at the biophysical characterization included in recently published work.\(^ {11}\) All repeats investigated were ~90 nt long to simplify their comparison. Repeats examined in this work are presented in Table 4.

Table 4. DNA repeats investigated in this work. Each repeat was analyzed via RPA ChIP Seq which reports fold enrichment at RPA bound sites compared to background genomic levels.\(^ {11,32}\)

<table>
<thead>
<tr>
<th>Monomer</th>
<th>Repeat #</th>
<th># nt</th>
<th>Fold over Input (RPA ChIP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAGG</td>
<td>22</td>
<td>88</td>
<td>13.7 ± 6.2</td>
</tr>
<tr>
<td>CACAG</td>
<td>18</td>
<td>90</td>
<td>22.2 ± 0.6</td>
</tr>
<tr>
<td>CACAGG</td>
<td>15</td>
<td>90</td>
<td>5.0 ± 0.5</td>
</tr>
<tr>
<td>CAGAGG</td>
<td>15</td>
<td>90</td>
<td>23.2 ± 3.8</td>
</tr>
<tr>
<td>CAGAGT</td>
<td>15</td>
<td>90</td>
<td>298.3 ± 125.2*</td>
</tr>
<tr>
<td>CAGAGGG</td>
<td>11</td>
<td>88</td>
<td>NM</td>
</tr>
<tr>
<td>CAGAGGGG</td>
<td>12</td>
<td>84</td>
<td>NM</td>
</tr>
</tbody>
</table>

NM = Not measured; repeats were not as enriched at fork-stalling sites compared to other repeats such as CAGAGG.

* Artifact only occurs at one input site (located within CAGAGG repeat), causing the fold over ratio to increase dramatically relative to other repeats.
Results

CAGG/CCTG repeats

The tetranucleotide repeat CAGG is the shortest repeat of this group and is linked to an expansion disease called Mytonic dystrophy type II which is the most common form of muscular dystrophy in adults. The repeat (CAGG)$_{22}$ had a CD signature with a peak at 261 nm and a trough at 284 nm (Figure 5). The CD signature of CAGG is similar to that of Z-DNA which has a trough at 290 nm and a peak at ~260 nm. The complementary oligonucleotide (CCTG)$_{22}$ had a CD signature with a peak at 284 nm and trough at 255 nm. The TDS of CAGG showed a peak and a shoulder at 269 and 256 nm, respectively (Figure 6). CCTG had a different TDS with a peak at 277 nm and shoulder at 241 nm. The Tm of CAGG is 60.6 ± 0.4 °C, while CCTG is unstable with a Tm < 4 °C (Figure 7). The data overall clearly suggest that CAGG and CCTG are structurally distinct and that only CAGG is stably folded. It was expected that CAGG could fold into a GQ structure because of its G-rich nature, but none of the biophysical data support this topology. Strikingly, the CD signatures of CAGG and CCTG are nearly mirror images of each other, which could reflect an enantiomeric relationship between the folds similar to that shown for Z-DNA and B-DNA.

CACAG/CTGTG repeats

CACAG was the second most intrinsically difficult-to-replicate repeat. This pentanucleotide repeat has been shown to play a role in the gene expression of Pastuerella haemolytica. CACAG repeat had a CD signature with a peak, shoulder, and trough at 280, 289, and 253 nm, respectively (Figure 5). The CD signature of CTGTG
has a two peaks (243 and 289 nm) and one trough (263 nm). CACAG and CTGTG have TDS peaks located at 265 and 276 nm, respectively (Figure 6). The T_m values of both repeats are identical within the experimental error, 41.2 °C but they differ in the value of enthalpy, Figure 7. Similar to the CAGG study, the data clearly show CACAG and CTGTG have different folds. Interestingly, CTGTG has the same stability of its complement yet is much less folded according to the intensity of its CD spectra. A shorter analog (9 repeats instead of 18) of CACAG is strongly linked to CTCF (zinc-finger) binding so we examined this sequence and its complement (CTGTG)_9 (Figure 5-7). The results for the 9-repeat and 18-repeat oligos were the rather similar suggesting that the characteristic structure of CACAG repeats is not strongly dependent on the repeat length or that the final structure is composed of small folded core units and 9 repeats is sufficient to form at least one of those units.

CACAG was one of the few difficult-to-replicate repeats that usually contained interruptions between repeats. Thus, we have also conducted biophysical studies on CACAG with intermittent and endogenous interruptions in the repeat sequence termed Endo-CACAG. The CD signatures and TDS data for both Endo-CACAG and its complement matched that of the native oligos (Figure 5 and 6). However, the thermal stability was effected where Endo-CACAG and Endo-CTGTG had T_m values of 24.7 ± 0.3 °C and 37.6 ± 0.3 °C, respectively. Interestingly, the T_m of CACAG was effected by the repeat interruptions more than its complement which suggest how crucial tandem repeats are for the stability of the CACAG structure.
CAGAGG/CCTCTG repeats

CAGAGG is the repeat most highly associated with replication stress. We studied (CAGAGG)$_{15}$ and found it had a CD signature with a peak at 261 nm, two shoulders at 276 and 291 nm, and trough at 241 nm (Figure 5). (CCTCTG)$_{15}$ has a CD signature with a peak at 282 nm and trough at 253 nm, which is reminiscent of the spectroscopic data for CCTG and CTGTG. The TDS spectra of CAGAGG and CCTCTG show peaks at 260 nm and 279 nm, respectively (Figure 6). In CD melting experiments, CCTCTG demonstrated no clear melting transition. CAGAGG had a clear melting transition with a $T_m$ of 55.6 ± 0.9 °C (Figure 8). Overall, the data shows CAGAGG and CCTCTG have distinct folds. CCTCTG is noticeably less folded than CAGAGG and this coincides with its lack of stability. The high degree of fold and moderately high stability of CAGAGG may explain why it was found in biological studies to be the most intrinsically difficult-to-replicate.\textsuperscript{11}

CAGAGAGG/CCTCTCTG repeats

The CD signature of CAGAGAGG has one dominant peak at 262 nm, a trough at 242 nm, and a shoulder at 289 nm (synonymous to CAGAGG). CCTCTCTG has a peak at 279 nm and a trough at 251 nm (Figure 5). The TDS for CAGAGAGG shows a peak at 260 nm, while TDS for CCTCTCTG shows at a peak at 280 nm (Figure 6). CCTCTCTG is unstable ($T_m$ of ~11 °C) and CAGAGAGG is moderately stable with a $T_m$ of 46.4 ± 0.6 °C (Figure 7). The data shows a clear difference between the structures where CAGAGAGG is both more folded and more stable.
**CACAGG/CCTGTG repeats**

(CACAGG)$_{15}$ has a CD signature with 277 nm peak a 243 nm trough and a 289 nm shoulder (Figure 5). (CCTGTG)$_{15}$ shows a CD signature with two peaks at 261 and 285 nm. The TDS of CACAGG comprised a 266 nm peak and a 259 nm shoulder; CCTGTG had a single TDS peak at 278 nm (Figure 6). CACAGG and CCTGTG have melting temperature of 35.8 ± 1.2 °C and 51.4 ± 0.8 °C, respectively (Figure 7). Unlike the other studied pyrimidine rich strands, (CCTGTG) is more stable than the purine rich strand (CACAGG). In addition, CCTGTG has a sharper and more well defined melting transition than its complement reflected in its high ΔH of 283 ± 9 kJ/mol (vs 66 ± 7 kJ/mol for CACAGG). This data provides further support that CCTGTG is more structured than CACAGG.

**CAGAGT/ACTCTG repeat**

The next difficult-to-replicate repeat is (CAGAGT)$_{15}$. It has low-amplitude CD signature with two peaks at 249 and 280 nm and a 266 nm trough (Figure 5). Its compliment, ACTCTG, has a signature with a peak and trough at 280 and 241 nm, respectively. TDS of CAGAGT shows a 263 nm peak, different from the TDS of ACTCTG with a 269 nm peak (Figure 6). The thermal stabilities of the two sequences also differ: CAGAGT and ACTCTG have T$_m$ values of 41.3 ± 0.6 °C and 29.9 ± 0.3 °C, respectively (see Figure 7). CAGAGT is more stable and less folded, as compared to ACTCTG. The CAGAGT repeats likely do not play any crucial role in the replicative stress due to their lack of structure and stability.
**CAGAGGG/CCCTCTG repeat**

CAGAGGG is different from the other oligos because the presence of multiple GGG tracts within this sequence is expected to promote GQ formation. The CD signature of CAGAGGG is consistent with this hypothesis as it shows a peak at 294 nm with a trough at 257 nm, indicative of a mixed hybrid GQ (Figure 5). TDS of this oligo includes a trough at 298 nm, signifying G:G Hoogsteen base-pairing (Figure 6).

CAGAGGG is the most stable oligo with a Tm of 72.0 ± 1.1 °C; its high stability is another marker of GQ topology (Figure 7). CCCTCTG has CD signature and TDS peaks both at 281 nm—not indicative of any known DNA structure (Figure 5 and 6). In addition, it has a poorly defined melting transition and low stability (~15 °C).

CAGAGGG has spectroscopic signatures which strongly suggests GQ formation, while CCCTCTG data resembles those of other the pyrimidine rich DNA.

![Figure 5. CD scans of the DNA repeats studied collected in 100K2Mg buffer at 4 °C.](image)
Figure 6. TDS scans of DNA repeats studied. DNA. The data were collected at 90 °C and at 4 °C, subtracted and normalized

Figure 7. UV-vis melting curves of the DNA repeats studied. Signal was monitored at 260 nm except for (CAGAGGG)\textsubscript{12} for which the signal was monitored at 295 nm (wavelength which signifies the presence G-quadruplex). All samples were identical to those used for CD studies.

Discussion

We thoroughly investigated a variety of difficult-to-replicate sequences for their structure and stability. One crucial observation is all the studied sequences adhere to the following structural framework: GCN\textsubscript{2-6}GCN\textsubscript{2-6}GCN\textsubscript{2-6}GCN\textsubscript{2-6} where the minimum number of GC pairs required for the formation of a stable structure is four and N\textsubscript{2-6} represents loops with lengths of 2-6 nt. From here on, we will take advantage of this
framework and discuss the oligos as a set of loop variants (where N varies in both composition and length) to better contextualize the data.

The oligos studied here can be partitioned into purine rich and pyrimidine rich groups. Every purine rich oligo showed folds distinct from its complement. While the pyrimidine rich oligos share one particular fold (characterized by a CD signature and TDS with dominant peaks at ~280 nm), the purine rich oligos were more structurally diverse. Besides CAGAGGG, these folds are both unusual and uncharacterized to-date.

Because the size for all of the oligos are about the same (88-92 nt), and each have at least five repeat units required to form stable structures, loops are likely the most crucial parameter that dictates folding and stability of the structures. This is especially important for oligos where loops are comprised of (AG)ₙ units such as CAGG, CAGAGG, CAGAGAGG. Based on the data presented here, these oligos are typically characterized by well-folded and stable structures relative to oligos with non-AG loops (CACAG and CACAGG). These observations suggest AG-based loops accommodate better folding which agrees with findings that show purine rich DNA can fold into unusual structures. While it may seem that CAGAGT is an exception to this framework, it is not. This oligo can be visualized as a CAGAG repeats connected by a T nucleotide; this single nucleotide insertion changes the structural framework from (GCN₂₆)₄ to (GTCN₂₆)₄. This change disrupts the stacking between contiguous GC base pair units which are probably crucial to the structural stability. T-T dimerization is another possibility, albeit it will likely provide minimal contribution to the fold. The biophysical data for CAGAGT suggests that single T interruptions result in lower CD signal amplitude (hence lower folding) and lower stability. CAGAGGG operates outside of the
proposed structural framework and clearly folds into a GQ. This sequence should be seen as GGGCAGAGGGCAGAGGG where CAGA units are most likely loops.

The purine rich oligos are more structurally diverse compared to the pyrimidine rich oligos. The exception to trend is CCTGTG which seems to be not only stable but quite well-folded with a CD signature distinct from those of other members in the CTG group. ACTCTG (similar to its complement) can be thought of as CTGTG repeats with interrupting adenine residues, which likely destabilizes the structure as it disrupts the GC bp interaction.

In sum, most of the oligos in this study seem to operate under a single structural framework that helps explain our findings. Our biophysical characterization revealed CAGAGG is the most-folded of the purine rich strands. The biological work completed by the Brown lab (Table 4) showed this repeat is the most associated with replication stress in mouse cell lines. Our data shows structured DNA (such as CAGAGG) is more highly linked to replicative stress. These results support the Brown lab’s hypothesis on replication stress. To further understand the structural features of CAGAGG that make it potentially problematic for the replication process, we focus our biophysical efforts solely on this repeat.

**Building the model of (CAGAGG)_n**

The Yatsunyk lab (conducted by Liliya Yatsunyk and Jessica Chen ’17) carried out a variety of biophysical tests to understand the fold of CAGAGG. 1D ¹H NMR studies showed the presence of imino and amino protons suggestive of GG and GC bp. CD structural and stability studies on CAGAGG with varying repeat lengths showed at least 5 tandem repeats were needed for stability and structure. Thermal melting studies
showed the fold and stability of CAGAGG were independent of DNA concentration indicating that structures formed are monomolecular. In conjunction, PAGE experiments supported monomolecular fold and indicated the homogeneity of the samples in most cases (in some cases one can observe the presence of weak dimer). The structure formed by CAGAGG repeat is stabilized by 100 mM of monovalent ions (Li\(^+\), Na\(^+\), and K\(^+\)) independent of their identity, and greatly stabilized by Mg\(^{2+}\) - where 1 mM of Mg\(^{2+}\) produces stabilizing effect similar to that of 100 mM K\(^+\). The stability of CAGAGG is enhanced by higher ion concentrations and lower pH. After this extensive biophysical characterization, many details of the structure still remained unclear.

Due to the G-richness of the sequence (guanine content of 50\%), the Yatsunyk lab hypothesized that CAGAGG repeats could adopt a GQ fold. The 264 nm peak in the CD spectra of CAGAGG resembles a parallel GQ but the interactions of CAGAGG with a highly selective and stabilizing parallel GQ binder N-methyl mesoporphyrin IX (NMM) were negligible (completed by Liliya Yatsunyk and Jessica Chen ‘17).\(^{40}\) GQ are known to be strongly stabilized by K\(^+\), to a lesser extent by Na\(^+\) and usually not stabilized by Li\(^+\).\(^{19}\) As mentioned previously, CAGAGG does not discriminate between K\(^+\), Na\(^+\), and Li\(^+\). In addition, CAGAGG is strongly stabilized by Mg\(^{2+}\) which is uncharacteristic for GQ and more characteristic of duplex DNA. These results ultimately do not support the GQ-forming hypothesis.

Next, we look towards the purine rich loops of CAGAGG to understand the structure. Depending on the conditions, purine rich DNA can adopt duplex, triplex, and even quadruplex topologies.\(^{38,41,42}\) Kypr et. al demonstrated formation and stabilization of d\((AG)\)_n homoduplexes with aqueous ethanol and cations such as Li\(^+\), Na\(^+\), K\(^+\).\(^{41,43}\) The
CD spectra of the parallel homoduplex d(AG)_n contains a peak at 260 nm and troughs at 240 and 290 nm. Interestingly, the single-stranded conformer of d(AG)_n which is promoted at low pH (by protonation of adenine) shares the CD spectra as the parallel homoduplex. The CD spectra of CAGAGG contains two shoulders at 280 nm and 290 nm, different from either of the d(AG)_n conformers (Figure 5). Kypr et. al reported stabilization the d(AG)_n antiparallel homoduplex with Zn^{2+} enhancing the 250 nm trough observed in its CD spectra. Unlike (AG)_n, the structure of CAGAGG unfolds in the presence Zn^{2+} as suggested by CD (Appendix, Figures S1 and S2). The data shows d(AG)_n and (CAGAGG)_n fold into distinct structures despite shared CD spectral characteristics. Previous findings have shown AG loops within canonical duplexes cause structural distortions that may explain the unusual CD spectra of CAGAGG.38,44

One plausible structure for CAGAGG is the stem-loop hairpin. The stem section of this fold would be comprised of canonical (GC) and non-canonical (AG) base pairs. The trough-peak-two shoulder CD spectra of CAGAGG does not resemble the 250 nm trough and 270 nm peak CD spectra of a typical right-handed duplex.34 The CAGAGG CD signature does, however, resemble a hairpin (particularly CAG here) whose signature includes a 240 nm trough, 260 nm peak, and 280 nm shoulder.45 Though the results rule out a classic WC duplex, whether CAGAGG repeats fold into a stem-loop hairpin remains unknown.

The CD signature of CAGAGG was compared with those of other well-known DNA structures using a program developed by the Chaires laboratory (U of Louisville).46 This comparative study shows CAGAGG is only structurally similar to GGGAGCG or GCGAGGG repeats as shown in Figure 8. The structure of these repeats (named
VK1/VK2 or VK34) was solved using NMR in the Plavec laboratory.⁴⁷,⁴⁸ According to the structures, these repeats form a multimeric and antiparallel tetrastranded structure, comprised of G:G Hoogsteen base pairs and non-canonical AG base pairs sandwiched by GC WC base pairs. The VK1 and VK2 have a characteristic cavity within the core of the structures. Meanwhile VK34 contains GCGC, GGGG, and AGAG tetrads that fill the core of this fold.

![Figure 8](image)

**Figure 8.** CD spectra (normalized by sequences length) of CAGAGG (1Core+2, introduced in next paragraph) with VK1 and VK2.

At this point, a new structural hypothesis is necessary. Our previous studies demonstrated the need for least 5 CAGAGG repeats for the fold. These 5 repeats allow four GC base pairs, which we hypothesized are essential for stability. Minimizing the number of unnecessary nucleotides within the terminal repeats will provide the most stable structure. Barrett Powell ’18 discovered two nucleotide terminal tails provide the most stable structure (Barrett Powell, Honors thesis). This optimized sequence 5’-AGG(CAGAGG)₃CAG-3’ is termed 1Core+2; the structure shown has one core and two nucleotides in each tail. Based on our previous findings, we proposed a model for
1Core+2 (Figure 9A). 1Core+2 is a monomolecular tetrastranded, antiparallel structure with a core made up of two GCGC tetrads #-stacked on each other. The tetrad could be constructed from two WC GC base pairs connected by bifurcated hydrogen bonding like in Leonard’s structure or two adjacent GC base pairs separated by a cavity as in Plavec’s VK1 or VK2 structures.47–49 The core is connected by three lateral AGAG loops which may be involved in structural interactions that help stabilize the fold. The rest of the work presented here will focus on testing and refining our proposed model using biophysical methods.

**Figure 9.** (A) Two possible models for 1Core+2. Hypothesized model based on an antiparallel, tetrahelical fold composed of GCGC tetrads that are connected by three AGAG loops. (B) Stem-loop hairpin that includes non-canonical base pairing in internal AGAG loop regions that are sandwiched by GC base pairs. In this structure, there is only one primary loop region.
Chapter 4: Ligand screen of (CAGAGG)_n with DNA binders

Interactions of 1Core with small molecule ligands

Introduction

Ligand screening presented in this chapter had two goals. The first one concerns crystallographic studies as one of the best ways to elucidate the secondary structure of CAGAGG. Another student in the lab, Barrett Powell ’18, dedicated to this task to his thesis work. In order to aid in crystallization trials, I screened a variety of known small molecules nucleic acid binders to find one that binds tightly to DNA while stabilizing and maintaining its secondary structure. Such a ligand might lead to a successful DNA crystallization and provide high resolution X-ray diffraction data necessary to solve the structure. Our second goal in performing the ligand screen was to shed light on the secondary structure of CAGAGG by identifying the molecules or class of molecules, which interact with this DNA structure. For instance, if we learn that CAGAGG interacts with exclusively GQ binders, then it is plausible to conclude CAGAGG has GQ character.

We started the ligand screen study using 1Core, the smallest structural representative for CAGAGG repeat. Overall, 23 ligands were studied; for most of them experiments were repeated just once. Interesting ligands were screened at least twice. The screening included the following three experiments: 1) UV-Vis titrations for all colored ligands to determine the binding interaction via the shift in the main UV-vis peak of a
ligand (in the case of a porphyrin, it is the Soret band), and when possible, binding constants were determined; 2) CD annealing studies demonstrate whether the presence of a ligand affects the secondary structure of 1Core; and 3) CD melting studies of annealed samples reveal the ligand-induced stabilization of the 1Core structure.

Prior to this work, the interactions of NMM with CA10 and CA15 (10 and 15 repeats of CAGAGG) were studied by Liliya Yatsunyk and Jessica Chen ‘17. UV-vis titration, CD scans, and CD melting studies were conducted and all experiments showed that NMM does not interact with either CA10 or CA15 (unpublished work). Furthermore, studies of the interactions of TMPyP4, TMPyP3, TMPyP2, RHPS4, and Braco19 with 1Core were completed by Barrett Powell ‘18.

Results

According to results presented in Table S1, the ligands can be classified as non-interactive, structurally disruptive, or stabilizing (and not structurally disruptive) as shown in Table 5. In the following section, we will present the data for one representative ligand from each class; the rest of the data can be found in the Appendix, Figures S3-S15.

Table 5. List of tested ligands categorized according to their interactions with 1Core.

<table>
<thead>
<tr>
<th>Non-interactive</th>
<th>Structurally disruptive</th>
<th>Stabilizing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethidium Bromide</td>
<td>Phen-DC3</td>
<td>Methylene Blue</td>
</tr>
<tr>
<td>Berberine Chloride</td>
<td>CuTMPyP4</td>
<td>Coralyn Chloride</td>
</tr>
<tr>
<td>Berenil</td>
<td>ZnTMPyP4</td>
<td>RHPS4</td>
</tr>
<tr>
<td>Commissie Brilliant Blue</td>
<td>Crystal Violet</td>
<td></td>
</tr>
<tr>
<td>Hoechst 33258</td>
<td>TMPyP4</td>
<td></td>
</tr>
<tr>
<td>Malachite Green</td>
<td>TMPyP3</td>
<td></td>
</tr>
</tbody>
</table>
Non-interactive Ligands: Ethidium Bromide (EB)

The non-interactive ligands from this study are NMM, TMPyP2, PDS, Braco-19, Neutral Red, Methyl Green, Malachite Green, Coomasie Brilliant Blue, Hoechst 33258, Berberine Chloride, Pentamidine Isethionate Salt, and EB. EB is known as a duplex binder that can engage in all three modes of binding: intercalation, minor, and major groove binding. The structure of EB and the results of the key experiments are shown in Figure 10. CD scans shows minimal change in the shape and intensity of CD signal of 1Core. Consistent with this data, EB does not affect the stability of 1Core (ΔT = -0.4 °C at 2 eq. and 0.5 °C at 4 eq, all within the experimental error). The Nicolini lab showed that when EB intercalates into DNA one can observe an induced signal at 308 nm. The lack of 308 nm spectral features in the CD signature of 1Core with EB provides support of no binding interaction (through intercalation specifically).
Figure 10. (A) CD scans and (B) normalized CD melting curves of 4.9 μM 1Core with 2 and 4 eq. of EB. (C) Structure of EB.

Structurally Disruptive Ligands: Phen-DC3

Ligands deemed disruptive are TMPyP4, TMPyP3, CuTMPyP4, Phen-DC3, and Crystal Violet. Phen-DC3 and ZnTMPyP4 will be used as representatives of this group. Phen-DC3, a derivative of bisquinolinium, is a well-known and widely studied selective GQ stabilizer.\textsuperscript{52,53} Its exceptional binding capabilities with GQ are due to its crescent shape shown in Figure 11A (encouraged by intramolecular H-bonding), which allows is full overlap with the terminal G-tetrad of a GQ. CD scans of 1Core with Phen-DC3 (Figure 11B) demonstrate dramatic alteration of the CD signature with two new bands emerging at \%250 nm and 290 nm indicating change in the secondary structure of 1Core. This signature is maintained, although the relatively intensity of the two bands change upon further increase of Phen-DC3 concentration. The dramatic change in the CD signature clearly shows that Phen-DC3 interacts with 1Core.
Destabilizing (and not structurally disruptive) ligands: ZnTMPyP4

ZnTMPyP4, a metallated version of the widely studied TMPyP4 duplex and GQ binder. Importantly, Zn(II) is coordinated by a water molecule in the fifth position, which is assumed to prevent intercalation of this ligand and thus confer its greater selectivity for GQ. In reality, such improvement of the selectivity was not observed in GQ binding studies. ZnTMPyP4 was shown to effectively interact with human telomeric DNA and induce the formation of parallel GQ under otherwise unfavorable condition. Unlike most other ligands that were structurally disruptive, ZnTMPyP4 binds and unfolds 1Core without changing its structure. A CD scan of 1Core in the presence of 2 and 4 eq. of ZnTMPyP4 showed overall decrease in CD signal intensity but preservation of the spectroscopic signature (Figure 12A).

CD melt showed significant destabilization of 1Core by -15.3 °C and -18.5 °C at 2 and 4 eq., respectively (Figure 12B). This data corroborates studies of CA10.
(CAGAGG<sub>10</sub>) with Zn<sup>2+</sup> ions which showed decrease in fold and stability proportional to the concentration of the ions (see Appendix, Figure S1). Previous findings showed Zn<sup>2+</sup> ions induce unusual structural changes to d(AG)<sub>n</sub>.<sup>34</sup> Coordination of Zn<sup>2+</sup> to the AG section of CAGAGG may explain the interactions between Zn<sup>2+</sup> ions and ZnTMPyP4 with 1Core. The metal ion of ZnTMPyP4 is suggested to bind to exposed N<sub>7</sub> sites on guanines such as those in Z-DNA.<sup>55</sup> These guanine rich repeats may be susceptible to such interactions with Zn<sup>2+</sup>, causing structural change. The porphyrin scaffold of ZnTMPyP4 may cause more steric hindrance upon interactions with 1Core, furthering destabilizing the structure. Su et. al demonstrated no structural transition or change in fold of B-DNA with the addition of ZnTMPyP4.<sup>56</sup>

![Figure 12](image)

**Figure 12.** (A) CD scans and (B) normalized CD melting curves of 4.21 µM 1Core with 2 and 4 eq. of ZnTMPyP4. (C) Structure of ZnTMPyP4

**Stabilizing Ligands: Methylene Blue (MB)**

Stabilizing ligands are RHSP4, Coralyne Chloride, and MB. MB is known as a duplex binder and is commonly used as a biological dye. CD scans of 1Core with MB (Figure 13A) showed decrease in the CD signal at 242, 265, 290, and 297 nm, main features of the 1Core CD signature. CD melting demonstrated the stabilization of 6.8 °C and 9.7 °C at 2 and 4 eq., respectively (Figure 13B). This data is noteworthy as MB
provided the highest stabilization compared to the rest of the ligands studied thus far. The changes in both CD signature amplitude and stability suggest the MB interacts with ICore. These results corroborate previous findings that have shown intercalated binding of MB with GC rich DNA via linear and circular dichroism spectroscopies and computational studies. The intercalation may explain the decrease in the CD spectrum as the MB interrupts contiguous base pairing in ICore which contributes to the CD signal amplitude.

![Figure 13](image)

**Figure 13.** (A) CD scans and (B) normalized CD melting curves of 5 μM ICore with 2 and 4 eq. of MB. (C) Structure of MB.

**Discussion**

Most of the GQ binders change the secondary structure of ICore while most duplex binders are non-interactive. These results suggest ICore lacks duplex character yet has G-quadruplex character. Prominent GQ features include G:G hoogsteen bonding and structure made of four strands. Our previous data shows ICore contains no G:G hoogsteen bonding as indicated in the lack of 295 nm trough within the TDS (discussed further in Chapter 5). The results suggest that helicity based on four strands is the primary structural similarity between GQ and ICore. This data corroborates our proposed
four-stranded model and similarities between the CD signature of 1Core and those of tetrahelical DNA VK1 and VK2.

Ligands that favorably interact with 1Core do not come from one class of DNA binders, suggesting 1Core is a fold that does not fit well into any of the well-known structural classes. In addition, these stabilizing ligands are structurally distinct so there are no clear motifs that enhance the binding preference for 1Core. While the triplex binders Berenil and Carolyne Chloride provide minimal stabilization to 1Core, the argument for 1Core being triplex is weakened by the lack of stabilization (and in fact destabilization) of CA10 caused by Zn$^{2+}$ as expected for triplexes.$^{59}$

MB was the only attractive co-crystallization agent from this screen. Despite the stability it provides the system, it did not provide high-quality crystals as a crystallization reagent needed for structure determination as shown by Barrett Powell ‘18. Among the stabilizing ligands from this study, there is no clear structural moiety that explains the stabilizing ability of these ligands.

While ZnTMPyP4 was not a hit according to the criteria of this study, it provides interesting results as it destabilizes the structure of 1Core without changing the fold. These results corroborate the ion dependence studies (Appendix, Figures S1 and S2) that show Zn$^{2+}$ ions alone destabilize the fold of CA10 and 1Core+2.

**Conclusion**

Most of the studied GQ binders changed the fold of 1Core into sometime unusual structure. The data suggest 1Core resembles a GQ more than a classic duplex. This corroborates our proposed model of 1Core, which is based on a tetraplex. Generally, 1Core is not interactive (according to our criteria) with well-known nucleic acid binders.
1Core demonstrates an ability to engage in various binding modes: intercalation, minor groove, and a major groove binding. This study only includes 23 ligands so a larger screen may yield more fruitful results and candidates for crystallization studies. This study would also benefit from further characterization of the binding kinetics of GQ ligands to 1Core.

This work presents one of the first ligands screens for a disease associated repeat besides telomeric DNA and other GQ forming oligos. Using a different approach, the Nakatani group synthesized a ligand naphthyridine—azaquinolone to stabilize the (CAG)$_n$ hairpin.60 Despite efforts such as those of the Nakatani group, there remain only few known binders to repeat DNA structures. This study provides an approach towards characterizing DNA—ligand interactions in the absence of structural knowledge of the DNA.

**Small molecule microarray screen of 1Core+2**

**Introduction**

In pursuit of the structure of (CAGAGG)$_n$, the interactions of 1Core with 23 known nucleic acid binders were tested. To expand the chemical space of this ligand screen, we turn to a small molecule microarray (SMM) designed by the Schneekloth lab (Figure 14). The library includes thousands of compounds that are chemically diverse, drug-like, and commercially available. This method was used to successfully identify small molecules that bind selectively GQ and HIV related RNA.31,61 This method allowed us to expand our screen from 23 to 7065 ligands (7042 new ligands). Through statistical
and visual analyses of the SMM data, a list of 23 hits was created (Table 3). The interactions of 1Core+2 with the twelve commercially available hits from this list were tested using CD spectroscopy.

![Figure 14](image)

**Figure 14.** Schematic of SMM of 1Core+2 Cy5 with untreated and treated slides. Red spots on the slide indicate 1Core+2 Cy5 binding.

### Results

Before starting the SMM, a fluorescently labeled analog of 1Core+2 was required to allow fluorescent imaging. A cyanine dye (specifically Cy5) suited for fluorescent imaging at 633 nm was added to the 5’ tail of 1Core+2 termed 1Core+2 Cy5. The chromophore was added to one of the terminal tails as they were shown to contribute little to the fold and would cause minimal structural perturbation (discussed later in Chapter 5). Both CD scans and TDS of the 1Core+2 and its labeled analog show nearly superimposable spectroscopic signatures as shown in **Figure 15A** and **15B**. CD melting
studies provided stabilities of 49.0 ± 0.5 °C and 47.7 ± 0.5 °C for 1Core+2 and 1Core+2 Cy5, respectively (Figure 15C). Expectedly, the data suggest adding the dye to the terminal tail did not influence the fold or stability to a significant degree. The fluorescent analog of 1Core+2 was subsequently used for SMM studies.

**Figure 15.** (A) CD scans, (B) normalized TDS, (C) normalized CD melting curves of 3.25 #M 1Core+2 and 3.33 #M 1Core+2 Cy5.

The fluorescent images for the interactions of SMM each compound with 1Core were analyzed as shown in **Figure 16.** All fluorescent images are shown in **Appendix, Figure S16.** These images show spot morphology and the degree of the fluorescent response, which are good measures of data quality. Some of the compounds included in this library are auto fluorescent (shown as green fluorescence). Visual analysis allowed elimination of these auto fluorescent molecules from the group of candidates that show good statistics (high Z-score in presence of 1Core+2 Cy5) despite lacking interactions with 1Core+2 Cy5.
Figure 16. Representative fluorescent images for SMM compounds. Red spots likely indicate 1Core+2 Cy5 binding. G702-4315 represents a molecule that showed no fluorescent response and hence no interaction with 1Core+2 Cy5 in either condition (PBS or 100K2Mg). G818-0013 represents a molecule that shows a fluorescent response irrespective of 1Core+2 Cy5 binding (shown in the buffer case). K405-3169 represents a molecule that shows a fluorescent response in the presence of 1Core+2 Cy5 in 100K2Mg (and not in PBS). Similar data has been retrieved where a molecule shows a fluorescent response in PBS and not 100K2Mg or both PBS and 100K2Mg (not shown). All three of these latter cases suggest 1Core+2 Cy5 binding with the compound of interest.

For each compound, a Z-score was calculated based on the increase in fluorescence (monitored at 648 nm) due to 1Core+2 Cy5 binding (Appendix, Figure S17). A compound was deemed a hit if its respective Z-score was higher than 3. The Z-scores for all 7042 molecules are shown in Figure 17. This threshold eliminated most of the compounds from the list of possible hits, leaving 33 compounds as reliable candidates. After a closer look at the visual and statistical data for each compound from this list of 33 candidates, 10 compounds were eliminated based on their Z-score in buffer alone being at or above the threshold and the quality of raw data.
Figure 17. Z-scores of all 7042 compounds for treated versus untreated microarrays in (A) PBS and (B) 100K2Mg. (C) Z-scores for all compounds for both treated cases (PBS and 100K2Mg). (D) Representative examples of unqualified (E946-0328) and qualified (E919-2270) compounds according to the Z-scores. Because the Z-score of E946-0328 in buffer alone is higher than 3, it is not a good candidate for a hit. The Z-score threshold of 3 standard deviations is shown with dashed lines on the figures.

The interactions of 1Core+2 with 8 of the commercially available hits from the final list of compounds were tested using CD. These ligands were chosen based on meeting the statistical criteria for a hit in both PBS and 100K2Mg (discussed in Chapter 2). CD signatures of 1Core+2 alone and with all 8 ligands tested were rather similar as shown in Figure 18A. CD melting analysis of 1Core+2 with the ligands show similar data with nearly superimimposable melting curves (Figure 18B).
Figure 18. CD scans (A) and CD melting (B) of 3.65 #M 1Core+2 with 2 eq. of each ligand, which was dissolved in DMSO.

**Discussion**

The SMM results suggest there are a variety of compounds that interact with 1Core+2 to a statistically significant extent. Monitoring these binding interactions via CD revealed no significant change in CD signature or stability for 1Core+2. These results show the tested compounds fit into the class of non-interactive ligands (according to the previous ligand screen). CD studies of the same ligands at 20 and 40 eq. show similar results (see Appendix, Figure S19) but highlight L827-1823 as a special case, providing minimal stability to 1Core+2.

L827-1823 is not currently well studied and has an interesting structure (see Table 3) that has not been observed amongst the ligands studied in the initial ligand screen. The fluorophenyl, oxadiazol, and pyridinyl parts of this ligand likely provide the main sites for interactions with 1Core through ' -stacking as observed with other DNA binders. Most of the other hits contain aromatic moieties (such as phenyl groups) that explain the interactions of these ligands with 1Core. L827-1823 is the only ligand of the
tested hits with the 1,4’-bipiperidine attachment. What is unclear at the moment is possible role of the 1,4’-bipiperidine for the binding interaction of L827-1823 with 1Core. Competitive studies of this ligand against other oligos revealed it is most selective for particular riboswitches and GQ (Appendix, Figure S18). These results may suggest some similarities between GQ and 1Core, which corroborates data from the previous ligand screen. Ultimately, L827-1823 possesses some set of characteristics that makes it especially interactive with 1Core compared to the other tested ligands.

Conclusion

Similar to our initial ligand screen, 1Core+2 seems to be non-interactive with most ligands. An exceptional ligand was L827-1823, showing structural preservation and stabilization of 1Core+2. Structural analogs can be designed based on L827-1823 and other notable hits that bind to 1Core+2 to figure out what structural moieties essential for establishing ligand-DNA interactions. Future studies should pursue more precise characterization of the binding interactions via UV-vis titrations. Studying the interactions between L827-1823 and longer (CAGAGG)ₙ repeats may also be fruitful. In general, this SMM method should be extended to longer analogs of (CAGAGG)ₙ. This work presents one of the first studies of a DNA repeat (excluding GQ repeats) in a wide ligand screen.
Chapter 5: Mutagenesis study of 1Core+2

Point mutation study

Introduction

To understand how each nucleotide contributes to the fold and stability of 1Core+2, we tested the model using single-point mutants. These mutants were designed by replacing each of the 24 nucleotides in 1Core+2 with thymine (N→T), the only nucleotide missing from the native sequence. If a nucleotide is crucial for the fold or stability, notable differences in both the spectroscopic signatures and thermodynamic parameters relative to the native structure should be observed. This study will help unravel the importance of core and loop nucleotides for the fold of 1Core+2. Each mutant is named based on the residue mutated and its place within the sequences, i.e. the point mutant of the first adenine in the sequence is termed A1T.

Results

All single point mutants of 1Core+2 were tested in CD melting studies. Melting curves were analyzed to yield melting temperatures (shown in Figure 19A and Appendix, Figure S20), which were used to calculate destabilization temperatures caused by each individual point mutation using 1Core+2 as a reference (T_m = 50.5 ± 1.2 °C).

Our results show that mutating any nucleotide within the GC core destabilizes the 1Core+2 structure significantly, by 12 - 28 °C. Interestingly, mutating core guanines
destabilized the structure to a lesser extent as compared to mutating core cytosine residues – an average destabilization temperature is -15.28 °C vs -20.15 °C, respectively. On the other hand, the stability of AGAG loop mutants is comparable to that of 1Core+2 (± 5 °C), Figure 19B and 19C. Four particular loop mutants (G6T, A7T, G18T, A19T) had non-sigmoidal melting curves that made analysis using the two-state model difficult. These mutants and G21T were also tested using UV melting, which provided stability data that disagreed with CD melting experiments (see Appendix, Figure S21). Due to the fact that UV melting provided smoother melting transitions, we deemed the $T_m$ values from these experiments more reliable than those from CD for all the sequences.

![Figure 19](image_url)

**Figure 19.** (A) Representative CD melting data for mutants that increased (A5T), decreased (G3T, G9T), and did not change (A13T) the $T_m$ of 1Core+2. (B) Heat map of 1Core+2 using rainbow color-coding scheme that shows $\Delta T_m$ due to mutation. (C) Data from B represented as destabilization temperature vs. point mutation. Bars for loops and overhangs are colored in black; bars for core GC nucleotides are colored in red.

CD scans were used to test structural changes caused by each mutation with respect to 1Core+2. According to the severity of the change in the CD signature, all mutants were categorized into three groups. Mutants in Group 1 maintain all notable features of the CD signature of 1Core+2 with minimal changes; Group 2 mutants have decreased intensity and/or decreased presence of the features of 1Core+2; Group 3
mutants show significant decrease in intensity and/or absence of 1Core+2 features. These three groups are represented on a 1Core+2 model using a color gradient from white (Group 1) to black (Group 3) as shown in Figure 20A. Group 1 consists of all nucleotides in loop 2 (bottom loop) and A17T from loop (see Figure 20B). The most drastic change in CD signature (Group 3) is observed due to mutations in two middle nucleotides of the two top loops: G6T, A7T, G18T, and A19T (see Figure 20C).

Figure 20. Summary of CD signature change due to N&T mutation (N = C, A, and G). (A) Color-coded model of 1Core+2 where white indicates no change (group 1) and black indicates drastic change in CD signature (group 3). CD scans for mutants in group 1 (B) and group 3 (C).

Because TDS does not provide a definitive answer to what the structure of 1Core+2 may be, we only focused on the differences between TDS signature of 1Core+2 and the mutants to monitor structural deviations. The TDS spectra were grouped according to structural elements (loop/core) and nucleotides (A vs G vs C): G-core mutants, C-core mutants, G-loop mutants, and A-loop mutants, Figure 21. Interestingly, G- or C-core mutants display little deviation from the TDS signature of 1Core+2; this is especially true for C-core mutants (Figure 21A and 21B). Meanwhile, TDS of A- and G-loop mutants show some variations, especially for G6T, G18T, A7T, and A19T and somewhat lesser yet still pronounced differences for A5T and A17T (Figure 21C and
These differences correlated directly with dramatic changes in CD signature for these mutants (Figure 20C).

Figure 21. TDS of 1Core+2 and (A) G-core mutants, (B) C-core mutants, (C) G-loop mutants, and (D) A-loop mutants.

Discussion

The biophysical data from this study suggests the core and the top two loops of 1Core+2 play crucial roles to the fold. Mutation of any nt within the core of the structure caused noticeable decrease in $T_m$ while mutations of top loop nt induced drastic topological changes observed by both CD and TDS. It is unsurprising the core nt
contribute the most to stability as they probably participate in some form of GC WC bp interactions which are among the most stabilizing base pairs.38

Interestingly, this study shed light on an asymmetrical contribution between nt in the bottom and top tetrads. G-to-T mutations in the bottom and top tetrads resulted in $\Delta T_m$ of -15.6 and -11.7 °C, respectively (averaged from the four mutants). Data for C-to-T mutations match these findings where these mutations in the bottom and top tetrads cause $\Delta T_m$ of -14.8 and -22.5 °C. This asymmetric contribution of the bottom and top GCGC tetrads to the stability of the core may be explained by asymmetry in structural contributions between the bottom and top loops. The top loops have been found to be important for the 1Core+2 fold either through inter-loop interactions such as non-canonical AG bp or base-stacking with the core. The GCGC tetrad data suggest the latter possibility, which explain why the top GCGC tetrad (and not the bottom tetrad) can undergo mutation and keep the structure stable. The bottom region as seen through the data is not as well structured and hence does not provide stability via stacking to the bottom tetrad.

Another peculiar finding of this study is G-T and C-T to mutations in the core cause different degrees of destabilization (13.7 and 18.7 °C, respectively). It is important to highlight the presence of contiguous guanines within the structure, which may act as “spare tires” when core guanines are mutated to maintain the fold and stability. Similar mechanisms have been found for GQ folds where ‘spare’ G-tracts can replace damaged G-tracts that are a few nt away.62 Cytosines within 1Core+2 do not have this “spare tire” option, which explains why mutating them causes significant destabilization. This hypothesis has been supported by testing double mutants where the contiguous guanine
tracts where mutated and caused marked destabilization and unfolding (see Appendix, Figure S22).

Interestingly, the top loop nucleotides play a particular role for the fold; they are engaged in non-canonical AG bp or AGAG tetrad formation. The top AGAG tetrad is the most structurally important as mutations to any nt in it cause loss of the unique 1Core+2 fold. With respect to the model, the importance of these nt is likely based on their role in facilitating tetrahelical nature of 1Core+2 by stabilizing the two loop turns through non-canonical base pairing. The bottom AGAG tetrad is not involved in the loop turn (where backbone polarity changes) and likely plays a less important role for the fold. The bottom tetrad is still probably engaged in base-stacking with both the top AGAG tetrad and the top GCGC tetrad which helps stabilize the overall fold. In addition, guanines within the bottom AGAG tetrad can act as spare tires which highlights the bottom tetrad is less structurally important. Removal of one nt from the top AGAG tetrad would cause either instability and structural deviation. The latter result was shown for loop mutants where the second guanine (AGAG) is removed (see the next section of Chapter 5).

**Conclusion**

We thoroughly tested the structure of 1Core+2 utilizing single point mutations. CD and UV spectroscopies were primary tools for monitoring changes caused by N→T mutation to 1Core+2 folding and stability. Nucleotides within the two-tetrad GCGC core of the structure caused marked destabilization and structural change upon mutation. Mutating nucleotides A7, G8, G18, and A19 within the two top loops caused great conformational change. These results suggest there are important interactions either between the nucleotides in the loops or stacking of loop nucleotides onto the terminal
GCGC tetrad that our current model is not accounting for. Meanwhile, the mutation of nucleotides in the bottom loop had no effect on structure and stability of 1Core+2 suggesting that it is not engaged into any. This data supports existence of non-canonical base pairing in 1Core+2. Amending our current model to show these AG base pairing interactions will more clearly represent the fold. These results corroborate the non-canonical AG interactions surmised in the hairpin model of 1Core+2. While this study provided crucial structural information on 1Core+2, it does not rule out either our tetrahelical model or the hairpin model. Further studies need to pursued to distinguish between the two structural possibilities of 1Core+2.

Loop variations

Introduction

The point mutation study of 1Core+2 suggested that top loops (according to our proposed model not the hairpin model) play structurally important roles. These results show the need to understand how the loops contribute to the fold of 1Core+2. Studying the loops provide a useful handle for testing the proposed tetrahelical model vs. the hairpin model. Previous CD studies on hairpins show that as the internal loops of hairpins change in length the overall fold is maintained. If 1Core+2 is a hairpin, we expect the fold to remain unchanged as loop length varies (see Figure 22). For this study, we designed 1Core+2 variants with different loop lengths termed Ln where n is the number of bases in each loop.
We expended this study by including two variants, TGAT, and TGAGAT. The point mutant study of 1Core+2 showed the central nucleotides within loops 1 and 3 are crucial for the fold. We hypothesized that the peripheral loop nucleotides (first and fourth in the loop) should play less significant roles and should not influence the fold and stability of the 1Core+2 structure. Thus we designed one variant where the loop length was maintained but two peripheral nucleotides were converted to T (TGAT loop); and the second variant which contained two spacer thymine residues added at the extremities of AGAG loop (TAGAGT loop). These two variants allow us to investigate whether the loop nucleotides participate in base stacking with the GCGC core. The TAGAGT variant
is expected to be either unstable if the G/A nt stack on top of GCGC tetrad or stable if this is not the case.

Results

Mutant L0 with its sequence 5-AGG CG CG CG CAG-3’ contains no predicted loop nucleotides. We expect it to form a Watson-Crick bimolecular duplex. The CD signature (peak at 280 nm and trough at 250 nm) and TDS signature (274.5 nm peak, indicative of 100% GC base pairing) confirm its duplex character (Figure 23A). PAGE results suggests a bimolecular nature of the L0 DNA structure based on its mobility which is more consistent with a dimer than with a monomer. Specifically, its mobility is the same as that of 1Core+2, which contains twice as many nucleotides. L0 has high thermal stability (Tm = 70.3 ± 2.5 °C), which is consistent with high degree of GC base pairing and supports the TDS data.

Figure 23. (A) CD scans and (B) TDS of 1Core+2 loop mutants L0, L1, and L2.

The inclusion of one internal AG bp going from L0 to L1 causes structural change observed in CD and TDS. The peak in CD at 274 nm is different from that for L0 at 280 nm; the same is true about TDS, Figure 23B. TDS reveals the peak of L1 is blue shifted (264 nm) with respect to that of L0 (274.5 nm). Because AG base pairs correspond to TDS features at 260 nm, the observed blue shift is unsurprising.28 On PAGE, L1 has a
The next mutant is L2 which contains three AG loops. This mutant could be
represented as a tetranucleotide repeat (CAGG)n which belongs to a larger family of
repeats.

![Figure 24](image_url)

Figure 24. Native 12% gel run in 1xTBE buffer with 3 mM MgCl₂. Ladders and
1Core+2 are included for comparison. Mobility of mutants clearly depends on loop
length. Previous studies have shown that duplexes with a single internal non-canonical
transitions. Previous studies have shown that duplexes with a single internal non-canonical
transitions, the secondary structure of L1 is unstable as it produces an unstacked
ladder.
locations (259 and 266 nm for L2 versus 256 and 269 nm for (CAGG)_{22}) as shown in (Figure 6). The TDS may be detecting subtle difference between the folds, which may be related to the 293 nm peak present in the CD spectra of L2 but missing in (CAGG)_{22}. Oligos L2 nd (CAGG)_{22} have very similar stabilities (60.6 ± 0.4 and 60 ± 1.5 °C for CAGG_{22} and L2, respectively), highlighting similar structures.

Mutants for L3 comprises three variants based on single nt deletion from each of the AGAG loops of 1Core+2. Mutant L3-AGG was excluded from this study as it would form a G-tract (CAGGG) and promote the formation of an alternative secondary structure, a GQ.\textsuperscript{19} CD scans of L3-AAG and L3-AGA show they share similar spectroscopic signatures with peaks at 279 and 276 nm respectively and troughs at 248.3 and 251.1 nm, respectively (Fig. 25A). In TDS, L3-AAG has a two-peak signature (257.7 and 265.8 nm) while the signature of L3-AGA has single peak at 264 nm (Fig. 25B). The stabilities of these two variants are the same (55.1 ± 0.3 °C for L3-AAG and 55.6 ± 0.5 °C for L3-AGA). These results suggest these mutants have mostly duplex character.

**Figure 25.** (A) CD scans and (B) normalized TDS of 1Core+2 L3 loop mutants.

Unlike the last two L3 variants, L3-GAG showed the same trough-peak-two shoulder CD signature as 1Core+2 though both its peak and trough are red-shifted by ~4
nm shown in Figure 25A. The TDS of L3-GAG has a peak at 260 nm (instead of 262 nm like 1Core+2) and a dip at 296 nm which is indicative of G:G hoogsteen bp (Figure 25B). L3-GAG has T<sub>m</sub> of 67 ± 1 °C which is more stable relative to the other variants and can be partially justified by presence of G:G bp. 38 Interestingly, PAGE results for L3-GAG showed the presence of higher order species, which differs from that of the other L3 variants that showed monomeric folds (Figure 24). This data suggests CD data is reporting the average of monomer and dimer species in solution (causing the degree stacking and stability to be enhanced relative to the other mutants).

It is important to note L3-GAG is sequentially similar to GCG-rich DNA which are known for having unique tetrahelical folds. 48 These structures are built from GGG and GCG tracts separated by single adenine residues. Looking at the sequence of L3-GAG (5'-AGGCGAGCGAGCGAG-3'), it is clear to see the presence of GCG tracts, which may be a driving force for a tetrahelical folds similar to those shown by the Plavec laboratory. In addition, the presence of the G-quartets (or G:G base pairs) in Plavec’s NMR structures are consistent with the G:G hoogsteen bp suggested by the TDS data for L3-GAG. The Plavec structures VK1, VK2, and VK34 all share the same CD signature (trough-peak-two shoulder) as L3-GAG. These unusual folds existed in monomers, dimers, or tetramers based on their sequence, corroborating the presence of a higher ordered species observed for L3-GAG in PAGE. The parallels between Plavec’s structures and L3-GAG suggest that they have similar folds. By extension, the similarity in the spectroscopic signatures of L3-GAG and 1Core+2 hints at the possibility that the latter also adopts a tetrahelical fold.
The effects of lengthening loops have been studied in mutants in L6 and L8. Increasing the AG pairs in the loops of 1Core+2 does not change the characteristic CD or TDS signatures, suggesting that the native fold is maintained. Importantly, both mutants have enhanced character at 264 and 244 nm in the CD signature. The TDS of L6 and L8 is unchanged (see Figure 26A and B). Mutants L6 and L8 have $T_m$ values of 44.5 ± 0.3 °C and 40.1 ± 0.4 °C, respectively, both lower than that of 1Core+2 ($T_m = 50.1 ± 1.2$ °C).

![Figure 26. (A) CD scans and (B) TDS of 1Core+2 loop mutants L6 and L8.](image)

Using mutants TGAT and TAGAGT, the influence of the central and peripheral loop nucleotides on the fold and stability were tested. Mutant TGAT has a CD signature with a peak at 278 nm, trough at 249 nm, and shoulder at 290 nm and thermal stability of 51.6 ± 0.3 °C (Figure 27A). This spectroscopic signature of TGAT suggests duplex character. The CD spectrum of TAGAGT variant has extremely low intensity and the peaks at 290 and 244 nm and a trough at 260 nm suggesting a poorly ordered assembly with inversion of handedness (Figure 27A). Consistent with this data, TAGAGT displays no clear melting transition (Figure 27C). PAGE results show that TGAT has the same mobility as 1Core+2 (Fig. 24). TAGAGT has a much slower mobility than 1Core+2, further supporting its lack of structure.
Discussion

The loop study corroborates the point mutant study, showing the loops are key to the unique fold observed for 1Core+2. Building the native structure (1Core+2) from no loop nucleotides show a clear structural transition from a dimeric WC duplex (L0) through unfolded L1, to a hairpin (L2), to a spectroscopically unique fold (L3-GAG). The structures of L3-L8 are identical and only vary in the degree of base stacking shown through CD signal amplitude. Based on the data, we surmise odd numbered loop variants such as L5 and L7 will have the same fold as the native structure since they meet the minimum loop length requirement of three nt. There is another clear trend where the stability is inversely proportional to the loop length, which is justified by the increased flexibility introduced with lengthened loops (Figure 28).
Figure 28. Effect of changing the number of loop nt on stability where X represents absence of clear melting transition.

The results show that mutants with loop length shorter than three (besides L3-GAG) do not fold into a structure similar to 1Core+2. And even in the case of L3-GAG, unlike 1Core+2, it exists in both monomeric and dimeric forms, thus all the data presented are the average of the two states. Judging by the CD signature amplitude (corrected for the number of nt in the sequences) which reflects the extent of base-stacking, 1Core+2 and the other L3 variants compare well to L3-GAG suggesting the same extent of folding. Shorter variants have lower signals and are, hence, less folded. Thus we conclude that loops with minimum 3 nt are required for the correct folding of 1Core+2.

Interestingly, 1Core+2 does not meet the sequential requirements for the fold of Plavec’s tetrahelical structures. These assemblies are based on GCG and GGG tracts. L3-GAG showed the fold is obtainable without GGG repeats. Strikingly, 1Core+2 contains no GCG tracts yet folds into the a similar structure to that of VK34 and L3-GAG. One
possible explanation for this is one of the adenines in the AGAG region become bulges that allow GCG tracts (AGGAGGAGCAG) as found in GQ.\textsuperscript{48} However, this possibility is unlikely as our data shows the adenines within the loops play critical structural roles and thus likely form hydrogen bonding with other bases.

We also studied two additional loop mutants TGAT and TAGAGT which showed severe effects on the fold, supporting the importance of the AGAG loop to 1Core+2. The fold of TGAT variant is distinctly different from that of 1Core+2 showing that the internal loop nt (AGAG) are not the only structurally important features of the top loops. The peripheral loop nt (AGAG) are likely base-stacked on top of the GCGC tetrad. The weak CD signature of TAGAGT shows it is unfolded. Inserting the spacer-T removed such interactions and completely destabilizes and unfolds the structure. This unfolding is probably the result of the enhanced loop length which being unstructured causes unraveling of the structure.

As the loop length increased, drastic structural changes were observed from L0-L3, which do not support the hairpin hypothesis for 1Core+2 (\textbf{Figure 22}). If 1Core+2 folded into a hairpin, it would likely resemble the spectroscopic signature of L2 only with increased spectral character corresponding to AG base pairs (260 on both CD and TDS).\textsuperscript{28} In addition, the irrefutable parallels between the CD signature of studied tetrahelical folds (such as VK34 and 1Core+2 support a tetrahelical structure for 1Core+2.
Conclusion

These results corroborate the single-point mutant study that showed the loops played an important structural role. The tendency of d(AG)$_n$ to fold into high-order structures makes it unsurprising that they act as structural elements in the CAGAGG repeats. These results support the DNA repeat screen from Chapter 3 where the purine rich loops were suspected to be important for structure formation. The specifics of how these AGAG loops facilitate the structure fold remain unclear such as whether these loops remain unstructured or form AG bp or AGAG tetrads. Most importantly, the study showed the structure of 1Core+2 changes drastically from L0-L3 which is rather unexpected if it formed a simple hairpin. This work does not support the hairpin model and gives weight to our proposed tetrahelical model.

Fluorescence studies of 1Core+2 2-amino purine oligos

Introduction

The point mutant study of 1Core+2 highlighted the importance of the loop nt for the fold and stability. Further investigation of role of 1Core+2 loop nt in the loop mutant study showed loop composition was crucial to the structure. To survey the environment of specific bases located in the loops, we utilized fluorescence spectroscopy (FL). Adenine can be replaced with a fluorescent probe called 2-amino purine (2-AP) that reports on the degree of solvent-accessibility of its chemical environment (Figure 29A). Previous studies of 2-AP modified human telomeric GQ have been successful in studying its different topologies. Several adenine sites within 1Core+2 were modified with 2-AP
as shown in Figure 29B. Based on our previous biophysical data, adenines likely participate in a variety structurally important interactions based on their position. More solvent accessible bases will fluoresce stronger as compared to the bases involved in base-pair formation or base-stacked.

**Figure 29.** (A) Molecular structures of adenine (*top left*) and 2-amino purine (*top right*). (*Bottom*) Possible base pairing schemes for AG interactions. (B) Model of 1Core+2 with modified adenine sites labeled and numbered.

**Results**

To test whether 2AP modification affected the structure of 1Core+2, CD and UV spectroscopic studies were completed. CD scans show that some of 2-AP modifications induced structural change (**Figure 30A**). Oligos 1-13 and 1-23 maintain the same trough-peak-two shoulder signature as 1Core+2 though both 1-13 and 1-23 have lower CD signal amplitude (at the 242 nm trough and 264 nm peak). Oligo 1-5 has a similar CD signature with a red-shifted peak that is of lower intensity than that of 1Core+2. Oligos 1-7 and 1-19 have similar CD signatures with broad peaks in the 275 nm range and a subtle shoulder at 290 nm. In addition, 1-19 has less CD signal amplitude compared to 1-7. TDS
of the oligos show they all have the same signature as 1Core+2 (Figure 30B). There are subtle shifts in the TDS of 1-5, 1-7, and 1-19, reflecting changes in the secondary structures. These results coincide with the CD signature data that showed these oligos were structurally distinct from 1Core+2. Unsurprisingly, point mutants A5T, A7T, and A19T were also structurally distinct from 1Core+2 (shown via CD and TDS). CD melting experiments provided Tm for 1-5, 1-13, and 1-23 of 45.3 ± 0.2 °C, 45.9 ± 0.5 °C, and 48.1 ± 0.4 °C, respectively with the Tm for 1Core+2 of 50.1 ± 1.2 °C (Figure 30C). On the other hand, 1-7 and 1-19 did not have clear melting transitions.

The secondary structures of 1Core+2 2-AP oligos were also studied via denaturing and native PAGE. Denaturing gel shows that all the oligos have the same mobility indicating that as synthesized the oligos are pure (Figure 31A). Interestingly, the native gel showed that the oligos have varying motilities reflecting their different folds (Figure 31B). Oligos 1-13 and 1-23 have the same mobility as 1Core+2, while oligos 1-5, 1-7, and 1-19 have slower mobilities. The order in decreasing mobility is: 1-5 > 1-7 > 1-19. These results corroborate CD data that showed 1-5, 1-7, and 1-19 were (differently and) less folded compared to 1Core+2 in the exact same order.
likely only effects AG bp interactions, the data suggests the loop interactions are

suggest they have the same secondary structure. Since subtituting adenine with 2-AP
13 and 1-23 have the same fold as 1-Core+2 and are consistent with the gel results, which
oligos are less folded and are much more slower on the gel. CD and TDS data showed 1-
in both CD signatures and TDS were observed. Native PAGE also showed these three
substitutions. Sites A5, A7, and A19 seem to be especially crucial to the fold as changes

These results suggest the fold and stability of 1-Core+2 is sensitive to the 2-AP

In the 1 x TBE buffer with 3 mM MgCl2, impurities do not contribute to the observed biophysical data. (B) Native 12% PAGE gel

imputities were not observed. Instead, biophysical data were consistent with the

different in the two syntheses but biophysical data were consistent, we assume that the

re-synthesized gel. The impurities observed because the amount of impurities were

comparison, Note, 1-7, 1-9 and 1-23 have some impurities. These oligos were

Figure 31. Characterization of 1-Core+2 and 2-AP oligos with PAGE. (A) Denaturing
crucial to the stability and fold. This observation is consistent with the point mutant study where loops 1 and 3 seemed to play a more important role for the fold. The data suggest A13 and A23 do not participate in any AG base pairing.

2AP-modified 1Core+2 mutants were tested via fluorescence spectroscopy. The fluorescence responses for these oligos in increasing order is 1-5 < 1-7 < 1-19 < 1-13 < 1-23 (Figure 32).

![Figure 32. Characterization of 1Core+2 2-AP oligos with FL. (A) Emissions scans and (B) summary of fluorescence responses monitored at 365 nm (with error bars included as an average of three measurements) for 2-AP-modified 1Core+2 mutants. Data were collected at 0.05 Abs of oligonucleotides in 100K2Mg buffer at 4 °C.](image)

**Discussion**

Oligos 1-23 is located at the 3’ tail of 1Core+2 and is expected to be highly flexible and solvent exposed. Introduction of 2-AP into this oligo did not affect its fold and stability to a great extent and yielded the highest fluorescence response consistent with its solvent exposed environment. 1-13 is located in the bottom loop of 1Core+2, which is expected to be rather relatively flexible and solvent accessible. Our other data (Chapter 5, **Figure 19A and 20A**) suggest that this loop does not form any interaction
and is fully flexible. In agreement this 2-AP mutant shows the second highest
fluorescence response. Lower value of FL as compared to 1-23 is most likely due to
restrained motion of the loop as compared to the overhang but is not indicative of the
presence of any base-pairing or base-stacking.

Oligos 1-5, 1-7, and 1-19 are part of the structurally-important top loops of
1Core+2. 2-AP modification in those oligos lead to modified CD signature suggesting
that the A in these oligos were engaged in important secondary interactions. Interestingly,
single point mutant studies suggested that A7 and A19 are engaged in crucial base-
pairing interactions while A5 is not (Chapter 5, Figure 20A and 21A). Thus our original
expectation was that 1-7 and 1-19 will produce the lowest FL response and the response
for 1-5 will be high. In addition, 1-7 and 1-19 substitutions are symmetrical and were
expected to produce the same changes in CD, UV, and FL. The observed data in Figure
32 are not in line with our expectation and can be interpreted in the following way. While
A7 and A19 are expected to be engaged in A:G base-pairing interaction, one face of these
bases should be solvent-accessible. This type of chemical environment is called partially
stacked; opposed to a fully stacked environment where both faces of adenine are shielded
from the solvent as could happen for example, in the middle of duplex DNA. The
intensity of CD scan for oligo 1-19 suggests that it is less folded than its proposed
symmetry mate 1-7. This loss of fold is reflected in the higher FL response of 1-19. Oligo
1-7 has the second lowest FL response and thus supports our expectation that A7 is
stacked or base-paired. Oligo 1-5 has the lowest FL intensity. Combining the data from
single-point mutant studies where replacing this oligo with T did not lead to
destabilization of 1Core+2, we can conclude that A5 does not participate in base-pairing
(where the identity of the base will be crucial) rather it participates in stabilizing base-stacking and is in a solvent-restricted environment. Specifically, it stacks on top of GCGC tetrad on one face and G-A base-pair on another. It is important to take into account the possible FL quenching of 2-AP due to neighboring guanines. However, in our case, it does not seem to be of great importance. Specifically, all oligos besides 1-5 and 1-23 have two neighboring guanines, yet FL of oligo 1-5 is the lowest possibly reflecting the high degree of stacking experienced by A5.

**Conclusion**

In sum, the FL results for 2-AP modified 1Core+2 mutants have provided further validation of the model and shed the light on the roles of adenines for the fold. The study confirmed that loop 2 is not engaged in any interaction as well as the overhangs. Unlike previous expectations it seems that A7 and A19 are not perfectly symmetric yet both are needed to maintain the stability. A5 is in the most solvent-restricted environment suggesting that it is base-stacked which was not obvious from our single-point mutant studies.
Chapter 6: Conclusion

This work delineated (CAGAGG)$_n$ as stable and structurally unique amongst other DNA repeats linked to replication stress. The ligand study showed 1Core interactions with predominately GQ binders and not duplex or triplex binders. While 1Core does not fold into a GQ, it probably possesses similar characteristics such as being a tetraplex. The data from the various newly designed structural studies in this work supported our proposed model and provided gained tremendous structural information for (CAGAGG)$_n$. The GCGC core of the structure is essential for the stability as anticipated. Unexpectedly, we found the top AGAG loops are required for unique fold. We expect AG base pairing interactions in the top loops, which explain their structural role. The bottom half of the structure contributes less to the stability and the fold. Both the length and composition of the loops are crucial to 1Core+2. The minimum loop length required for the fold is three nucleotides; lengths below that threshold result in drastically different, less folded structures. The structural sensitivity to single nucleotide additions to the loops does not support the hairpin model for 1Core+2. 2AP studies showed the top half of the 1Core+2 structure to be more crowded (likely due to base-stacking) as expected based on our refined model (accounting for AG base pairs). The data presented in this work suggest 1Core+2 is not a simple duplex, or hairpin, but rather a tetraplex as depicted in our model.

Despite the vast structural knowledge we gained for (CAGAGG)$_n$, there is still much to learn. The core of the structure is suggested to be a made up of GCGC tetrads of unspecified nature. These tetrads can come in various forms such as minor groove, major groove, and slipped.$^{21,48,69,70}$ Studies using either cytosine or guanine analogs may be able
to capture specific base pairing configurations, which can shed light on the GCGC tetrad conformation. Pursuing the crystal structure is another goal for this project (led by Barrett Powell ’18, Honors Thesis) but finding a structural solution has proved elusive. Expanding our ligand study may produce binders that help produce more stable, higher quality crystals.

Interestingly, CAGAGG is similar in sequence to other disease associated DNA repeats, CAG and CAGG. Each of these repeats adopts their own unique folds and consequently has their own biological function. Unlike CAG or CAGG, CAGAGG does not form a hairpin but rather some other unique structure. This work overall supports the hypothesis of the Brown lab. The repeat most highly associated with replication stress (CAGAGG) is structurally stable and novel, supporting the idea that stable DNA structures can physically block DNA polymerase during the replication process. Expanding our knowledge of the types of structures these repeats can adopt will be critical to understanding their biological roles. Furthermore, this structural information can be utilized to design drugs with anti-cancer activity by destabilizing problematic structures such as CAGAGG.
References


(32) Barlow, J. H.; Faryabi, R. B.; Callén, E.; Wong, N.; Malhowski, A.; Chen, H. T.; Gutierrez-Cruz, G.; Sun, H.-W.; McKinnon, P.; Wright, G.; et al. Identification of


Chapter 3: Ion dependence study of (CAGAGG)$_n$

To understand the fold of (CAGAGG)$_n$, the ion dependence of (CAGAGG)$_{10}$ (termed CA10) was tested for various monovalent and divalent ions. While Na$^+$, Li$^+$, and K$^+$ provide about the same degree of stabilization of the CA10 fold, Mg$^{2+}$ provided the most notable degree of stabilization. This study is extended through testing the effects of Zn$^{2+}$ and Pb$^{2+}$ on the CA10 fold. Addition of both of the divalent ions decrease the CD signal amplitude of CA10 (more prominent for Zn$^{2+}$) and diminish the 280 nm shoulder (Figure S1A and S1B). According to CD melting studies, Zn$^{2+}$ at 1 mM and 2 mM cause $T_m$ of -5.4 and -1.0 °C, respectively, while Pb$^{2+}$ at 1 mM and 2 mM cause $T_m$ of 13.4 and 7.4 °C, respectively compared to CA10 in 10 mM lithium cacodylate (see Figure S1C).

**Figure S1.** CD scans of CA10 with (A) ZnCl$_2$ and (B) Pb(NO$_3$)$_2$. (C) Bar-graphs summarizing results from CD melts of CA10 with both ZnCl$_2$ and Pb(NO$_3$)$_2$. All experiments were conducted in buffers that contained 10 mM lithium cacodylate pH 7.2 with specified ions at 1, 2, or 5 mM.
The Zn$^{2+}$ data for CA10 corroborate results from the Zn$^{2+}$ dependence studies of 1Core+2. Zn$^{2+}$ causes gradual structural degradation of 1Core+2 as shown in the CD scans (Figure S2A). Interestingly, the CD melting data shows destabilization of 1Core+2 and subsequent stabilization of the less folded structure (Figure S2B). This unusual T$_m$ pattern is also observed in the CA10 study. CD titration of 1Core+2 with Zn$^{2+}$ shows the gradual loss of the 276 nm shoulder DNA structure (Figure S2C).

Figure S2. (A) CD scans and (B) normalized melting curves of 2.5 µM 1Core+2 with ZnCl$_2$. (C) Titration of 1Core with ZnCl$_2$ monitored via CD.

The CD data suggest Zn$^{2+}$ and Pb$^{2+}$ play two opposing roles for CA10. Zn$^{2+}$ destabilized the native fold, then stabilizes a different, less folded structure. This other secondary structure resembles (CAGG)$_n$ which has a CD signature that contains two peaks at 260 and 290 nm. These results are unsurprising as Kypr et. al demonstrated Zn dependence of d(AG)$_n$ repeat structures. Zn$^{2+}$ interacts with d(AG)$_n$ and likely influences the AGAG regions of the CAGAGG repeat. As CA10 undergoes a structural transition from its native form to ssDNA, the AG sites are liberated from the fold and permitted to more effectively interact with Zn$^{2+}$. Kypr showed that d(AG)$_n$ homoduplexes were stabilized by Zn$^{2+}$ but this is not observed spectroscopically for (CAGAGG)$_n$. Due to the other residues within the sequence that cannot engage in AG base pairing, this homoduplex structure may be rather unfavorable and therefore unobserved via CD.
Pb\(^{2+}\) seems to stabilize the fold that resembles (CAGG)\(_n\) yet unfolds the structure to less of a degree at 2 mM compared to Zn\(^{2+}\) at 2 mM. Of the tested divalent cations, Mg\(^{2+}\) is the only one to stabilize and maintain the secondary structure of CA10. Whether the ability of Zn\(^{2+}\) to stabilize (CAGG)\(_n\) like folds sheds light on the relationship of the tetranucleotide repeat to the zinc finger protein 9 gene is still unknown.\(^{33,66}\)

**Chapter 4: Interactions of 1Core with small molecule ligands**

**Table S1.** List of ligands investigated in this study and their effects on 1Core. \(\Delta T_m\) indicates the degree of stabilization (with the positive sign) or destabilization (with the negative sign) of 1Core.

<table>
<thead>
<tr>
<th>Ligand</th>
<th>UV-Vis P:DNA (K(_d), nM)</th>
<th>CD Scan</th>
<th>(\Delta T_m) (°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-quadruplex binders</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 NMM(*)</td>
<td>Yes, no change in Soret max</td>
<td>No change in the CD signature of CA10 or CA15</td>
<td>~0 for both CA10 and CA15</td>
</tr>
<tr>
<td>2 TMPyP4(**)</td>
<td>3:1 (60)</td>
<td>Significant structural change; signal intensity at 260 nm decreased, signal intensity at 280 nm increased.</td>
<td>nm</td>
</tr>
<tr>
<td>3 TMPyP3(**)</td>
<td>4:1 (103)</td>
<td>Significant structural change; red shift in signals.</td>
<td>nm</td>
</tr>
<tr>
<td>4 TMPyP2(**)</td>
<td>nm</td>
<td>No structural change; minimal decrease in signal intensity.</td>
<td>-0.6 at 1 eq. -1.1 at 2 eq. -1.9 at 4 eq. -2.2 at 6 eq. -15.3 at 2 eq. -18.5 at 4 eq. nm</td>
</tr>
<tr>
<td>5 ZnTMPyP4</td>
<td>nm</td>
<td>Minimal structural change; signal intensity decreased.</td>
<td>-0.3 at 2 eq. -0.9 at 4 eq.</td>
</tr>
<tr>
<td>6 CuTMPyP4</td>
<td>nm</td>
<td>Significant structural change; decrease in signal intensity.</td>
<td>+2.9 at 2 eq. +3.5 at 4 eq. +4.8 at 6 eq.</td>
</tr>
<tr>
<td>7 PM3P</td>
<td>Yes, change in Soret but inconsistent data</td>
<td>Minimal structural change; decrease in signal intensity.</td>
<td></td>
</tr>
<tr>
<td>8 RHSP4(**)</td>
<td>5:1 (4)</td>
<td>Minimal structural change; signal intensity increased</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Compound</td>
<td>nm</td>
<td>Description</td>
</tr>
<tr>
<td>----</td>
<td>-----------------------------------</td>
<td>----------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>9</td>
<td>Braco19**</td>
<td>nm</td>
<td>Minimal structural change; signal intensity decreased at 260 nm.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Phen-DC3</td>
<td>nm</td>
<td>Significant structural change; signal intensity decrease.</td>
</tr>
<tr>
<td>11</td>
<td>Pyridostatin (PDS)</td>
<td>nm</td>
<td>Minimal structural change; decrease in signal intensity.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Crystal Violet</td>
<td>nm</td>
<td>Minimal structural change; small red shift of signal and decrease of signal intensity.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Malachite Green</td>
<td>nm</td>
<td>Minimal structural change; very small increase in signal intensity.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Duplex binders**

<table>
<thead>
<tr>
<th></th>
<th>Compound</th>
<th>nm</th>
<th>Description</th>
<th>Changes at Eq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>Coomassie Brilliant Blue</td>
<td>nm</td>
<td>Minimal structural change; decrease in signal intensity.</td>
<td>-1 at 2 eq.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.1 at 4 eq.</td>
</tr>
<tr>
<td>1</td>
<td>Methylene Blue</td>
<td>nm</td>
<td>Minimal structural change; decrease in signal intensity.</td>
<td>+6.8 at 2 eq.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+9.7 at 4 eq.</td>
</tr>
<tr>
<td>15</td>
<td>Methyl Green</td>
<td>nm</td>
<td>Minimal structural change; decrease in signal intensity.</td>
<td>+0.6 at 2 eq.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+3 at 4 eq.</td>
</tr>
<tr>
<td>17</td>
<td>Neutral Red</td>
<td>nm</td>
<td>Minimal structural change; decrease in signal intensity.</td>
<td>+0.9 at 2 eq.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.2 at 4 eq.</td>
</tr>
<tr>
<td>18</td>
<td>Ethidium Bromide</td>
<td>nm</td>
<td>Minimal structural change; very small decrease in signal intensity.</td>
<td>-0.4 at 2 eq.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+0.5 at 4 eq.</td>
</tr>
</tbody>
</table>

**Triplex Binders**

<table>
<thead>
<tr>
<th></th>
<th>Compound</th>
<th>nm</th>
<th>Description</th>
<th>Changes at Eq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>Coralyne Chloride Form Berberine Chloride</td>
<td>nm</td>
<td>Minimal structural change; decrease in signal intensity.</td>
<td>+0.4 at 2 eq.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.9 at 4 eq.</td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td>Minimal structural change; almost no change in signal intensity.</td>
<td>-0.9 at 2 eq.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+5.9 at 4 eq.</td>
</tr>
</tbody>
</table>

**Major Groove Binders**

<table>
<thead>
<tr>
<th></th>
<th>Compound</th>
<th>nm</th>
<th>Description</th>
<th>Changes at Eq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>Berenil</td>
<td>nm</td>
<td>Minimal structural change; decrease in signal intensity.</td>
<td>+0.3 at 2 eq.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+1.6 at 4 eq.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>+1.5 at 6 eq.</td>
</tr>
<tr>
<td>22</td>
<td>Pentamidine Isethionate Salt Hoechst 33258</td>
<td>nm</td>
<td>Minimal structural change; decrease in signal intensity.</td>
<td>-0.5 at 2 eq.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-5.2 at 4 eq.</td>
</tr>
</tbody>
</table>

*Berberine Chloride*
Berberine Chloride is an alkaloid isolated from Chinese herbs and is used frequently as medicinal herb (Figure S3A). Previous studies showed Berberine Chloride preferentially binds to H-DNA but it can engage in intercalative binding with dsDNA.\textsuperscript{71,72} CD showed little signature change for 1Core in the presence of 2 and 4 eq. of Berberine Chloride (Figure S3B). Similarly, no stabilization was noticed in the CD melt, with the stabilization temperature, $-T_m$ equal to 0.4 °C at 2 eq. and -1.0 °C at 4 eq. of Berberine Chloride (Figure S3C).

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure.png}
\caption{(A) Structure of Berberine Chloride. (B) CD scans and (C) normalized CD melts of 4.25 µM 1Core with 2 and 4 eq. of Berberine Chloride.}
\end{figure}

\textit{Berenil}

Berenil is among many diarylamidine molecules that posses antimicrobial and antiparistic activities (Figure S4A). It is a known groove binder of DNA particularly at AT base pairs and has been shown to stabilize triplex DNA.\textsuperscript{71,73} CD scans show modest decrease in the CD signal intensity without change in the signal shape of 1Core (Figure S4B). No stabilization was observed in the CD melt, where $-T_m$ was 0.5 °C, 1.8°C, and 1.7 at 2, 4, and 6 eq. of Berenil (Figure S4C). The CD data suggest minimal interaction between 1Core and Berenil. These results corroborate previous findings that showed a significant preference of Berenil to bind to AT-rich DNA over GC-rich DNA.
**Figure S4.** (A) Structure of Berenil. (B) CD scans of and (C) normalized CD melts of 5 µM 1Core with 2 and 4 eq. of Berenil.

*Comassie Brilliant Blue*

Comassie Brilliant Blue is a known dye that has been suggested to have DNA binding properties ([Figure S5A](#)). CD scans in the presence of ligand showed only a minor decrease in the CD signal of 1Core ([Figure S5B](#)). CD melting studies show the ligand did not stabilize the structure of 1Core (T_m equals -1 °C and -0.1 °C at 2 and 4 eq. respectively) as shown in [Figure S5C](#). The lack of change in the CD wavelength scan and absence of stabilization combined suggest absence of binding interactions between 1Core and Comassie Brilliant Blue R-250.

**Figure S5.** (A) Structure of Comassie Brilliant Blue R-250. (B) Overlay of pre and post-melt scans of 5 µM 1Core with 2 and 4 eq. Comassie Brilliant Blue R-250. (C) Normalized CD Signal for CD melt of 5 µM 1Core with 2 and 4 eq. of Comassie Brilliant Blue R-250.
**Coralyne Chloride**

Coralyne Chloride is a four-ringed DNA binder that has shown antitumor activity (Figure S6A). Previous findings demonstrated selectivity of Coralyne Chloride for triplex DNA.\(^71\) There is a modest decrease in CD signal intensity without the change in the signal shape as shown in Figure S6B. Minimal stabilization was observed in the CD melt: \(-T_m\) equal to -0.4 °C at 2 eq. and 3.5 °C at 4 eq. of the ligand (see Figure S6C). The results suggest Coralyne Chloride have weak binding interactions with 1Core. The molecule has been shown to stabilize triplex DNA (more than duplex DNA) through intercalative binding.\(^75\) The lack of interaction and stabilization suggest 1Core lacks triplex DNA character.

**Figure S6.** (A) Structure of Coralyne Chloride. (B) CD Scan, and (C) normalized CD melts of 4.25 µM 1Core with 2 and 4 eq. of Coralyne Chloride.

**CuTMPyP4**

CuTMPyP4 is a metallated version of TMPyP4 and a known, modestly selective GQ binder. Similar to TMPyP4, CuTMPyP4 is mostly planar with 4-coordinated Cu\(^{2+}\) in the center of the porphyrin ring as shown in Figure S7A. Previous studies suggest CuTMPyP4 weakly binds to parallel GQ through end-stacking and intercalation.\(^76,77\) CD scans show diminishment of the characteristic 265 nm peak and 279 nm shoulder of the 1Core CD signature with preservation of the 290 nm shoulder (Figure S7B). The spectral
change according to CD is corroborated by UV-vis titration experiments that demonstrate two-state binding as shown in Figure S7C. The results indicate CuTMPyP4 interacts with 1Core. The species formed upon complexation of the ligand with 1Core resembles an antiparallel GQ suggested by the 295 nm peak and 260 nm trough in the CD signature though the identity is still unknown. These results suggest 1Core is not a parallel GQ despite its 265 nm peak as CuTMPyP4 does not cause structural change of parallel GQ. The strong interactions of 1Core with CuTMPyP4 suggest it contains GQ characteristics.

![Figure S7](image)

**Figure S7.** (A) Structure of CuTMPyP4. (B) CD scans of 4.7 µM 1Core with 2 and 4 eq. of CuTMPyP4. (C)

**Crystal Violet**

Crystal Violet (CV) is a triphenylmethane dye that has been shown to bind selectively with GQ (Figure S8A). CD scans showed a decrease in signal intensity at 265 and at 240 nm along with appearance of a new negative signal 310 nm (Figure S8B). The CD melting studies shows modest stabilization of 4.3 and 6.0 °C at 2 and 4 eq. respectively (Figure S8C). Modest stabilization accompanied by clear conformational change suggests interactions between 1Core and CV. As suggested in the study of 1Core with EB, the negative bands near 300 nm may provide further indication of CV binding.
with 1Core.

**Figure S8.** (A) Structure of CV. (B) CD scans, and (C) normalized CD melt for 4.7 µM 1Core with 2 and 4 eq. of CV.

*Hoechst 33258*

Hoechst 33258 is bisbenzimidazole derivative that is known for its DNA binder ability (**Figure S9A**). Bailly and coworkers demonstrated minor groove and intercalative binding of Hoechst 3258 with AT and GC-rich DNA.\(^{79}\) There is modest decrease in the signal intensity upon addition of 2 and 4 eq. of Hoechst 33258 to the 1Core (**Figure S9B**). Hoechst 33258 is an achiral molecule and was demonstrated to induce a positive CD band at 353 nm upon DNA binding.\(^{80}\) The lack of spectral change in the 300-330 nm range suggests only free Hoechst 33248 exists in solution. No stabilization occurred in the CD melt as observed - \(T_m\) is 0.8 °C at 2 eq. and 1.5 °C at 4 eq. of Hoechst 33258 (**Figure S9C**). The results suggest no interactions of Hoechst 33258 with 1Core and 1Core lacks duplex character.
Figure S9. (A) Structure of Hoechst 33258. (B) CD scans and (C) normalized CD melts of 4.9 µM 1Core with 2 and 4 eq. of Hoechst 33258.

Malachite Green

Malachite Green is another triphenylmethane dye that has been shown to preferentially bind to GQ (Figure S10A). Minimal CD signal disruption was shown in CD scans of 1Core in the presence of Malachite Green (Figure S10B). CD melts showed no stabilization: $T_m$ of -0.9 °C at 2 eq. and -0.7 °C at 4 eq (Figure S10C). The results suggest Malachite Green does not interact with 1Core. CV another triphenylmethane showed interactions with 1Core. Compared to CV, Malachite Green is only missing an amino group on one of the phenyl rings. These results suggest the presence of two amino groups is necessary for the binding interactions of 1Core with triphenylmethanes.

Figure S10. (A) Structure of Malachite Green. (B) CD scans and (C) normalized CD scans of 5.4 µM 1Core with 2 and 4 eq. of Malachite Green.

Methyl Green
Methyl Green is a triphenylmethane and the only major groove binder of this ligand screen (Figure S11A). CD scans showed decreased signal amplitude at 265 and 290 nm (Figure S11B). CD melting studies revealed minimal stabilization: $T_m$ of 0.6 °C and 3 °C at 2 and 4 eq., respectively (Figure S11C). The results suggest minimal binding interactions between Methyl Green with 1Core. While minor groove binding is disallowed for G-rich DNA such as 1Core due to the N2 amino These results support lacking duplex character of 1Core.

![Figure S11](image)

**Figure S11.** (A) Structure of Methyl Green. (B) CD scans and (C) normalized CD melts of 5.4 μM 1Core with 2 and 4 eq. of Methyl Green.

**Neutral Red**

Neutral Red (NR) is phenazine dye that is known for intercalative binding to duplex DNA (Figure S12A). Modest signal decrease without overall change in the CD signal shape was observed in the CD scans as shown in Figure S12B. No stabilization was noticed in the CD melt; changes of 0.9 °C at 2 eq. and -0.2 °C at 4 eq. were measured (Figure S12C). Thus, we conclude that Neutral Red does not interact with 1Core.
Pyridostatin

Pyridostatin (PDS) is a molecule well-known for its ability to bind and stabilize GQ (Figure S13A).\textsuperscript{84–86} CD scans show modest signal decrease of 1Core in the presence of PDS (Figure S13B). No stabilization of 1Core was observed at 2 eq. (- T_{m} = -0.1 \, ^{\circ}\text{C}) and 4 eq. (- T_{m} = -0.8 \, ^{\circ}\text{C}) of PDS as shown in Figure S13C. The data suggest PDS unfolds 1Core yet maintains the stability of the structure. The data shows the interactions between 1Core and GQ are different as stabilization of 1Core would have occurred in the presence of PDS. These findings are not unsurprising, as 1Core has demonstrated unusual interactions with other GQ binders.

Figure 12. (A) Structure of Neutral Red. (B) CD scans and (C) normalized CD scans of 5.2 \, \mu\text{M} 1Core with 2 and 4 eq. of Neutral Red.

Figure S13. (A) Structure of PDS. (B) CD scans and (C) normalized CD scans of 4.9 \, \mu\text{M} 1Core with 2 and 4 eq. of PDS.
**Pentamidine Isethionate Salt**

Pentamidine Isethionate Salt is an antitrypanosomal drug and minor groove DNA binder (Figure S14A). CD scans show no change in the CD signature or signal amplitude in the presence of Pentamidine Isethionate Salt (Figure S14B). The ligand induced no stabilization of 1Core: $T_m$ equal -0.6 °C at 2 eq. and -1.1 °C at 4 eq. of ligand (Figure S14C). The results suggest Pentamidine Isethionate Salt does not bind to 1Core. This result is unsurprising as its binding has been demonstrated only for AT-rich DNA, similar to other minor groove binders presented in this study.87,88

![Figure S14](image)

**Figure S14.** (A) Structure of Pentamidine Isethionate Salt. (B) CD scan and (C) normalized CD melts of 4.9 µM 1Core with 2 and 4 eq. of Pentamidine Isethionate Salt.

**Pentacationic manganese (III) porphyrin**

Pentacationic manganese (III) porphyrin (PM3P) was reported to be a highly specific G-quadruplex ligand (Figure S15A).89 CD scans showed preservation of the secondary structure of 1Core with decrease in the CD signal intensity (Figure S15B). Virtually no stabilization was observed from the CD melting study: $T_m$ of -0.25 °C and -0.93 °C at 2 and 4 eq. respectively (Figure S15C). A clear isosbestic point was not observed in UV-vis titrations, suggesting either a complicated binding model or possible aggregation (Figure S15D). Two titrations of PM3P with 1Core were completed; the two
data sets yielded overall similar looking data but inconsistent binding models. One data set favored the 8:1 binding model for PM3P:DNA, while another favored the 5:1 binding model. The data suggests PM3P unfolds the DNA but does not destabilize it.

Figure S15. (A) Structure of PM3P. (B) CD scans and (C) normalized melting curves of 5.37 µM 1Core with 2 and 4 eq. of PM3P. (D) UV-Vis Titration of 5.4 µM PM3P with 106.6 µM 1Core. A total volume of 46 µL of 1Core was added to a 1.0 mL sample of PM3P. Concentration of ligand was kept constant throughout the titration by adding 5.4 µM PM3P to the 1Core aliquot. Data were zeroed from 675-700 nm.
Chapter 4: Small molecule microarray screen of 1Core+2

Figure S16. Fluorescent images for select SMM compounds from buffer only (no 1Core+2), 100K2Mg, and PBS trials. Side-by-side comparison of fluorescent images is shown for each of the two trials completed in each case. Red spots likely indicate 1Core+2 Cy5 binding. Green spots indicate auto fluorescence of SMM compounds.
Figure S17. Z-scores of 66 SMM compounds for treated (1Core+2 in PBS or 100K2Mg) and untreated (PBS alone) microarrays.

Figure S18. Hit profiles of top 13 compounds against the other oligos screened via SMM. Included is a description of the type of structure of each oligo.
The interactions of 1Core+2 with ligands identified from the SMM study were studied via CD. For most of the ligands, no changes in the CD signature were observed with the exception of L827-1823, which caused decrease in the signal amplitude (Figure S19A and B). These results match the stability studies, which show most ligands cause no stabilization besides L827-1823 (Figure S19C and D). The data suggest most of the SMM ligands can be categorized as non-interactive. L827-1823 has more pronounced interactions with 1Core+2 as observed in the CD signature change and slight increase in stability. Further structural studies should be pursued with 1Core+2 and analogs of L827-1823.
Chapter 5: Point mutant study of 1Core+2

CD melting studies of 1Core+2 mutants show variable structural responses to single mutations. As expected mutating core nucleotides destabilize the structure significantly (Figure S20B). Mutation of loop nucleotides enhanced the stability of the structure (Figure S20A and C). Mutations within the core replace structurally essential GC base pairs with noticeably weaker GT and CT wobble pairs. Computational studies by the Buck group showed CT mismatches centralized within a DNA duplex is destabilizing though mismatches at the end of the duplex provide minor stability to the
Our data corroborate these results showing any G\(\text{T}\) in the core of 1Core+2 is destabilizing, while G\(\text{T}\) within the loop or tails regions may provide minor stability.

Figure S20. CD melting curves of 1Core+2 single point mutants for (A) loop adenines, (B) core guanines and cytosines, and (C) loop guanines.

Studies of some mutants provided unclear melting transitions according to CD, which made stability determination elusive as shown in Figure S21A. These mutants include A7T, A19T, G6T, and G18T, which were all identified as very structurally important (evidenced by drastic change in CD signature compared to 1Core+2). G21T, a core mutant, also provided unclear thermal melting data via CD. The CD data suggest the mutants are relatively unstable which corroborate their CD signatures which lack prominent 1Core+2 character. UV melting studies of the specified mutants provided higher quality melting data, which made stability determination possible (Figure S21B): A7T (48.4 ± 0.3 °C), A19T (47.8 ± 0.8 °C), G6T (48.6 ± 0.3 °C), G18T (47.5 ± 0.3 °C), and G21T (37.2 ± 0.6 °C). While the average \(ST_m\) for the loop mutants compare to 1Core+2 is -2.4 °C, the structural change observed for these mutants is still significant and a better measure for the effects of mutations within the loop region.
Figure S21. (A) CD and (B) UV melting curves of the structurally important 1Core+2 loop mutants A7T, A19T, G6T, and G18T and core mutant G21T.

Chapter 5: Double mutant study of 1Core+2

The single point mutant study of 1Core+2 revealed an uneven contribution of the core guanines to the stability. We hypothesized that core guanines upon mutation can be replaced by neighboring guanines in a spare tire fashion (as shown with G tracts in GQ). We designed four double mutant constructs to test this spare tire mechanism: G2TG3C (G2 mutated to T and G3 mutated to C), G8CG9T, G14CG15T, G20CG21T. All of the double mutants show low amplitude, non-1Core+2 CD signatures with the exception of G2TG3T (Figure S22A). TDS shows G3CG3T and G14CG15T have signatures that match 1Core+2 while G8CG9T and G20CG21T have distinct signatures with a peak at 275 nm and shoulder at 260 nm (Figure S22B). CD melting of curves of the double mutants provided unclear melting transitions as shown in Figure S22C. UV melting studies provided melting data for the mutants (Figure S22D): G2CG3T (36.1 ± 0.9 °C), G8CG9T (50.0 ± 0.5 °C), G14CG15T (39.1 ± 3.8 °C), G20CG21T (51.0 ± 1.7 °C). This mirrors the single point mutant study, which showed mutations to guanines in the top
GCGC tetrad caused less destabilization compared to mutations to guanines in the bottom tetrad. Though melting data could be retrieved from the UV study, the melting curves were linear in nature so the data is less reliable. The data overall sheds light on the role of the neighboring guanines and their role in stability maintenance when core guanines are mutated. This is an interesting artifact of the CAGAGG repeat which may provide further evidence for why it is so problematic within cells. Residues of contiguous guanine runs may participate in formation of the core.

Figure S22. (A) CD scans, (B) TDS, (C) CD melts, and (D) UV melts of 1Core+2 double mutants.

Understanding the core plays a crucial role for 1Core+2, we tested the effect of varying the number of GCGC tetrads. The CD signature of 1GC1 displays a peak at 264 nm but the two shoulders at 275 and 290 nm are missing (Figure S23A). Meanwhile, the
signature of 1GC2 does not show the characteristic 264 nm peak of 1Core+2 yet maintains the two shoulders. Adding a GCGC tetrad to 1Core+2 either above or below results in a new signature, rather similar for both constructs with three peaks at 290, 277 and 268 nm peaks; the peak at 264 nm is lost. 3GCa and 1GC1 share the same TDS signature as 1Core+2, Figure S23C. The TDS signatures of 1GC2 and 3GCb are similar to each other but are different from that of 1Core+2: a peak at 267 nm is observed. The thermal stability of mutants was tested in CD melting studies and representative melting curves are shown in Figure S23B. Our results demonstrate that removal of bottom or top GCGC tetrad from 1Core+2 leads to 38.1 and to 21.8 °C destabilization, respectively. At the same time, addition of GCGC tetrad above or below the GCGC core of 1Core+2 leads to stabilization by 15.4 and 9.6 °C. One GCGC contributes to ~20°C of stability to the structure of 1Core+2.

Changing the number of tetrads greatly impact the fold and stability of 1Core+2. Addition or removal of a tetrad removes 264 nm peak which is characteristic for CAGAGG. 3GCa and 3GCb maintain a peak and two-shoulder signature similar to 1Core+2 yet the TDS of the two mutants vary significantly. These results suggest they are structurally similar but sensitive to the position of the tetrad; the farther the TDS peak is from 260 nm the less AG bp contribution is provided to the structure. Strikingly, both the top and bottom tetrads provide different structural features as suggested by the loss of either the shoulders or the peak in the CD signal for 1GC1 or 1GC2, respectively. TDS also shows a clear difference between the two single tetrad mutants where 1GC1 has more 262 nm character compared to 1GC2, which is indicative of structural contributions from AG bp. These results suggest the 264 nm peak can be attributed to AG bp.
Unsurprisingly, the number of GCGC tetrads of 1Core+2 is proportional to the structural stability, which likely due to more GC bp interactions. The unique native fold is highly sensitive to both single nt contribution and the number of GCGC tetrads and their placement within the structure.

Figure S23. (A) CD scans, (B) normalized representative CD melting curves monitored at 266 or 277, and (C) TDS of 1Core+2 tetrad mutants.

Table S2. Summary of stabilization data for all sequences: $T_m$, $\Delta T_m$, and hysteresis for 1Core+2 mutants.

<table>
<thead>
<tr>
<th>Name</th>
<th>$T_m$, °C</th>
<th>$\Delta T_m$, °C</th>
<th>Hysteresis, °C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1Core+2</td>
<td>50.5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>A1T</td>
<td>49.4</td>
<td>-0.7</td>
<td>3.7 ± 0.9</td>
</tr>
<tr>
<td>G2T</td>
<td>51.6</td>
<td>1.5</td>
<td>2.4 ± 0.7</td>
</tr>
<tr>
<td>G3T</td>
<td>37.5</td>
<td>-12.7</td>
<td>3.4 ± 0.4</td>
</tr>
<tr>
<td>C4T</td>
<td>24.0</td>
<td>-26.2</td>
<td>1.0 ± 0.9</td>
</tr>
<tr>
<td>A5T</td>
<td>55.7</td>
<td>5.6</td>
<td>2.6 ± 0.1</td>
</tr>
<tr>
<td>G6T</td>
<td>46.3</td>
<td>-3.8</td>
<td>2.0 ± 0.1</td>
</tr>
<tr>
<td>A7T</td>
<td>52.0</td>
<td>1.9</td>
<td>7.1 ± 3.3</td>
</tr>
<tr>
<td>G8T</td>
<td>55.6</td>
<td>5.5</td>
<td>2.7 ± 0.4</td>
</tr>
<tr>
<td>G9T</td>
<td>32.7</td>
<td>-17.5</td>
<td>2.9 ± 0.8</td>
</tr>
<tr>
<td>C10T</td>
<td>35.5</td>
<td>-14.6</td>
<td>2.2 ± 0.9</td>
</tr>
<tr>
<td>A11T</td>
<td>52.9</td>
<td>2.8</td>
<td>2.9 ± 0.1</td>
</tr>
<tr>
<td>G12T</td>
<td>49.8</td>
<td>-0.4</td>
<td>2.9 ± 0.5</td>
</tr>
<tr>
<td>A13T</td>
<td>49.8</td>
<td>-0.3</td>
<td>2.2 ± 0.5</td>
</tr>
<tr>
<td>G14T</td>
<td>49.5</td>
<td>-0.7</td>
<td>2.4 ± 0.2</td>
</tr>
<tr>
<td>G15T</td>
<td>40.1</td>
<td>-10.1</td>
<td>2.0 ± 0.7</td>
</tr>
<tr>
<td>C16T</td>
<td>32.0</td>
<td>-18.2</td>
<td>3.2 ± 2.2</td>
</tr>
<tr>
<td>A17T</td>
<td>56.0</td>
<td>5.9</td>
<td>2.4 ± 0.5</td>
</tr>
<tr>
<td>G18T</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>A19T</td>
<td>52.9</td>
<td>2.8</td>
<td>2.1 ± 0.8</td>
</tr>
<tr>
<td>G20T</td>
<td>54.6</td>
<td>4.5</td>
<td>1.7 ± 1.7</td>
</tr>
<tr>
<td>G21T</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>
Chapter 5: Three-dimensional melting studies of 1Core+2

Here the thermodynamics of melting process for CAGAGG is investigated. Previously, it was assumed that CAGAGG underwent a one-state melting transition (folded to unfolded). This assumption will be tested because CAGAGG is suggested to fold into a non-canonical assembly; the complicated features of the structure of CAGAGG shown by our results may cause a more complex melting transition.

Monochromatic thermal stability studies for CAGAGG are not sensitive to specific components of the melting transition such as intermediates.

To further explore other key species within the melting transition for CAGAGG, full wavelength CD melts were completed for 1Core. Using SVD, the number of significant spectral species can be elucidated from the full wavelength melting data. The number of significant species was determined using several criteria including the values for the autocorrelation functions of U and V matrices where species from each matrix with values higher than 0.8 were considered significant. A study of the three-dimensional melting transition of 1Core revealed the presence of three significant species within the melting process where the intermediate species is subtle yet real (see Figure S24B).
Figure S24. (A) Melting surface of 1Core and (B) autocorrelation of U and V matrices for 1Core. (C) Calculated CD spectra generated from the two-state transition model (F $\rightarrow$ I $\rightarrow$ U). This CD spectra is the product of the basis spectra (from SVD) and the weight of each significant species (from global fits). F is folded species, I is intermediate species, and U is unfolded species. (D) Kinetic eigenvectors (EV) derived from the results of fitting data to the two-state transition model.

While the melting process of 1Core has been shown to be complex, the actual mechanism behind its melting transition remains unknown. The 3D melting data was fit to 3D models corresponding to various mechanisms (two-state, three-state, and four-state transitions). The statistics for these fits are shown in Table S3. The models that best-fit the data for 1Core was the two-state transition and the parallel intermediate transition. These results coincide with the SVD analysis, which determined three species for the melting transition. Though the parallel intermediate transition was statistically plausible
from the fitting results, our previous native PAGE and AUC studies showed 1Core samples were a homogenous so this particular model does not fit this system With the two-state model considered, the next plausible step is to understand the identity of the intermediate species as demonstrated by the Chaires lab.\textsuperscript{27}

Table S3. Summary of statistical analysis of global fitting results for each model with 1Core CD data, showing Chi squared and AIC values. (Note: lower values for both Chi Squared and AIC correspond to better fits). Data averaged from three trials.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\chi^2$</th>
<th>AIC Value</th>
<th>$\chi^2*$</th>
<th>AIC Value*</th>
<th>$\Delta \chi^2$</th>
<th>$\Delta$AIC Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Transition</td>
<td>888</td>
<td>232</td>
<td>728</td>
<td>166</td>
<td>-161</td>
<td>-66</td>
</tr>
<tr>
<td>Two Transitions</td>
<td>362</td>
<td>-25</td>
<td>260</td>
<td>-176</td>
<td>-102</td>
<td>-151</td>
</tr>
<tr>
<td>Parallel Transitions</td>
<td>1220</td>
<td>288</td>
<td>1064</td>
<td>125</td>
<td>-156</td>
<td>-164</td>
</tr>
<tr>
<td>Three Transitions</td>
<td>330</td>
<td>-120</td>
<td>388</td>
<td>-172</td>
<td>58</td>
<td>-52</td>
</tr>
<tr>
<td>Parallel Intermediate Transitions</td>
<td>136</td>
<td>-228</td>
<td>113</td>
<td>-310</td>
<td>-23</td>
<td>-82</td>
</tr>
</tbody>
</table>

These studies also yielded CD spectra of the all the significant spectral species for each model tested. CD spectra for each significant species within a melting transition were generated by the basis spectra (determined by SVD) and the weight of each spectral species (determined by the global fits). Our goal here is to elucidate the identity of the intermediate species as this knowledge could shed light on the secondary structure of 1Core. One spectral signature was generated corresponding to the intermediate of the two-state melting transition for 1Core (Figure S24C). The spectra was very similar to that of 1Core besides the decreased amplitude, suggesting less base stacking in the intermediate structure compared to that of fully folded 1Core.
Longer more biologically relevant repeats of \((CAGAGG)_n:\)

2Core+2

Point mutation study of 2Core+2

The Brown lab determined CAGAGG repeats 80-100 times within the mouse genome. Because longer CAGAGG repeats may be more biologically relevant, we shift our focus to characterizing these larger structures. Our lab initially believed the folds of longer repeats (or multicore structures) are simply 1Core units connected by AGAG loops as exhibited in the beads-on-a-string (or beads) model (Figure S25A). Barrett Powell ’18 proposed an elongated structure of multicore CAGAGG that is constructed of AGAG tetrads sandwiched by GCGC cores as shown in Figure S25B. This model was supported by his AUC studies that showed the number of repeats or cores has a linear relationship to the axial ratio of the secondary structure (the fold increased in linearly in the vertical directions as more cores were added). Additionally, the NMR resolved structure of VK34 by the Plavec laboratory showed a tetrastranded assembly that contained both GCGC and AGAG tetrads.\textsuperscript{27}
There are currently two contending models for the fold of multicore CAGAGG. Multicore studies by Barrett Powell ’18 show that the secondary structure of CAGAGG is the same irrespective of the number of cores. Our current data does not elucidate which of the two models is more appropriate. These two models have noticeable structural differences. For instance, some of loop nt in the beads model and the elongated model occupy different environments (on the top or bottom of the structure). We strive to test these models and take advantage of the differences between them to resolve which one is
correct. This chapter will focus on further biophysical studies of the simplest of these multicore systems, 2Core+2.

The point mutant study for 1Core+2 (Chapter 5) showed that the bottom loops provided minimal contribution to the fold and stability while the top loops were structurally important. Here we investigate specific loop regions of 2Core+2 to test which model better fits our previous findings with 1Core+2. We surveyed loops 2, 4, and 6; in the beads model they are the bottom loops so they are expected to have minimal contribution to the structure. Conversely, loops 2 and 6 in the elongated 2Core+2 model are similar to 1Core+2 as they are essentially top loops while loop 4 is a bottom loop. If the elongated structure is the appropriate model, we expect the results to parallel those of 1Core+2 where the top loops are structurally essential and the bottom loop is not. The other loops are excluded from testing in this model as in either model these loops seem to play structurally important roles (top loops in the beads model or AGAG core in the elongated model). We designed 12 single N→T point mutants for loops 2, 4, and 6 (listed in Table 4) and tested them via PAGE, CD, and UV spectroscopies.

**Table S4. List of mutants and their sequences, including extinction coefficients for each oligonucleotide.**

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
<th>Extinction Coefficient (M⁻¹cm⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2Core+2</td>
<td>AGG CAGAGG CAGAGG CAGAGG CAGAGG CAGAGG CAGAGG CAGAGG CAGAGG CAGAGG CAGAGG CAGAGG CAG</td>
<td>504900</td>
</tr>
<tr>
<td>A11T</td>
<td>AGG CAGAGG CTGAGG CAGAGG CAGAGG CAGAGG CAGAGG CAGAGG CAGAGG CAG</td>
<td>499600</td>
</tr>
<tr>
<td>G12T</td>
<td>AGG CAGAGG CATAGG CAGAGG CAGAGG CAGAGG CAGAGG CAGAGG CAGAGG CAG</td>
<td>503700</td>
</tr>
<tr>
<td>A13T</td>
<td>AGG CAGAGG CAGTGG CAGAGG CAGAGG CAGAGG CAGAGG CAGAGG CAGAGG CAG</td>
<td>500400</td>
</tr>
<tr>
<td>G14T</td>
<td>AGG CAGAGG CAGATG CAGAGG CAGAGG CAGAGG CAGAGG CAGAGG CAGAGG CAG</td>
<td>502900</td>
</tr>
<tr>
<td>A23T</td>
<td>AGG CAGAGG CAGAGG CAGAGG CTGAGG CAGAGG CAGAGG CAGAGG CAGAGG CAG</td>
<td>499600</td>
</tr>
</tbody>
</table>
Results

Here we examined point mutants of 2Core+2 for their CD signature and stability to determine the effect of N→T mutation. In CD, all of the point mutants share the same signature as 2Core+2 but have varying degrees of CD signal amplitude. The data shows that nt in loop 2 and 6 caused a decrease in the CD signal amplitude when mutated (Figure S26). The decrease in CD signal amplitude for nt in loop 4 is minimal in comparison. Interestingly, mutations to the middle loop nt in loops 2 and 6 caused the most decrease in CD signal amplitude.
Figure S26. CD scans of in molar ellipticity of mutants in (A) loop 1, (B) loop 2, and (C) loop 3.

TDS results complement the CD findings (Figure S27). The data shows that loop 4 mutants have a nearly superimposable signature with 2Core+2 whereas loop 2 and 6 mutants have shifts in their TDS signatures relative to 2Core+2. CD melting studies showed that the stabilities of all point mutants are about the same as the stability of 2Core+2 (all within 3 °C), see Figure S27 and Table S5.

Figure S27. TDS of mutants in (A) loop 1, (B) loop 2, and (C) loop 3.
Figure S28. Normalized melting curves of mutants in (A) loop 1, (B) loop 2, and (C) loop 3. All melts were monitored at 264 nm.

Native PAGE studies were completed to determine the mobility and number of possible species in solution for each sample (see Figure S28). The results show that 2Core+2 and its loop mutants have rather similar mobility. Interestingly, G24 and G26T samples show the presence of dimers.

Figure S29. Native 12% PAGE gel of 2Core+2 (termed CA2Core+2 here) and its point mutants run in 1xTBE with 3 mM MgCl₂ with dT length markers included. All mutants
show very similar nobilities to that of 2Core+2 with well-defined bands. Mutants G24T and G26T seem to have presence of dimer.

Table S5. Summary of the melting studies of 2Core+2 and point mutants, describing stabilities, stability change, and hysteresis.

<table>
<thead>
<tr>
<th>Name</th>
<th>$T_m$ (°C)</th>
<th>$\Delta T_m$</th>
<th>Hysteresis</th>
</tr>
</thead>
<tbody>
<tr>
<td>2Core+2</td>
<td>54.4</td>
<td>--</td>
<td>3.4</td>
</tr>
<tr>
<td>A11T</td>
<td>56.6</td>
<td>2.2</td>
<td>2.8</td>
</tr>
<tr>
<td>G12T*</td>
<td>53.4</td>
<td>-1.4</td>
<td>2.5</td>
</tr>
<tr>
<td>A13T</td>
<td>52.7</td>
<td>-1.8</td>
<td>2.8</td>
</tr>
<tr>
<td>G14T</td>
<td>56.1</td>
<td>1.7</td>
<td>2.1</td>
</tr>
<tr>
<td>A23T</td>
<td>56.0</td>
<td>1.6</td>
<td>3.3</td>
</tr>
<tr>
<td>G24T</td>
<td>54.0</td>
<td>-0.5</td>
<td>2.8</td>
</tr>
<tr>
<td>A25T</td>
<td>53.7</td>
<td>-0.8</td>
<td>2.4</td>
</tr>
<tr>
<td>G26T</td>
<td>54.4</td>
<td>-0.1</td>
<td>0.9</td>
</tr>
<tr>
<td>A35T*</td>
<td>57.2</td>
<td>2.6</td>
<td>2.5</td>
</tr>
<tr>
<td>G36T</td>
<td>53.2</td>
<td>-1.2</td>
<td>3.4</td>
</tr>
<tr>
<td>A37T</td>
<td>52.2</td>
<td>-2.2</td>
<td>2.7</td>
</tr>
<tr>
<td>G38T</td>
<td>56.6</td>
<td>2.2</td>
<td>2.3</td>
</tr>
</tbody>
</table>

*Tm* retrieved from taking derivative. Stability for the other mutants determined by fitting melting curves.

CD and TDS studies show that the point mutants in loops 2 and 6 changed the degree of fold for 2Core+2. Strikingly, this effect is most notable for the middle nt in either of the two loops. Mutants from loop 4 show minimal structural deviation for 2Core+2 relative to the other mutants. These results mirrors biophysical studies for 1Core+2 point mutants that showed the two middle nt in loops 1 and 3 (top loops) caused the most structural change. This data suggest that the loops in 2Core+2 are structurally crucial which discredits the beads model based on the assumption loops 2, 4, and 6 (bottom loops in this model) are unimportant. The results also show loops 2 and 6 play similar to roles to the top loops in 1Core+2, which supports the elongated model where the two specified loops act as top loops. All of the mutations caused minimal structural change. This result may be explained by the larger assembly in 2Core+2, which is comprised of more structural elements compared to 1Core+2. PAGE results corroborate
the other findings understanding that PAGE is accounting for the globular structure and is less sensitive to the degree base stacking compared to CD. The presence of the small amount of dimers in G24T and G26T can be explained by since the mutations allow intermolecular AT bp interactions between loops of different monomer. These dimers may contribute to the CD signal amplitude observed for loop 4 mutants. Because no dimers were observed for the other loop mutants, perhaps these results show that loop 4 nt are free to engage in dimerization, while the nt in loops 2 and 6 are engaged in stronger structural interactions (such as stacking with a GCGC core unit or formation of AGAG tetrad as seen in two structures from Plavec’s laboratory which disallows participation in dimerization.48

The loop contributions to the secondary structure of 2core were tested with 12 N→T single-point mutants of 2Core+2. This study is consistent with similarly designed study of 1Core+2 and shows similarity in top-loop contributions to CAGAGG repeats. There is clear structural dependence on the selected loops (specifically 2 and 6), which would be unexpected according to the beads model. Ultimately, the data support the elongated core model.

**Fluorescence studies of 2Core+2 2-AP oligos**

To determine which of the two models for 2Core+2 is most probable, we investigate the fold, stability and FL response of a series of 2AP modified 2Core+2 in a manner similar to that described for 1Core+2 (Chapter 5, Figure 32) The 2-AP modifications sites in 2Core+2 match those used in the 1Core+2 FL study. This allows direct comparison between 1Core+2 and 2Core+2 and thus a reliable testing of the beads
model. If the 2Core+2 structure adopts the beads model which assumes that 1Core units are independent, we anticipate the same FL response pattern for the five adenine sites as observed in the 1Core+2 case. In the elongated model, 2-5, 2-7, and 2-19 make up the core of the structure, comprised of suggested stacked AGAG tetrads and thus these oligos are expected to produce lower FL response. Oligos 2-13 and 2-23 are located in one of the top and bottom loops, respectively. Bottom loop is expected to be completely flexible, just like the case for 1Core+2, and thus 2-23 is expected to produce the highest FL response.

**Figure S30.** (Left) Beads-on-a-string model and (right) elongated model of 2Core+2 with 2AP-modified adenine sites indicated and numbered.

The fold and stability of 2-AP modified 2Core+2 oligos were studied using CD and UV spectroscopies. CD scans of the oligos show very minimal changes, with slightly reduced amplitude, **Figure S31A.** TDS data shows the 2-AP-modified oligos have similar
but red shifted spectra with respect to 2Core+2 (Figure S31B). Finally, all 2-AP modified oligos have Tm within experimental error (Tm of the 2-5, 2-7, 2-13, 2-19, and 2-23 are 52.9 ± 0.5 °C, 52.4 ± 0.5 °C, 52.1 ± 0.5 °C, 52.0 ± 0.5 °C, 53.4 ± 0.5 °C, respectively), Figure S31C. 2AP oligos were also studied using denaturing and native PAGE where they displayed the same mobility, Figure S32. Combined, biophysical data suggest that 2-AP had minimal effect on the structure and stability of 2Core+2. The fluorescence of all the 2-AP mutants of 2Core+2 is displayed in Figure S31 and increases in the following order: 2-5 < 2-13 < 2-7 ≈ 2-23 ≈ 2-19.

Figure S31. Biophysical characterization of 2Core+2 and 2-AP oligos. (A) CD scans, (B) TDS, and (C) normalized CD melting curves monitored at 264 nm.
Figure S32. Characterization of 2Core+2 and 2-AP oligos with PAGE. (A) Native 12% PAGE gel run in 1× TBE buffer with 3 mM MgCl₂ (B) Denaturing 12% PAGE gel run in 1× TBE buffer with 3 mM MgCl₂. 12% PAGE gel run in 1× TBE buffer with 3 mM MgCl₂. 

Note, 2-5 includes 1Core+2, and all of 2-AP analogs. Thus, the lowest FL intensity of these oligos is the exception. The oligos that are faster than expected, suggesting a higher degree of structuration amongst 2Core+2 and 2-AP oligos. All of the oligos in the native PAGE have mobilities that correspond to ~30 nt.
exposed to the solvent and most likely are not particularly strongly engaged in either
base-pairing or base-stacking.

Figure S33. Characterization of 2Core+2 2-AP oligos with FL. (A) Emissions scans
and (B) summary of fluorescence responses monitored at 365 nm (with error bars
included) for 1Core+2 (gray) and 2Core+2 (red) 2-AP oligos (n=3). Data were collected
at 0.1 Abs of oligonucleotides in at 4 °C.

This investigation ultimately did not provide a clear answer as to which of the two
contending models is more relevant. This study might have also revealed that our
assumptions regarding 2Core+2 in context of either of the model may need to be refined.
Compared to 1Core+2, the sensitivity of 2Core+2 to single adenine substitutions
decreased immensely (shown by both structural and stability data). This might suggest
that overall the 2Core+2 structure is more robust as compared to 1Core+2 structure.

Three-dimensional melting studies of 2Core

A study of the three-dimensional melting transition of 2Core revealed the
presence of three significant species within the melting process (see Figure S24B). The
statistics for these fits are shown in Table S6. The models that best-fit the data for 2Core
were the three-state transition and the parallel intermediate transition. These results match
the SVD analysis, which determined three species for the melting transition. Though the parallel intermediate transition was statistically plausible from the fitting results, our native PAGE and AUC studies showed 2Core samples were homogenous. This parallel intermediate transition is therefore inappropriate for this system.

Focusing on the three-state transition model, we identified a possible intermediate species. Interestingly, the generated CD spectrum for the intermediate species shows a peak at 270 nm and trough at 250 nm, indicative of a duplex. Other observed intermediate species for 2Core resembles that of 1Core which is a simply a less folded structure. The difference between the results of the 3D melting studies for 1Core and 2Core suggest structural dissimilarity, supporting the elongated model for 2Core. This result corroborates data from the point mutation and 2-AP studies of 2Core. One possible explanation for the duplex intermediate for the 2Core melting transition is the elongated unfolds into two hairpins connected by a single AGAG loop prior to its transition to the unfolded state. It is interesting to note that Plavec proposed a similar folding transition for VK34.48
Figure S34. (A) Melting surface of 2Core and (B) autocorrelation of U and V matrices for 1Core. (C) Calculated CD spectra generated from the two-state transition model (F I U). This CD spectra is the product of the basis spectra (from SVD) and the weight of each significant species (from global fits). F is folded species, I is intermediate species, and U is unfolded species. (D) Kinetic EV derived from the results of fitting data to the two-state transition model.

Table S6. Summary of statistical analysis of global fitting results for each model with 1Core CD data, showing Chi squared and AIC values. (Note: lower values for both Chi Squared and AIC correspond to better fits). Data averaged from three trials.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\chi^2$</th>
<th>AIC Value</th>
<th>$\chi^2$*</th>
<th>AIC Value*</th>
<th>$\gamma \chi^2$</th>
<th>$\gamma$ AIC Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Transition</td>
<td>2209</td>
<td>629</td>
<td>2127</td>
<td>615</td>
<td>-82</td>
<td>-14</td>
</tr>
<tr>
<td>Two Transitions</td>
<td>270</td>
<td>-39</td>
<td>193</td>
<td>-166</td>
<td>-76</td>
<td>-127</td>
</tr>
<tr>
<td>Parallel Transitions</td>
<td>1988</td>
<td>477</td>
<td>2127</td>
<td>627</td>
<td>139</td>
<td>150</td>
</tr>
<tr>
<td>Three Transitions</td>
<td>121</td>
<td>-291</td>
<td>74</td>
<td>-439</td>
<td>-47</td>
<td>-148</td>
</tr>
<tr>
<td>Parallel Intermediate</td>
<td>116</td>
<td>-293</td>
<td>88</td>
<td>-389</td>
<td>-28</td>
<td>-97</td>
</tr>
</tbody>
</table>
Next Steps/To-Do:

- AUC for 1Core+2 Ln mutants (only one trial conducted for L0 and L1)
- Test 1Core+2 L5 and/or L7 mutants
- Test effect of Co$^{2+}$ on CAGAGG
- Fresh CD scans of L0-L8 (for monitoring change in 265 nm vs. loop length)
- UV titrations of MB with 1Core
- CD studies of 2Core+2 with L827-1823
- Test 1Core with structural analogs of L827-1823
- EMSA experiment with 2Core+2 and RPA
- FL acrylamide titrations for 2AP analogs of 1Core+2 and 2Core+2
- $^{31}$P NMR on 1Core+2 (looking at backbone angles, especially for top loops)
- 2Core+2 TTTT loop mutants (should have larger effect than single N$\rightarrow$T mutation)
- Gel testing tetrahelical nature of CAGAGG$^{20}$
- Further work on intermediate in 2Core unfolding (is it a duplex?)
- Produce higher quality crystals of 2Core
- Re-do 2Core+2 loop mutant gel
- Need at least one more trial of UV/CD studies for 1Core+2 RS variants
Copy of published work

“Genome-wide Identification of Structure-Forming Repeats as Principal Sites of Fork Collapse upon ATR Inhibition”

Reference:

SUMMARY

DNA polymerase stalling activates the ATR checkpoint kinase, which in turn suppresses fork collapse and breakage. Herein, we describe use of ATR inhibition (ATRi) as a means to identify genomic sites of problematic DNA replication in murine and human cells. Over 500 high-resolution ATR-dependent sites were ascertained using two distinct methods: replication protein A (RPA)-chromatin immunoprecipitation (ChIP) and breaks identified by TdT labeling (BrITL). The genomic feature most strongly associated with ATR dependence was repetitive DNA that exhibited high structure-forming potential. Repeats most reliant on ATR for stability included structure-forming microsatellites, inverted retroelement repeats, and quasi-palindromic AT-rich repeats. Notably, these distinct categories of repeats differed in the structures they formed and their ability to stimulate RPA accumulation and breakage, implying that the causes and character of replication fork collapse under ATR inhibition can vary in a DNA-structure-specific manner. Collectively, these studies identify key sources of endogenous replication stress that rely on ATR for stability.

INTRODUCTION

Numerous sensory and repair networks are required to safeguard genome integrity during duplication. These networks operate when progressing replication forks encounter impediments, such as damaged bases and difficult-to-replicate sequences. Replication protein A (RPA) is an immediate responder to such events by protecting the unwound template strands. RPA-coated single-stranded DNA (ssDNA) recruits and activates numerous DNA repair and cell cycle checkpoint regulators, including the ATR checkpoint kinase. ATR activation at stalled replication forks stabilizes these intermediates and prevents progression into M phase (Saldivar et al., 2017). Although steady progress has been made in defining the signal transduction cascades that regulate ATR and operate downstream of it, substantially less is known about the genomic sequences that promote fork stalling and make ATR an essential gene.

Although it has long been known that ATR loss of function is sufficient to cause chromosome breaks (Brown and Baltimore, 2000), pinpointing the sites of genomic breakage has primarily relied on candidate approaches. For example, ATR suppression increases chromosome breakage at common fragile sites (CFSs), which were originally defined as breaks observed in mitotic cells following partial DNA synthesis inhibition (Casper et al., 2002; Glover et al., 2017). Other candidate genomic features examined for ATR dependence include expanded microsatellite repeats. CAG/CTG trinucleotide repeats, the causative feature in Huntington’s disease and myotonic dystrophy, form non-B DNA structures (Mirkin, 2007; Neil et al., 2017) and are stabilized by the ATR ortholog, Mec1, in S. cerevisae (Lahiri et al., 2004). A similar dependence has been reported for expanded CGG/CCG repeats, which causes Fragile X syndrome (Entezam and Usdin, 2008). Finally, telomere repeats are reliant on ATR for stability, putatively through replication perturbation at G-quadruplexes (Johnson et al., 2008; Sfeir et al., 2009). Nevertheless, these repeat sequences were examined as candidates for dependence on ATR based on their proven or speculated inhibitory effects on DNA replication, not through an unbiased screen for sequences that are the most reliant on ATR for stability.
Herein, we use ATR inhibition and two distinct methods, RPA-ChIP and BrITL, to detect replication fork collapse. Using this approach, over 500 sites of problematic replication across the mouse and human genomes were defined. Although numerous genomic and transcriptional features have been proposed to cause a dependence on ATR, the elements most commonly associated with fork collapse were microsatellite and inverted repeats that form stable secondary structures. Notably, the repeats identified as most reliant on ATR for stability are distinct from those previously examined and include microsatellite repeats that form purine-rich non-B DNA structures and hairpin-forming inverted retrotransposable elements and quasi-palindromic AT-rich minisatellite repeats. Notably, the type of structure formed by these repeats impacts their propensity to cause breakage and accumulate RPA, indicating that the character of fork collapse differs in a repeat-specific manner. These observations have identified both new sites of problematic DNA replication that rely on ATR for stability and distinct categories of replication fork collapse that are delineated by the DNA structure formed. These findings indicate that such repeats may comprise part of the mechanism of action of ATR inhibitors, which have entered clinical trials as cancer therapeutics.

RESULTS

Genome-wide Identification of RPA-Enriched Sites following ATR Inhibition

RPA accumulates at sites of replication fork collapse (Figure 1A). To identify such sites following ATR inhibition, RPA-ChIP sequencing (RPA-ChIP-seq) (Yamane et al., 2013) was performed on passage-immortalized Atrflox/C0 Cre-ERT2+/mouse embryonic fibroblasts (Ruzankina et al., 2007; Smith et al., 2009) that were either treated or left untreated with ATR inhibitor and a partially inhibitory concentration of the DNA polymerase antagonist, aphidicolin (0.2 μM), as a fork stalling enhancer (ATRi+aph18hrs). As shown (Figure 1B), chromatin was sonicated to relatively large fragment sizes (500–2,000 bp) prior to RPA-ChIP retrievals. This approach facilitated mapping of sites that contained repetitive DNA by co-retrieving adjacent unique sequences. Isolated DNA was subsequently re-sonicated (200–300 bp), subjected to next generation sequencing (NGS), and mapped to the reference genome (Figure 1B). Several measures were taken to assure validity of ATR-dependent sites, including (1) normalization of retrievals by genomic representation (pre-ChIP inputs), (2) selection of sites with a minimum read enrichment of >4-fold over normalized input and a p value of <10−20, and (3) site identification in two independent biological replicates. Peaks called in both biological replicates and not observed in DMSO controls were defined as ATR dependent.

A total of 168 sites of significant and specific RPA enrichment were identified in the ATRi+aph18hrs condition with an average peak signal intensity of 13.6-fold over background (Figures 1C and 1D). As expected, the average peak footprint (~1.5 kb) was similar to the size of the sonicated chromatin fragments used for RPA-ChIP. In aggregate, these peaks comprised a small fraction of the total murine genome (10−6). These metrics indicated a high degree of resolution and specificity in ATR-dependent site identification.

To expand and prioritize ATR-dependent sites, three additional ATR inhibition (ATRi) conditions were examined by RPA-ChIP: (1) 9-hr ATRi + aph treatment (ATRi+aph18hrs); (2) ATRi treatment alone (ATRi18hrs); and (3) ATRi combined with suppression of the replication fork protection complex factor, TIMELESS (ATRi+shTIM1; Smith et al., 2009). Notably, of the 87 sites observed using these additional conditions, 84 (97%) overlapped with those identified with ATRi+aph18hrs treatment (Figures 1D and 1E; Table S1), and co-occurrence correlated well with RPA signal intensity. The 29 sites found in all 4 conditions (ATRi29) exhibited an average signal intensity that was 22.5-fold over background in the ATRi+aph18hrs condition. The 56 sites found in just 2 or 3 conditions (ATRi56) and those found exclusively with ATRi+aph18hrs exhibited a downward trend in signal intensity (13.3-fold and 10.8-fold over background, respectively; Figure 1D; Tables S1 and S2). RPA-ChIP-seq using ATR-deleted cells also preferentially identified these higher-priority ATRi29 and ATRi56 sites (Table S2). The identification of these RPA-accumulation sites, 171 in total (ATRi71), and their prioritization by signal intensity and common occurrence provided a means to investigate the dominant causes of replication fork collapse following ATR inhibition. Because these sites met two key criteria of defective replication—RPA accumulation and dependence on ATR for stability—they were collectively deemed to be replication perturbed locations (RPLs).

RPLs and Large-Scale Genomic Features

To investigate the mechanisms underlying dependence on ATR, RPLs were examined for their overlap with large-scale genomic characteristics: chromatin state (euchromatic, heterochromatic, and boundary elements); gene location (transcription start sites and gene bodies); and sequence elements (transcription factor binding sites and repetitive sequences). RPL sites were found outside of gene bodies, promoters, and terminators at a frequency that roughly mirrored the aggregate percentage of

Figure 1. Genome-wide Identification of Fork Collapse Sites by RPA-ChIP-Seq

(A) RPA-ChIP-seq detection of replication fork collapse from ATR inhibition (ATRi) (red diamond).
(B) Schematic of RPA-ChIP-seq experimental approach. Cross-linked chromatin was sonicated into large fragments (1.5 kb average) prior to immunoprecipitation with RPA2 antibody. Retrieved DNA was sonicated into smaller fragments (200–300 bp) for NGS.
(C) RPLs identified in the mouse genome (red marks).
(D) Venn diagram depicting overlap of peaks identified from different conditions.
(E) Representative peaks in RPA-ChIP-seq coverage and ratio tracks (ATRi+aph18hrs and DMSO-treated control; UT). Symbols above select peaks indicate identification under additional experimental conditions.

See also Tables S1 and S2.
Figure 2. Short Tandem Repeats Are Enriched in RPLs

(A) Genomic features associated with RPLs. Percent and number of RPLs overlapping with noted features and the expected overlap based on fraction of the genome comprised of these features are shown. Statistical significance (p value) was calculated by permutation test.

(B) Example of repetitive DNA in RPL peaks. Top track: Representative ratio track of RPA-ChIP Seq reads over input reads from ATRi+aph18hrs-treated cells. First track: RPA-ChIP of ATRi+aph18hrs; second track: input of ATRi+aph18hrs; third track: RPA-ChIP of DMSO-treated control (UT). Bottom: RepeatMasker annotations of repetitive elements within the peak region. Arrows detail examples of repetitive elements present. Middle and bottom track: Zoomed-in RPL peak. First track: RPA-ChIP of ATRi+aph18hrs; second track: RPA-ChIP of DMSO-treated control (UT). Bottom: RepeatMasker annotations of repetitive elements within the peak region. Statistical significance (p value) was calculated by permutation test.

non-coding DNA in the mammalian genome (Figure 2A). Similarly, there was no significant correlation with the transcription state of coding genes, as determined by co-incidence with H3K4<sup>me3</sup> and H3K27ac (Figure 2A). Collectively, these data indicate that the transcriptional state of coding genes does not strongly influence the localization of RPLs.

Although statistically significant, overlap of RPLs with early replicating fragile sites (ERFSs) (Barlow et al., 2013) and replication timing transition regions (TTRs) (Yue et al., 2014) was minimal when considering their broad genomic coverage (Figure 2A). Indeed, only 22 out of the 171 RPLs (12.9%) were found within ERFSs, with 8 expected from random co-occurrence (Figure 2A). No sub-peaks within the broad ERFS footprints (208 kb on average) were observed to align with overlapping RPLs (Figure S1). Similarly, the greater-than-expected co-occurrence of RPLs with TTRs was relatively small given the genomic coverage.
of TTRs (Figure 2A). However, upon querying the association of RPLs with known characteristics of TTRs (Yue et al., 2014), one TTR marker in particular, CTCF binding, showed significant overlap with RPLs, suggesting it as the underlying cause of TTR association (Figure 2A).

CTCF, which functions in transcription, chromatin anchoring, and boundaries between topologically associated domains (TADs), is enriched in the early replicating and midpoint regions of TTRs (Pope et al., 2014; Yue et al., 2014). Notably, 24 of the 171 RPL sites were superimposable with previously identified CTCF-binding sites in mouse embryonic fibroblasts (MEFs) (Figures 2A and S2A), and this association was highly significant (p value < 10^-12). Of the 24 CTCF-associated RPL sites, 10 (42%) were within TTRs, which could account for the greater-than-expected association of RPLs with TTRs. Although a small fraction of RPLs were strongly associated with CTCF binding sites, >99.9% of CTCF binding sites throughout the genome do not overlap with RPLs, indicating that CTCF binding is not sufficient to cause dependence on ATR for stability.

In a search for possible determinants within these CTCF sites, we noted that 21 of the 24 RPLs that overlapped with CTCF binding were characterized by centrally located microsatellite repeat sequences that did not fit the typical consensus for CTCF binding, specifically, (CACAG)n and (CAGAGG)n, and similar repeats (Figure S2A). Notably, most CTCF-associated RPLs that contained repeats were frequently observed under multiple ATRi conditions (5 in the ATRi subgroup and 12 in the ATRi subgroup), indicating that such repeats correlated well with RPA-ChIP detection. Collectively, the association of RPLs with these repeat-containing CTCF sites was stronger than association with any other large-scale chromatin feature.

An ATR-TIMELESS-Dependent RPL in rDNA Is Linked to Microsatellites

Previous studies have demonstrated an important role for ATR in fork stability following TIMELESS suppression (Smith et al., 2009), a finding in accord with TIMELESS limiting ssDNA formation at replication forks (Katou et al., 2003; Smith et al., 2009). Consistent with these studies, fork collapse sites identified by ATRi+shTIM1 and ATRi+aph18hrs overlapped significantly (Figure 1; Table S1). Notably, deletion of the TIMELESS orthologs TOF1 and SWI1 in S. cerevisiae and S. pombe, respectively, abrogates replication fork barrier function at the rRNA transcriptional terminator, FOB1, an effect that has been proposed to increase genomic instability through replication-transcription conflicts (Krings and Bastia, 2004; Mohanty et al., 2006).

Although no enrichment of RPA was observed in the rRNA transcription unit, one RPL was found in the intergenic spacer region of the 45-kb rDNA repeat, and its signal intensity was 10-fold greater in the ATRi+shTIM1 condition than in any other ATRi condition (Figure S2B). Once again, this rDNA-associated RPL was characterized by a centrally located microsatellite, (GAAATTTTCC), which can form H-DNA (Follieron et al., 2013; Mirkin, 2007) or stem-loop structures upon pairing with the extensive TTT-rich repeats downstream. These data are consistent with previous findings indicating that TIMELESS orthologs in yeast counteract replication fork stalling and fragility at structure-forming repeats (Voineagu et al., 2008; Zhang et al., 2012) and suggest a specific role for TIMELESS in fostering replication through this type of microsatellite. Collectively, the localization of RPLs correlated better with repetitive DNA than any other genomic feature analyzed.

Strong Association of RPLs with Microsatellite Repeats

The association of CTCF- and rDNA-overlapping RPL sites with microsatellite repeats suggested that repetitive DNA may be the main mechanism underlying ATR dependence at these sites. As shown in Figure 2B, a high frequency of various repetitive sequences were observed within RPA-ChIP-seq enrichment sites, with microsatellite repeats frequently observed near the peak center (Figure 2B). These microsatellite regions were often characterized by a gap in NGS read accumulation due to standard filtering of reads that align to multiple genomic loci (Figure 2B; “CAGAG” repeat). However, the position of microsatellite repeats at the center of RPL peaks suggested that such features might strongly contribute to replication fork collapse upon ATR inhibition.

To determine which repeats found within RPL peaks most consistently associated with RPA accumulation, the number of reads containing these sequences was quantified in RPA-ChIP retrievals relative to the number in sequenced input DNA. Consistent with prior studies in yeast (Admire et al., 2006; Sziard et al., 2010), rDNA-encoding sequences were slightly enriched in RPA-ChIP retrievals from ATR-inhibited cells (Figure S3A). However, other well-known satellite repeats and retrotransposable elements on the whole were not enriched following ATR inhibition (Figure S3A). Although other factors might influence the stability of a subset of these transposable elements, such as rare inverted orientations, these elements did not appear to be sufficient to drive RPA accumulation in ATR-inhibited cells.

The microsatellite repeats found within RPA-accumulation regions were also queried for enrichment in RPA-ChIP retrievals. To do so, we developed a repeat-counting program called REQer (repeat enrichment quantifier) that identifies the number of tandem and aggregate repeat units within NGS reads. This approach is more accurate than quantifying repeats from the reference genome, which is subject to polymorphisms in repeat length and incomplete genomic assemblies. Using REQer, reads containing high numbers of repeat units were counted in RPA-ChIP retrievals and normalized by similar quantifications using sequenced input DNA (Figures 2C–2E).

As expected, unexpanded trinucleotide microsatellites were not significantly enriched in ATRi+aph18hrs RPA-ChIP reads (Figures 2E and S3B), and telomere repeats (TTAGGG)n, were enriched approximately 2.5-fold (Figures 2E and S3B). Strikingly, the microsatellites that exhibited greatest enrichment following ATR inhibition had not previously been reported as difficult to replicate (Figures 2C–2E and S3C). These repeats included the hexameric and pentameric microsatellites (CAGAGG/CCTCTG)n and (CAG/CTG)n with other variants of CAGAGG such as (CAGAGG/CCTCTG in.) (CAGAGG/CCTCTG)n and (CAG/CTG)n, were also enriched in RPLs. Notably, the latter of these, (CAG/CTG)n, is associated with myotonic dystrophy type 2 (Liquori et al., 2001). Most repeats enriched in RPLs occurred in tandem iterations without interruptions, with the exception of (CAGAGG) in, which showed a precipitous
The Microsatellite Repeats Observed in RPLs Form Stable Intrastrand Structures

We hypothesized that RPL-associated microsatellite repeats form stable intramolecular secondary structures, which in turn could limit replicative polymerase progression. To test this hypothesis, synthetic single-stranded oligos of RPL repeats (Figure 2; Table S3) were first examined for increased mobility during native PAGE after a single heating and cooling cycle (Figure 3A). With the exception of (CAGAGT)15, a repeat variant found imbedded within (CAGAGG)n repeats, repeat-containing oligos migrated as single-dominant bands with mobilities greater than expected from their lengths (Figure 3A), consistent with the formation of compact, uniquely folded structures. Notably, (CAGGG)22, (CAGAGG)15, and (CAGAGGAGG)11 exhibited greater electrophoretic mobility than oligos encoding their complementary strands (Figure 3A), suggesting that the purine-rich strands of these repeats have more structure-forming potential than their pyrimidine-rich complements.

To determine the strength and types of secondary structures formed, repeat-containing oligos were examined by circular dichroism (CD), UV-vis spectrometry, and thermal difference spectra (TDS). By CD, the oligos that exhibited the highest molar ellipticity, a proxy for the extent of DNA folding, were (CAGAGG)15 (Δε = 1,100 ± 70 M⁻¹cm⁻¹), (CAGAGGAGG)11 (Δε = 730 ± 70 M⁻¹cm⁻¹), and (CACAG)18 (Δε = 650 ± 40 M⁻¹cm⁻¹; Figures 3B and S4; Table S3). These values and the reversibility of melting transitions (Figure S4) supported the formation of unimolecular (intrastrand) structures. These purine-rich repeats also exhibited high thermal stabilities (Figures 3B, 3C, and S4; Table S3). However, consistent with PAGE analysis, oligos encoding the pyrimidine-rich complementary strand of these repeats were neither well-folded nor stable (Figures 3A–3C; Table S3), again indicating that the structure-forming potential of these RPL repeats resides primarily in their purine-rich strands.

(CAGAGG)n structure formation was further characterized as it was the most frequently observed RPL-associated repeat (Figures 2C–2F). These studies demonstrated that five tandem units of CAGAGG were sufficient to generate stable-secondary structures, with structure formation, CD signature, and stability remaining relatively unaffected by additional units of up to 15 (Figures 3D–3G). The unimolecular nature of the (CAGAGG)n structure was further supported by analytical ultracentrifugation (not shown) and identical CD spectral and thermal stability measurements across a >20-fold concentration range (Figures 3H and 3I). Notably, the CD signature of (CAGAGG)n, which includes a major peak at ~260 nm and two prominent shoulders at 276 and 291 nm, is not consistent with either B-form or G-quadruplex DNA, indicating a potentially novel secondary structure. Collectively, these results demonstrate that purine-rich strands of RPL microsatellites form intranarrow secondary structures that could impact DNA replication.

RPL-Associated Microsatellite Impedes DNA Replication

We next investigated whether the structure-forming repeat most commonly observed in RPLs, (CAGAGG)n, is sufficient to slow DNA replication in vitro and in cultured cells. For in vitro studies, purine-rich (CAGAGG)15, the complementary pyrimidine-rich (CCTCTG)15, and scrambled control sequences were examined for their ability to impede progression of the DNA polymerase δ holoenzyme (Polδ-PCNA-RFC; Pol δHE). Pol δHE synthesis pausing was quantified as accumulations of reaction products relative to the inserted sequences: far upstream (68–11 nt); immediately upstream (10–1 nt); within the insert; and downstream (3’) of the insert (Figure 4A; Hile and Eckert, 2004, 2008).

Whereas DNA synthesis from the (CCTCTG)15 pyrimidine-rich template occurred without pausing either before or within the repeat, a substantial accumulation of reaction products was observed immediately upstream of the purine-rich (CAGAGG)15 template (Figure 4B). No pause sites were observed within the repeat region itself, indicating that polymerase blockage at the repeat interface was substantial (Figure 4B). Termination of the Pol δHE complex immediately upstream of the (CAGAGG)15 insert was 22-fold greater than that observed upstream of the complementary (CCTCTG)15 repeat and 9- to 12-fold greater than that of the purine-rich scrambled controls (Figure 4C; p < 0.0001; two-way ANOVA). Notably, no increase in Pol δHE termination was observed within the repetitive (CAGAGG)15 inserts themselves (Figure 4C), indicating that DNA replication was not slowed randomly throughout the repeat region but rather was barred specifically at the start of the folded structure. In contrast, low-level sequence-specific Occasional Pol δHE pausing was observed within the purine-rich scrambled insert (Figures 4B and S5A), as expected based on nucleotide content (Walsh et al., 2013). Importantly, these data indicate that (CAGAGG)15 repeats are fundamentally different from other microsatellite sequences, which cause replicative polymerase pausing within the repeat sequence, not at the start of it (Hile and Eckert, 2004, 2008; Walsh et al., 2013).

To determine whether the duplex (CAGAGG/CCTCTG)n repeat impedes DNA replication in cells, a 105-unit repeat was amplified from an endogenous RPL (Chr7:35159697–35161220; mm10) and subcloned into the pML113 SV40 plasmid replication system (Follonier et al., 2013) at two distinct sites (Figures 4D, 4E, and S5B–S5D). Vectors containing scrambled (CAGAGG/CCTCTG)n synthetic sequences of similar length were also generated as controls. These plasmids were replicated in large T antigen-expressing human osteosarcoma cells (U2OS), and...
Figure 3. RPL-Associated Repeats Form Unique Intrastrand Secondary Structures

(A) Non-denaturing PAGE gel of oligonucleotide repeats.
(B) CD molar ellipticity peak values as a proxy for DNA folding. Melting temperatures, shown above bars, were obtained using a non-linear fit assuming two-state system. (*) indicates melting transition is characterized by a non-sigmoidal melting curve.
(C) List of simple repeat sequences analyzed and respective melting temperatures.
(D) Repeat-normalized CD wavelength scans of (CAGAGG)n, with n = 2, 4, 5, 6, 10, and 15.
(E) Representative CD melting and cooling curves for (CAGAGG)n, n = 4, 5, 6, and 10.
(F) Graph of melting temperatures obtained by UV-vis with different CAGAGG monomer lengths. Tm values and change in enthalpy are summarized in the embedded table.
(G) Non-denaturing PAGE gel of (CAGAGG)n; n = 5, 10, and 15.
(H) Overlay of (H) normalized UV-vis melting and (I) CD scans (4°C) of (CAGAGG)10 at varying oligonucleotide concentrations.

For (B), the data are represented as mean ± SEM. For (A)–(I), all samples were prepared in 10 mM lithium cacodylate (pH 7.2), 100 mM KCl, and 2 mM MgCl2 buffer. For (A) and (G), DNA bands were visualized with Stains All. See also Figure S4 and Table S3.
Figure 4. CAGAGG Repeats Impede DNA Synthesis

(A) Schematic of in vitro Pol dHE primer-extension assay.

(B) Representative images of Pol dHE reaction products. Pol dHE DNA synthesis products from ssDNA templates containing (CAGAGG)15, (CCTCTG)15, or scrambled control inserts (purine-rich or pyrimidine-rich) with increasing reaction times (3–15 min, triangle) were separated by denaturing PAGE alongside a dideoxynucleotide sequencing of the same template (TACG). Left: (CCTCTG)15 and (CAGAGG)15 insert-containing templates are shown; right: two distinct purine-rich scrambled control insert-containing templates are shown, with numbers indicating random sequences of weak structure-forming potential. Control lanes are indicated (Hyb, primer-template hybridization; /C0 Pol, no polymerase). Percent extension, extended DNA over extended plus unextended primer-bound DNA. Also see Figure S5A for pyrimidine-rich scrambled control.

(C) Pol dHE termination probability. Termination probability, normalized by the number of nucleotides in each region, was quantified as the ratio of DNA molecules within a specific region over these plus all longer DNA molecules.

(D) Effect of (CAGAGG)n repeats on plasmid DNA synthesis in cells. Left: (CAGAGG)105 or a scrambled sequence of the same nucleotide composition and length (SCR) was inserted proximal to the bidirectional SV40 origin (triangles); SV40 large T-antigen (TAg) (Follonier et al., 2013). Right: representative 2D gels are shown. Plasmid-transfected cells were either untreated (UT) or treated with 0.6 μM aphidicolin (APH) for 24 hr. Isolated episomal DNA was digested with DpnI, EcoRI (RI), and Eco NI (NI), and replication intermediates were resolved by 2D neutral-neutral gel electrophoresis with Southern hybridization to the indicated probe. Arrows denote the point of divergence of the double-Y structure from the simple-Y arc.

(legend continued on next page)
Dpn I-resistant replication intermediates were resolved by neutral-neutral 2D gel electrophoresis.

Normal DNA replication through the scramble control insert was observed both with and without low-dose aph treatment (Figure 4D). However, replication of the (CAGAGG/CCTCTG)\textsubscript{105}-containing plasmids generated distinct replication intermediates at the top of the simple Y arc, regardless of the repeat orientation (Figures 4D and S5B). These intermediates are consistent with the formation of double-Y structures (Huberman, 1997), which demonstrates that fork stalling at the (CAGAGG/CCTCTG)\textsubscript{105} repeat was sufficiently persistent on the origin-proximal side to allow the opposite-moving fork to replicate around the plasmid and ultimately stall on the other side of the (CAGAGG/CCTCTG)\textsubscript{105} repeat (Figure 4F). Consistent with this interpretation, the point of divergence from simple Y arcs to the descending arm was altered when the (CAGAGG/CCTCTG)\textsubscript{105} repeat was inserted in an origin-distal location (Figures 4D and S5B).

Quantification of these replication fork barrier effects showed that the (CAGAGG/CCTCTG)\textsubscript{105} repeat was sufficient to impede fork progression by 2- to 3-fold over scrambled control inserts, and this inhibition was enhanced 5-fold by aph treatment (Figure 4G). Importantly, treatment of transfected cells with low-dose aph increased the abundance of double-Y migration products only in (CAGAGG/CCTCTG)\textsubscript{105} repeat-containing vectors, not scrambled controls (Figures 4D, 4E, and S5B–S5D). This selectivity suggests that polymerase slowing may increase (CAGAGG)n structure formation, which further limits polymerase progression. In aggregate, these findings demonstrate that the (CAGAGG)n repeat is sufficient to cause replicative polymerase stalling and impeded fork progression.

Development of BrITL: A Genome-wide Method to Identify Sites of DSB Formation

To validate RPL sites and to identify additional sites that accumulate little RPA, we developed a highly selective method that labels accessible DNA ends in the context of intact chromatin (Figure 5A). This method, abbreviated BrITL, attaches biotin-nucleotide adducts to 3’ ends of broken DNA in permeabilized cells, thus eliminating background-inducing chemical fixation and long incubation steps that generate breaks through depurination and β-elimination. Similar to RPA-ChIP, BrITL retrievals are performed on DNA that is sonicated to relatively large fragment sizes (200–2,000 bp), thus increasing opportunities to map repeat-containing sites of breakage through co-retrieval of adjacent unique sequences (Figure 5A).

To test this method, an I-PpoI endonuclease fusion protein was introduced into passage-immortalized MEFs, which were then treated to activate I-PpoI, and subjected to BrITL. The genomic regions proximal to the I-PpoI cleavage site were quantified by qRT-PCR with input DNA serving as a control, thus quantifying the amount of retrieved DNA relative to the total amount of genomic material present. I-PpoI induction led to a substantial increase in the BrITL detection of genomic DNA nearest the I-PpoI site (Figure 5B). Regions 20 kb away from the I-PpoI endonuclease site were not readily detected by BrITL, even in I-PpoI-induced cells, and the breadth of I-PpoI-enhanced qRT-PCR signal approximated the expected retrieved fragment size of 200 bp to 2,000 bp (Figure 5B). These data indicate that BrITL is capable of retrieving DSB ends both efficiently and specifically.

BrITL Confirms ATRi-Driven Instability of Key RPLs

Genome-wide BrITL-seq was performed on ATRi+aph\textsuperscript{18hrs}-treated cells to both validate RPA accumulation sites and potentially identify additional ATR-dependent sites that were not detectable by RPA-ChIP. Using similar bioinformatic criteria (>4-fold over background; IDR p-value < 10\textsuperscript{-8}; subtraction of DMSO control peaks), a total of 223 BrITL sites were identified as ATR dependent (Table S4). Importantly, RPL sites of highest signal intensity and detection under various conditions were validated as break sites using BrITL (Figures 6A and 6C). Furthermore, sites harboring the most common RPL-associated repeat, (CAGAGG/CCTCTG)n, were prominently detected by BrITL-seq and qRT-PCR, and enrichment of this repeat in BrITL-seq reads from ATR inhibited cells was directly detected by REQer (Figures 6B and 6C).

To determine whether (CAGAGG/CCTCTG)n is sufficient to cause collapse into DSBs, the (CAGAGG/CCTCTG)\textsubscript{105} and scrambled control inserts used for 2D gel analysis were subcloned and transfected into cells to generate stable lines (Figure 6D). These stable lines were then treated with ATRi+aph for 18 hr and assayed for breaks by BrITL-qRT-PCR (Figure 6D). Although breakage near scrambled control inserts was not significantly affected by ATRi+aph\textsuperscript{18hrs} treatment, a significant increase in BrITL detection near the stably inserted (CAGAGG/CCTCTG)n repeat was observed following drug treatment. Interestingly, the (CAGAGG/CCTCTG)\textsubscript{105} repeat was sufficient to cause instability in the absence of ATRi+aph\textsuperscript{18hrs} treatment based on both increased BrITL detection over scrambled controls and a reproducible selection against high-copy (CAGAGG/CCTCTG)n integrants following initial vector transfection (Figures 6D and S6). These data indicate that (CAGAGG/CCTCTG)n repeats can cause DSB formation when placed outside of their normal genomic context, and this breakage is amplified by ATR inhibition.
Although many BrITL sites overlapped with RPL sites, BrITL also identified new ATR-dependent sites that were characterized by the presence of inverted repeats and were not readily detectable by RPA-ChIP (Figures 6E, 6F, and S7). A total of 147 of the 223 ATR-dependent BrITL sites harbored long inverted repeats and quasi-palindromes with predicted melting temperatures of >70°C. Notably, these repeats were mainly comprised of complementary retrotransposable elements, including a variety of SINEs, LINEs, and LTRs (Figures 6E and S7; Table S5). It has long been known that inverted repeats are unstable through their ability to form stable stem-loop structures, and that breakage can occur at such structures in a manner that produces hairpin ends, which can alter processing (Lobachev et al., 2007). Indeed, sensitivity of quasi-palindromic retroelement repeats to checkpoint loss is consistent with engineered reporter system studies in yeast (Zhang et al., 2013). Although we cannot exclude the possibility that BrITL might also detect regressed replication forks, such events would still indicate replication abnormalities at these sites under ATR inhibition. These data suggest that not all ATR-dependent sites accumulate RPA and that inverted repeats and quasi-palindromes comprise an additional category of vulnerable sequences. Collectively, two independent genome-wide methodologies indicated that replication fork collapse caused by ATR inhibition is primarily associated with structure-forming repetitive sequences.

**ATR-Dependent BrITL Sites in Human Cells Are Associated with Structure-Forming Repeats**

Our findings demonstrate that purine-rich structure-forming repeats, such as (CAGAGG/CCTCTG)ₙ, and inverted repeats are key drivers of replication fork collapse following ATR inhibition in murine cells. Although it is possible a structural analog may exist, extensive tandem iterations of (CAGAGG/CCTCTG)ₙ are not common in the human genome (ENCODE). Therefore, we asked what repeat or feature might dominate sensitivity to ATR inhibition in human cells.

Human triple-negative breast cancer cells (MDA-MB-231) were treated with ATRi+aph for 9 hr (ATRi+aph₉hrs) and subjected to genome-wide BrITL-seq. A total of 167 highly significant ATR-dependent sites of expected peak breadth were identified (Table S6). Similar to murine sites, human BrITL sites were not strongly associated with coding genes (Table 1).
Figure 6. BrTL Sites Overlap with RPLs and Inverted Repeats

(A) Top: coverage and ratio tracks of BrTL retrievals and inputs from ATRi+aph18hrs and DMSO-treated cells at RPLs. Bottom: BrTL-qRT-PCR detection of RPL sites relative to peak-centric (CAGAGG)\textsubscript{n} repeats.

(B) Quantification of total repeat units in BrTL retrieval reads by REQer. X axis depicts the total repeat units counted within the total BrTL reads (ATRi+aph18hrs and DMSO control; UT) at indicated distances from the (CAGAGG)\textsubscript{n} and 630-bp scrambled control insertion sites.

(C) Table listing RPLs that overlap with BrTL peaks and the repeats associated with these sites. A total of 23 TAAGGAAGCCAGCTGCGG units (ChrX:135 Mb) and 12 CTRTAAACGCTCGCTGCTYAGGTGCCCGRCRTGAGGACG motifs (Chr4:10.1 Mb, Repeat Masker sequence shown) are present based on the mm10 reference genome.

(D) Left: schematic of (CAGAGG)\textsubscript{105}-containing vector for stable genomic integration and primer sets used in BrTL qRT-PCR analysis. Right: qRT-PCR analyses of genomic BrTL retrievals (ATRi+aph18hrs and DMSO control; UT) at indicated distances from the (CAGAGG)\textsubscript{n} and 630-bp scrambled control insertion sites. Data points represent independent biological replicates; red, scrambled 630 bp insert; blue, (CAGAGG)\textsubscript{105} insert. Hollow dots represent outliers. *p < 0.05; Student’s t test.

(E) Representative coverage tracks of RPA-ChIP and BrTL retrievals and inputs at RPA-positive and RPA-negative BrTL sites following ATRi+aph18hrs. RepeatMasker annotations of repetitive elements as well as a representative inverted retroelement repeat and its M-fold-predicted stem-loop structure are shown below.

(F) Fraction of ATRi+aph18hrs BrTL peaks associated with inverted repeats or AT-rich repeats. For (A), the data are represented as mean ± SEM. See also Figures S6 and S7 and Tables S4 and S5.
Nevertheless, a greater-than-expected fraction of human BrITL sites overlapped with CFSs (25 expected; 35 observed) and with H3K4me3 enrichment regions (1 expected; 15 observed), a marker of gene promoters (Tables 1 and S7). Notably, the 15 H3K4me3-overlapping BrITL sites represented only a minute fraction (<0.1%) of total H3K4me3 enrichment sites, indicating that this histone modification on its own does not strongly influence ATR sensitivity. However, analogous to the association of RPLs with CTCF sites in murine cells, association of BrITL sites with both H3K4me3 and CFSs correlated with the presence of peak-centric repeats, including minisatellites, microsatellites, and inverted repeats (Figures 7A–7C; Table S7). Such repeats were observed in 11 of 15 H3K4me3-overlapping BrITL sites and 24 of 35 sites that overlapped with CFSs (Table S7). Thus, similar to the murine genome, the dependence on ATR for stabilization of the human genome correlated best with the presence of repetitive DNA.

Although DNA repeats were a common feature of ATRi sensitivity, the repeat sequences associated with vulnerability in human cells were different from those identified in the mouse (Figures 7B and 7C). Indeed, nearly half of BrITL-associated minisatellite and microsatellite repeats were distinctly AT-rich, as defined by a greater-than-expected frequency of AT/TA content with respect to average genomic frequency (Figure 7B; p < 0.05; chi-square test). Such content was not a prevailing characteristic of murine BrITL sites (Figure 6F). Non-AT-rich microsatellite and minisatellite repeats were also observed within human BrITL peaks; however, these associated repeats were less common compared to AT-rich repeats (Figures 7B and 7C). Furthermore, in contrast to the prevalence of tandem (CAGAGG/CCTCTG)n repeats in murine RPL and BrITL sites, AT-rich satellites in human BrITL sites were more pleomorphic, sometimes occurring tandemly in only one BrITL site. These observations begged the questions whether specific motifs were present and enriched within these larger satellites and whether peak-centric repeats observed had structure-forming potential. The motif-finding program, Homer2, and the microsatellite prediction tool, MISA (Beier et al., 2017), were used to identify repeated sequences within BrITL peaks in an unsupervised fashion. Identified motifs were then quantified within ATR-dependent BrITL peaks and compared to similar quantifications within 10 independent sets of randomly selected “peak” regions of similar footprint size (Figure 7C). This unbiased approach once again identified AT-rich motifs as the most common sequence feature enriched in human BrITL peaks (Figure 7C). For example, TATATATTATATATAATATATATATATATAACGT were enriched 104-fold and 410-fold in BrITL peaks over randomly selected pseudo-peaks (Figure 7C). Remarkably, many of these BrITL-enriched motifs were found to be components of larger minisatellites. For example, TATATATGTA/TATATATACACGT, within the larger tandem minisatellite TATATATGTA/TATATATACACGT, is enriched 46-fold in ATR-dependent BrITL peaks (461 sites expected; 15 observed) in BrITL peak Chr11:102,89 Mb (Figures 7B and 7C). Many times, AT-rich repeats and other BrITL peak sequences were quasi- or perfect palindromes (e.g., TATATATGTA/TATATATACACGT, TATAATATA/TATAATATACACGT, in BrITL peak Chr11:102,89 Mb (Figures 7B and 7C). Moreover, the ATR-dependent BrITL peaks were also capable of forming stable secondary structures. In contrast, motifs that were not enriched in ATR-dependent BrITL peaks over randomly selected regions exhibited little hairpin-forming potential (Figure 7C). Indeed, some motifs, such as tandem polyA and polyT repeats, which do not form hairpins without a complementary sequence in cis, were significantly depleted from ATR-dependent BrITL peaks (Figure 7C), indicating that such stretches on their own do not cause dependence on ATR for stability.

Table 1. Genomic Features of BrITL Sites in Human Breast Cancer Cells

<table>
<thead>
<tr>
<th>Genic Feature</th>
<th>Intron</th>
<th>Exon</th>
<th>Inter-genic</th>
<th>H3 K4me3</th>
<th>H3 K27ac</th>
<th>H3 K36me3</th>
<th>TTRs</th>
<th>CTCF</th>
<th>CFS</th>
</tr>
</thead>
<tbody>
<tr>
<td>% of BrITL sites overlapping (sites observed)</td>
<td>37.7%</td>
<td>26.3%</td>
<td>3.9%</td>
<td>62.3%</td>
<td>11.3%</td>
<td>9.0% (15)</td>
<td>3.0% (5)</td>
<td>0% (0)</td>
<td>15.6% (26)</td>
</tr>
<tr>
<td>% coverage of human genome (sites expected)</td>
<td>53.3%a</td>
<td>46.7%a</td>
<td>0.34% (1)b</td>
<td>0.78% (1)b</td>
<td>1.4% (2)</td>
<td>11.6% (19)b</td>
<td>0.36% (1)b</td>
<td>14.9% (25)c</td>
<td></td>
</tr>
<tr>
<td>p value</td>
<td>&lt;10^-4</td>
<td>0.11</td>
<td>1.0</td>
<td>0.07</td>
<td>0.08</td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Percent and number of BrITL sites overlapping with noted features are compared to the expected overlap based on fraction of the genome comprised by these features. Statistical significance (p value) was calculated by permutation test. Also see Table S7.

*Yue et al., 2014

aENCODE database (MCF7)

bFungtammasan et al., 2012

cBacolla et al., 2016; Kato et al., 2012

Mfold (Zuker, 2003) was then used to examine the structure-forming potential of repeat-containing sequences identified within BrITL peaks. Overall, identified repeat sequences were predicted to form large and stable hairpin structures with melting temperatures greater than 55°C (Figures 7B–7D; data not shown). Notably, the MISA- and Homer2-identified sequence motifs that were enriched in ATR-dependent BrITL peaks were also capable of forming stable secondary structures. In contrast, motifs that were not enriched in ATR-dependent BrITL peaks over randomly selected regions exhibited little hairpin-forming potential (Figure 7C). Indeed, some motifs, such as tandem polyA and polyT repeats, which do not form hairpins without a complementary sequence in cis, were significantly depleted from ATR-dependent BrITL peaks (Figure 7C), indicating that such stretches on their own do not cause dependence on ATR for stability.

Notably, the large hairpin-forming structures generated by AT-rich BrITL peaks show numerous similarities to the inverted and palindromic AT-rich repeats (PATRRs) that are associated with translocation and deletion hotspots in cancer and developmental disorders, such as DiGeorge and Emanuel syndromes (Bacolla et al., 2016; Kato et al., 2012). Indeed, a sequence motif observed frequently in PATRRs, TATAATATA (Delhais, 2015), is enriched 46-fold in ATR-dependent BrITL peaks (461
**A**

Repetitive sequences in human BrTL peaks based on signal intensity (top 15):

<table>
<thead>
<tr>
<th>Chr</th>
<th>Start</th>
<th>End</th>
<th>Examples of repeated sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chr6</td>
<td>55,475</td>
<td>55,515</td>
<td>TATATA x 5</td>
</tr>
<tr>
<td></td>
<td>55,515</td>
<td>56,124</td>
<td>TATATA x 5</td>
</tr>
<tr>
<td></td>
<td>121,088</td>
<td>121,128</td>
<td>TATATA x 5</td>
</tr>
<tr>
<td>Chr12</td>
<td>55,515</td>
<td>55,559</td>
<td>TATATA x 5</td>
</tr>
<tr>
<td></td>
<td>55,559</td>
<td>55,604</td>
<td>TATATA x 5</td>
</tr>
<tr>
<td></td>
<td>121,128</td>
<td>121,178</td>
<td>TATATA x 5</td>
</tr>
<tr>
<td>Chr14</td>
<td>55,559</td>
<td>55,604</td>
<td>TATATA x 5</td>
</tr>
<tr>
<td></td>
<td>55,604</td>
<td>55,649</td>
<td>TATATA x 5</td>
</tr>
<tr>
<td>Chr12</td>
<td>55,649</td>
<td>55,694</td>
<td>TATATA x 5</td>
</tr>
<tr>
<td></td>
<td>55,694</td>
<td>55,739</td>
<td>TATATA x 5</td>
</tr>
<tr>
<td>Chr12</td>
<td>55,739</td>
<td>55,784</td>
<td>TATATA x 5</td>
</tr>
<tr>
<td></td>
<td>55,784</td>
<td>55,829</td>
<td>TATATA x 5</td>
</tr>
<tr>
<td>Chr12</td>
<td>55,829</td>
<td>55,874</td>
<td>TATATA x 5</td>
</tr>
<tr>
<td></td>
<td>55,874</td>
<td>55,919</td>
<td>TATATA x 5</td>
</tr>
<tr>
<td>Chr12</td>
<td>55,919</td>
<td>55,964</td>
<td>TATATA x 5</td>
</tr>
<tr>
<td></td>
<td>55,964</td>
<td>55,999</td>
<td>TATATA x 5</td>
</tr>
</tbody>
</table>

**B**

Human BrTL peak overlap:

<table>
<thead>
<tr>
<th>Genomic Feature</th>
<th>Sequence Characteristics</th>
<th>Chr</th>
<th>Start</th>
<th>End</th>
<th>Examples of repeated sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Chr1</td>
<td>102895030</td>
<td>102895635</td>
<td>TATATA x 5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chr1</td>
<td>102895030</td>
<td>102895635</td>
<td>TATATA x 5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chr3</td>
<td>28215773</td>
<td>28216497</td>
<td>TATATA x 5</td>
</tr>
</tbody>
</table>

**C**

Randomized Peaks

**D**

Chromosomal distribution of BrTL peak overlap

(legend on next page)
occurrences) compared to randomly chosen pseudo-peaks ($p < 10^{-25}$). In addition, akin to inverted retrotransposable repeats in mouse BrITL sites (Figures 6E and 6F; Table S5), 12 human BrITL sites were comprised of inverted Alu elements or similar sequences (Figure 7B; Table S7), further indicating that stem-loop structures play a key role in ATRi-driven fork collapse as detected by BrITL. Overall, these data indicate that human ATR-dependent BrITL sites, similar to murine sites, are strongly associated with repetitive sequences that form stable secondary structures.

**DISCUSSION**

Using two distinct detection methods in mouse and human cells, we have shown that ATR inhibition causes localized replication fork collapse preferentially at repetitive DNA. Whereas a variety of endogenous stresses can activate ATR, structure-forming repeats were the genomic feature most strongly associated with site-specific breakage under ATR inhibition. Collectively, the types of repeats associated with collapse and the optimal means of peak detection (RPA versus BrITL) have created a diverse prioritized catalog of ATR-dependent sites in the mammalian genome, which can putatively serve as specific genomic readouts of ATR dysfunction and therapeutic biomarkers of response to ATR inhibition. In summary, these findings have provided a more precise understanding of the role of ATR in genome stabilization as well as new tools to study it.

**Dependence of Structure-Forming Repeats on ATR for Stability**

The repeat categories most dependent on ATR for stability were purine-rich structure-forming repeats and those that generate stem-loop structures. Purine-rich structure-forming repeats exhibited a similar sequence pattern: C\textsubscript{N2-6}G, with the intervening regions made up of (AG)\textsubscript{n} or ACA segments (e.g., CAGG, CAGAGG, CACAGAGG, and CACAG). Repeats of these sequences generate strand-selective intramolecular structures that impede DNA replication (Figures 3 and 4). Notably, increased polymerase-helicase uncoupling and ssDNA generation through partial polymerase inhibition (aph) may facilitate the formation of such purine-rich structures, which further inhibits polymerase progression and increases fork collapse under ATR inhibition (Figures 1, 2, and 4). Interestingly, secondary structure formation and polymerase blockage on the purine-rich strand would not necessarily impede synthesis on the complementary pyrimidine-rich strand, producing daughter strand gaps ahead of the purine-rich repeat that would lack available complementary sequence for reannealing. Accordingly, significant repeat-strand bias was observed in 16 of 18 RPLs where asymmetric RPA-ChIP signal intensity was observed around the central (CAGAGG/CCTCTG)\textsubscript{N} repeat (Figure S4F; $p < 0.0009$). It seems likely that strand-selective structure formation may be the underlying cause of robust RPA accumulation upon collapse at this type of repeat, but other models are certainly plausible.

The structures formed by purine-rich repeats are non-B-form according to CD and TDS studies (Figures 3 and S4) and are unlikely to involve G quartets given G-quadruplex-promoting ions (e.g., Li) and small molecules did not foster structure formation (data not shown). Principal-component analyses of the CD signatures of over 60 known DNA structures were not consistent with the (CAGAGG)\textsubscript{N} signature (J. Brad Chaires, personal communication). However, a tetrahelical structure formed by a somewhat related purine-rich repeat in humans, (AGCGAGGG)\textsubscript{N}, (Kocman and Plavec, 2014, 2017) exhibits biophysical similarities to murine (CAGAGG)\textsubscript{N}, suggesting that the structures formed by these sequences may be conserved between species.

In contrast to the unusual strand-specific structures formed by purine-rich repeats, the other main category of ATR-dependent repeats, namely inverted repeats and AT-rich palindromes and quasi-palindromes, are predicted to form structures on both complementary strands (Figures 6, 7, and S7). These hairpin structures are well known to be sufficient to cause genomic instability (Kato et al., 2012; Lobachev et al., 2007). Notably, similar sequences not expected to form hairpins, such as poly(A)\textsubscript{N}; poly(T)\textsubscript{N}; and head-to-tail LINE, SINE, LTR, and Alu elements, were not enriched in BrITL retrievals after ATR inhibition (Figures 7C and S3A; data not shown). This correlation indicates that structural formation is a key aspect of dependence on ATR for stability. Additional research is required to determine the mechanism by which BrITL-associated repeats cause fork collapse and whether structure formation is sufficient for ATR dependence or requires other associated characteristics.

Importantly, hairpin-forming sites in mouse and human cells did not accumulate substantial amounts of RPA but were easily detected by BrITL (Figures 6, 7, and S7). These findings suggest that DNA resection is impaired at hairpin-forming sites, putatively by the close-ended structures themselves or that structure formation in RPLs somehow fosters unusually high levels of RPA accumulation compared to other breakage sites. In either case, the distinct structural characteristics of these two main categories of ATR-dependent repeats, strand-selective structures and hairpins, correlate with preferential detection by RPA and BrITL, respectively. This correlation indicates that both the
structural causes and biochemical consequences of fork collapse at these two categories of repeats may be distinct.

**RPLs, BrITL, and CFSs**

CFSs are associated with incomplete synthesis in M phase, which can lead to breaks either through endonuclease cleavage or physical strain at ultrafine bridges (Glover et al., 2017; Letessier et al., 2011; Wyatt et al., 2017; Ying et al., 2013). Accordingly, low origin density and difficult-to-replicate sequences have been proposed to cause CFS breakage (Glover et al., 2017; Letessier et al., 2011). In regards to the latter possibility, ATR-dependent BrITL sites were marginally enriched in human CFSs (25 expected and 35 observed; Table S7), with eight sites located in most expressive breast epithelial CFSs (FRA1D, FRA2I, and FRA5E; Hosseini et al., 2013). However, it is prudent to note that this enrichment is only 10 sites more than expected by chance (Tables 1 and S7). In addition, RPL and BrITL sites were not enriched in highly expressed murine CFSs (14 expected and 9 observed; Table S7).

Although it is possible that spreading of RPA and BrITL signals across these large regions was too weak to call, another possibility is that ATRi-driven fork collapse is not substantially more common in CFSs than in other parts of the genome. Notably, non-peak-associated sensitivity to ATR (ATRi versus DMSO; 20–30 kb from peak regions; Figures 6A and 7A; data not shown) was not remarkably different from the effect of ATR suppression on CFS breakage (Casper et al., 2002). This viewpoint is in line with replication rates in CFSs and the genome at large being similar and that CFS breakage correlates well with large inter-origin distances (Letessier et al., 2011). Regardless of these potential mechanisms of CFS breakage, our unbiased identification of sites that are most sensitive to ATR inhibition pinpoints new, specific vulnerabilities both within CFSs and throughout the genome.

**Potential Impact of RPL and BrITL Sites on Aging and Cancer Treatment**

A variety of recent studies have implicated “replication stress” as a cause of age-related pathologies (Burhans and Weinberger, 2007; Flach et al., 2014; Ruzankina et al., 2007). Although the molecular causes of such stress have remained relatively obscure (Burhans and Weinberger, 2007), both triplet repeat expansions and ribosomal repeat instability have been associated with age-related pathologies (Flach et al., 2014; Orr and Zoghbi, 2007). It is interesting to speculate that the difficult-to-repeat sequences identified herein might contribute to such replicative stress, where replication fork collapse and the ensuing DNA damage responses would degrade regenerative capacity. Identification of these ATR-dependent sites and causative repeat sequences provides the means to study their effects on tissue homeostasis.

The identification of ATRi-induced breakage sites also has significant implications for cancer treatment. ATR inhibitors have entered clinical trials for the treatment of a variety of cancers. Although it is evident that replication fork collapse and cell cycle checkpoint abrogation are key components of the mechanism of ATRi action, the difficult-to-repeat sequences that rely on ATR for stability are similarly part of that mechanism. Given that the lengths of microsatellite and minisatellite repeats are frequently polymorphic, it is intriguing to speculate that cancer-cell-associated expansions of such repetitive sequences may help predict benefit from ATRi-based therapies. Accordingly, clinical response to ATRi would not be solely based on lethal interactions with defective gene products, but also with the state of RPL- and BrITL-associated repeats.

**STAR METHODS**

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **CONTACT FOR REAGENT AND RESOURCE SHARING**
- **PRIMERS USED IN BRITL QRT-PCR, RELATED TO FIGURES 6 AND 7**
- **SCRAMBLLED AND (CAGAGG)105 INSERT SEQUENCES, RELATED TO FIGURES 4 AND 6, AND TO FIGURES S5 AND S6**
  - Scrambled insert sequence
  - (CAGAGG)105 insert sequence
- **EXPERIMENTAL MODEL AND SUBJECT DETAILS**
  - Cell lines –
  - Treatment –
  - Culture –
- **METHOD DETAILS**
  - RPA ChIP-Seq –
  - BrITL –
  - qRT-PCR analysis –
  - NGS library preparation –
  - Transfection of (CAGAGG)105 or scrambled sequence vectors –
  - Southern Blot –
  - Bioinformatics –
  - Fork-pausing Experiments –
  - Biophysical Characterization of DNA Secondary Structure –
- **QUANTIFICATION AND STATISTICAL ANALYSIS**
  - RPA ChIP-Seq –
  - BrITL –
  - Biophysical Characterization of DNA Secondary Structure –
  - Fork-pausing Experiments –
- **DATA AND SOFTWARE AVAILABILITY**
  - Mendeley Deposited Data
  - Accession Numbers
  - ENCODE Mouse Datasets –
  - ENCODE Human Datasets –

**SUPPLEMENTAL INFORMATION**

Supplemental Information includes seven figures and seven tables and can be found with this article online at https://doi.org/10.1016/j.molcel.2018.08.047.

**ACKNOWLEDGMENTS**

This project was funded by grants from the NIH (E.J.B.: 2R01AG027376 and 1R01CA189743; N.S.: 5R25CA101871; Y.-C.T.: T32ES019851; L.A.Y.: 1R15CA208678). Additional funding was provided through the Pennsylvania...
Department of Health (E.J.B.), Center of Excellence in Environmental Toxicology (P30ES013508, E.J.B.), Bassler Center for BRCA Research (E.J.B.), Abramson Family Cancer Research Institute (E.J.B.), Camille and Henry Dreyfus Teacher-Scholar Award (L.A.Y.), and Jake Gitten Cancer Research Foundation (K.A.E.). We are grateful to Dr. Jean-Louis Mergny (IECB, Bordeaux, France), in whose laboratory some of these studies were conducted (L.A.Y.). We would also like to thank Dr. J. Brad Chaires for CD/TDS principal-component comparisons to existing signatures and Dr. Beverly Emanuel (CHOP) for helpful insights. The contents of this manuscript are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

AUTHOR CONTRIBUTIONS

The goals of the project and overall study design were conceived and designed by N.S., Y.-C.T., and E.J.B. N.S., Y.-C.T., and R.C. analyzed sequencing data. D.J., B.P., J.C., and S.H. signed by N.S., Y.-C.T., and E.J.B. N.S., Y.-C.T., S.H., D.J., B.P., J.C., T.A., and R.C. generated the data. D.M., M.D., Y.L., T.K., S.S., P.B., K.A., X.W., K.D.S., J.S., and R.C. analyzed sequencing data. D.J., B.P., J.C., and L.A.Y. wrote the section regarding experiments in Figures 3 and S4. S.H. and K.A.E. wrote the section pertaining to Figures 4 and S6. N.S. and E.J.B. wrote and assembled the manuscript with input from all authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 27, 2017
Revised: July 11, 2018
Accepted: August 30, 2018
Published: October 4, 2018

REFERENCES

guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 22, 1813–1831.
## STAR METHODS

### KEY RESOURCES TABLE

<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antibodies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mouse monoclonal anti-RPA32 (RPA ChIP on MEF 4-3 cells)</td>
<td>Millipore</td>
<td>NA19L-100UG; RRID:AB_565123</td>
</tr>
<tr>
<td>Mouse IgG2a Isotype Control from murine myeloma antibody (Pre-clear for RPA ChIP on MEF 4-3 cells)</td>
<td>Sigma Aldrich</td>
<td>M5409; RRID:AB_1163691</td>
</tr>
<tr>
<td>Bacterial and Virus Strains</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R408 Helper Phage</td>
<td>Promega</td>
<td>Cat# P2291</td>
</tr>
<tr>
<td>SURE 2 Supercompetent Cells</td>
<td>Agilent Technologies</td>
<td>Cat# 200152</td>
</tr>
<tr>
<td>Chemicals, Peptides, and Recombinant Proteins</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4-OHT</td>
<td>EMD Chemicals/Calbiochem</td>
<td>579002-5MG</td>
</tr>
<tr>
<td>Shield-1</td>
<td>Laboratory of Tom Wandless, Stanford</td>
<td>N/A</td>
</tr>
<tr>
<td>ATR45 (ATR inhibitor for MEFs)</td>
<td>Ohio State University (OSU); Charrier et al., 2011</td>
<td>ATR45</td>
</tr>
<tr>
<td>VE-822 (ATR inhibitor for MDA-MB-231 cells)</td>
<td>Selleck Chemicals</td>
<td>Cat# S7102</td>
</tr>
<tr>
<td>Aphidicolin (for RPA ChIP and BrTL experiments)</td>
<td>Calbiochem</td>
<td>CAS 38966-21-1</td>
</tr>
<tr>
<td>Biotin-16-ddUTP</td>
<td>Enzo Life Sciences</td>
<td>ENZ-42813</td>
</tr>
<tr>
<td>TdT</td>
<td>Roche</td>
<td>Cat# 03333566001</td>
</tr>
<tr>
<td>Polyethylene Glycol (Avg Mol Wt of 8000)</td>
<td>Sigma</td>
<td>Cat# P5413; CAS 25322-68-3</td>
</tr>
<tr>
<td>Aphidicolin (for Fork Pausing experiments)</td>
<td>Sigma</td>
<td>Cat# A0781; CAS 38966-21-1</td>
</tr>
<tr>
<td>Equilibrated Phenol, pH 8.0, Ultrapure</td>
<td>Affymetrix/Thermo-Fisher</td>
<td>Cat# AAJ75829AN; CAS 108-95-2</td>
</tr>
<tr>
<td>γ&lt;sup&gt;32&lt;/sup&gt;P ATP (6000Ci/mmol)</td>
<td>Perkin-Elmer</td>
<td>Cat# BLU0022001MC</td>
</tr>
<tr>
<td>T4 Polynucleotide Kinase</td>
<td>Thermo-Fisher</td>
<td>Cat# 18004010</td>
</tr>
<tr>
<td>Recombinant human PCNA</td>
<td>Laboratory of Marietta Lee</td>
<td>Biochemistry 2001 40: 4512-4520</td>
</tr>
<tr>
<td>Recombinant yeast RFC</td>
<td>Laboratory of Linda Bloom</td>
<td>Thompson et al., 2012</td>
</tr>
<tr>
<td>Sequenase 2.0</td>
<td>Affymetrix/Thermo-Fisher</td>
<td>Cat# 70775Y</td>
</tr>
<tr>
<td>Exo- Klenow Polymerase</td>
<td>Affymetrix/Thermo-Fisher</td>
<td>Cat# 70057Z</td>
</tr>
<tr>
<td>Critical Commercial Assays</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NEBNext Library Kit</td>
<td>NEB</td>
<td>E6000S</td>
</tr>
<tr>
<td>Dynabeads KiloBaseBinder kit</td>
<td>Life Technologies</td>
<td>601-01</td>
</tr>
<tr>
<td>Ampure XP Beads purification kit</td>
<td>Beckman Coulter</td>
<td>A63880</td>
</tr>
<tr>
<td>Lipofectamine 3000 Transfection Reagent</td>
<td>Thermo Fisher Scientific</td>
<td>L3000001</td>
</tr>
<tr>
<td>Biotin Chromogenic Detection kit</td>
<td>Thermo Fisher Scientific</td>
<td>K0661</td>
</tr>
<tr>
<td>High Prime DNA labeling kit version 10</td>
<td>Roche Diagnostics</td>
<td>REF: 11585592001</td>
</tr>
<tr>
<td>Biotin DecaLabel DNA Labeling Kit</td>
<td>Thermo Fisher Scientific</td>
<td>K0651</td>
</tr>
<tr>
<td>Dynabeads Protein A for Immunoprecipitation</td>
<td>Invitrogen</td>
<td>100-02D</td>
</tr>
<tr>
<td>Deposited Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPA ChIP-Seq raw and processed files</td>
<td>This paper</td>
<td>GEO: GSE100790</td>
</tr>
<tr>
<td>BrTL raw and processed files</td>
<td>This paper</td>
<td>GEO: GSE115623; GSE115624</td>
</tr>
<tr>
<td>Mendeley deposited data</td>
<td>This paper</td>
<td><a href="https://doi.org/10.17632/9gxwy4vxt8.1">https://doi.org/10.17632/9gxwy4vxt8.1</a></td>
</tr>
<tr>
<td>Experimental Models: Cell Lines</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEF 4-3 cells (Atr&lt;sup&gt;Flox&lt;/sup&gt; Cre-ERT2&lt;sup&gt;+&lt;/sup&gt;)</td>
<td>Smith et al., 2009</td>
<td>N/A</td>
</tr>
<tr>
<td>TIM KD MEF 4-3 cells</td>
<td>Smith et al., 2009</td>
<td>N/A</td>
</tr>
</tbody>
</table>

(Continued on next page)
**CONTACT FOR REAGENT AND RESOURCE SHARING**

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Eric Brown (brownej@upenn.edu).

<table>
<thead>
<tr>
<th>REAGENT or RESOURCE</th>
<th>SOURCE</th>
<th>IDENTIFIER</th>
</tr>
</thead>
<tbody>
<tr>
<td>I-PpoI MEF 4-3 cells</td>
<td>This paper</td>
<td>N/A</td>
</tr>
<tr>
<td>MDA-MB-231 cells</td>
<td>Andy J. Minn, Univ. of Pennsylvania</td>
<td>N/A</td>
</tr>
<tr>
<td>U-2OS osteosarcoma cells</td>
<td>ATCC</td>
<td>HTB-96</td>
</tr>
<tr>
<td>Fetal Bovine Serum</td>
<td>GE Life Sciences/Hyclone</td>
<td>Cat. No. SH 30071.03</td>
</tr>
<tr>
<td>Oligonucleotides</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>dTn gel electrophoresis ladder: 5’-dT15, 5’-dT34, 5’-dT30, 5’-dT60, 5’-dT90</td>
<td>Integrated DNA Technology</td>
<td>N/A</td>
</tr>
<tr>
<td>G94-43-mer, PAGE-Purified 5’-GGA TAA CAA TTT CAC ACA GGA AAC AGC TAT GAC CAT GAT TAC G-3’</td>
<td>Integrated DNA Technologies</td>
<td>N/A</td>
</tr>
<tr>
<td>Scrambled and (CAGAGG)₁₀₅ insert sequences, see below</td>
<td>This paper</td>
<td>N/A</td>
</tr>
<tr>
<td>HFUGW probe primer set (against UBC promoter): Forward: 5’-CATAGAGCTCGCCCTTAGAACC-3’ Reverse: 5’-GCCAGAATTGCGGGACATTT-3’</td>
<td>Integrated DNA Technologies</td>
<td>N/A</td>
</tr>
<tr>
<td>Recombinant DNA</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>pBabe-ddIPpoI</td>
<td>Addgene, Goldstein et al., 2013 #49052</td>
<td></td>
</tr>
<tr>
<td>pMIG Bcl-XL</td>
<td>Addgene</td>
<td>#8790</td>
</tr>
<tr>
<td>pGEM-3Zf(-) Vector</td>
<td>Promega</td>
<td>Cat# P2261</td>
</tr>
<tr>
<td>pML113, pML114, pML115</td>
<td>Massimo Lopes, Univ. of Zurich</td>
<td>Follonier et al., 2013</td>
</tr>
<tr>
<td>Software and Algorithms</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>Trimomomatic</td>
<td>Bolger et al., 2014</td>
<td>N/A</td>
</tr>
<tr>
<td>Fastqc</td>
<td>Leggett et al., 2013</td>
<td>N/A</td>
</tr>
<tr>
<td>STAR-2.5.2a Aligner</td>
<td>Dobin et al., 2013</td>
<td>N/A</td>
</tr>
<tr>
<td>MACS2 peak-calling</td>
<td>Zhang et al., 2008</td>
<td>N/A</td>
</tr>
<tr>
<td>IDR analysis</td>
<td>Landt et al., 2012</td>
<td>N/A</td>
</tr>
<tr>
<td>Homer2 motif-finding</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>The M-fold Web Server (DNA folding form)</td>
<td>Zuker, 2003</td>
<td>N/A</td>
</tr>
<tr>
<td>MISA</td>
<td>Beier et al., 2017</td>
<td>N/A</td>
</tr>
<tr>
<td>ImageQuant version 5.2</td>
<td>GE Healthcare</td>
<td>N/A</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>Illustra Microspin G-50 column</td>
<td>GE Healthcare</td>
<td>Cat# 27-5330-1</td>
</tr>
<tr>
<td>Xtreme Gene XP version 8</td>
<td>Roche Diagnostics</td>
<td>REF: 06366236001</td>
</tr>
<tr>
<td>RPL Peak Genomic Location (mm9)</td>
<td>Distance relative to (CAGG)n repeat region</td>
<td>Forward Primer (5’ - &gt; 3’)</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>---------------------------------------------</td>
<td>----------------------------</td>
</tr>
<tr>
<td>Chr7: 35943530 – 35946581</td>
<td>20 kb 5’</td>
<td>GCAAGCCATGGAGACTCCTATA</td>
</tr>
<tr>
<td></td>
<td>1,260 bp 5’</td>
<td>CCGTGTGGGTCTGCTGTCATTATTA</td>
</tr>
<tr>
<td></td>
<td>220 bp 5’</td>
<td>ACACCCAGTGACCAAAGATATAG</td>
</tr>
<tr>
<td></td>
<td>230 bp 3’</td>
<td>GACAACTGGCTCAGGAGATATAT</td>
</tr>
<tr>
<td></td>
<td>530 bp 3’</td>
<td>GAATGCTGCCTGGTTGATATG</td>
</tr>
<tr>
<td></td>
<td>20 kb 3’</td>
<td>GCCGAGGGAGAGTTATG</td>
</tr>
<tr>
<td>Chr6: 87722438 – 87728725</td>
<td>20 kb 5’</td>
<td>CCGACTGAGCTCAGCTCTTATT</td>
</tr>
<tr>
<td></td>
<td>1,450 bp 5’</td>
<td>GTGATGAGCCCAAACACCTAAGT</td>
</tr>
<tr>
<td></td>
<td>285 bp 5’</td>
<td>CCCAGAGCTGTCACATGAGAGAAT</td>
</tr>
<tr>
<td></td>
<td>970 bp 3’</td>
<td>GCCGAGGACCTCCTACATTAA</td>
</tr>
<tr>
<td></td>
<td>20 kb 3’</td>
<td>CACAGACACCACAGCATAGAG</td>
</tr>
<tr>
<td>Chr11: 5641800-5644000</td>
<td>20 kb 5’</td>
<td>TCCTCCATTTCCTCCTATTCC</td>
</tr>
<tr>
<td></td>
<td>430 bp 5’</td>
<td>TGGGCTCTCAGCTACTTACTA</td>
</tr>
<tr>
<td></td>
<td>115 bp 5’</td>
<td>GCCTGATTCACAGCAGAGTAAAT</td>
</tr>
<tr>
<td></td>
<td>575 bp 3’</td>
<td>TGGGAGCAGGAGATAGAAAT</td>
</tr>
<tr>
<td></td>
<td>20 kb 3’</td>
<td>GTGATCAGAGACTATGAGGAAATG</td>
</tr>
<tr>
<td></td>
<td>20 kb 3’</td>
<td>CCGTTCAGGAATAGTCTGCAAA</td>
</tr>
<tr>
<td></td>
<td>(Continued on next page)</td>
<td></td>
</tr>
</tbody>
</table>

RPL Peak Genomic Location (mm9) | Distance relative to (CAGG)n repeat region | Forward Primer (5’ - > 3’) | Reverse Primer (5’ - > 3’) |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chr9: 121838100-121841700</td>
<td>20 kb 5’</td>
<td>GTGACTTATCTCCACCTGACTC</td>
<td>CTACTGAGGAGAAGCCTGAGATG</td>
</tr>
<tr>
<td></td>
<td>460 bp 5’</td>
<td>GCAAGTTCTGGGCTGTGATCTATT</td>
<td>GTGAGGTTGTTGCTCAGAGGTCTAT</td>
</tr>
<tr>
<td></td>
<td>165 bp 5’</td>
<td>CCTGCTCTCGACACAATGTACTC</td>
<td>CCGATACACCCGCTCCTTT</td>
</tr>
<tr>
<td></td>
<td>230 bp 3’</td>
<td>GCCTAGTGATACATCTCCTCTCTA</td>
<td>GTAAGGCTACACTGAGATGAT</td>
</tr>
<tr>
<td></td>
<td>360 bp 3’</td>
<td>CAAAGTGATGGCTGGAGGATG</td>
<td>GCCTAGGATTTGCTGATG</td>
</tr>
<tr>
<td></td>
<td>20 kb 3’</td>
<td>CCGTTCAGGAATAGTCTGCAAA</td>
<td>AGAGATGAGCAGGAAATC</td>
</tr>
</tbody>
</table>

RPL Peak Genomic Location (mm9) | Distance relative to (CACAG)n repeat region | Forward Primer (5’ - > 3’) | Reverse Primer (5’ - > 3’) |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chr17: 13737678-13751955</td>
<td>20 kb 5’</td>
<td>CTTTCTGTCGGTCCTCTCTATC</td>
<td>CACAGCTGAGAATCCTGAGATG</td>
</tr>
<tr>
<td></td>
<td>1,085 bp 5’</td>
<td>TTTCATGACCGAGAGATGAGAGTT</td>
<td>GCCTCTCCCTGTGCTGATGAT</td>
</tr>
<tr>
<td></td>
<td>600 bp 5’</td>
<td>GAGATACAGACCAGCAGAGGAA</td>
<td>CTACCTTGTACAGGAGGATGATA</td>
</tr>
<tr>
<td></td>
<td>0 bp 3’</td>
<td>GCTGTCTGCTGCTGCTGTAATG</td>
<td>GCCTCTTGTACGGCTCCTTCCT</td>
</tr>
<tr>
<td></td>
<td>500 bp 3’</td>
<td>GCTCTACTGGTGTCCTGAGTAAT</td>
<td>GCCTCTTGCTGCAGGGGATG</td>
</tr>
<tr>
<td></td>
<td>1,400 bp 3’</td>
<td>CCGCTAATTCCCTCTGAGATC</td>
<td>TCTCTGCTGACTGCTCCTATT</td>
</tr>
<tr>
<td>Chr6: 29694103-29706050</td>
<td>30 kb 5’</td>
<td>GCTTCTGTCAGAGACAGAGATTG</td>
<td>GAAAGTCCAGAAGGAGCACTAG</td>
</tr>
<tr>
<td></td>
<td>20 kb 5’</td>
<td>CCGACAGATGCTGCTCAGAGAAG</td>
<td>GCCTTTATGCTGAGATGAT</td>
</tr>
<tr>
<td></td>
<td>1,200 bp 5’</td>
<td>GCCACAGCTACAGCAGAGAAGA</td>
<td>TGGAGATACAGCAGATCTCTTT</td>
</tr>
<tr>
<td></td>
<td>170 bp 5’</td>
<td>ATCACACCCACTCCACAAGG</td>
<td>GAATTCCACCATGAGCTCATA</td>
</tr>
<tr>
<td></td>
<td>190 bp 3’</td>
<td>AGTGAGTACAGCATGAGTT</td>
<td>CCTATGAGTACCTGCTACTTCA</td>
</tr>
<tr>
<td></td>
<td>400 bp 3’</td>
<td>CCTCTAGAGGCTCTCTTTAATTGT</td>
<td>GAACCTCAGAGATACCATCCTCT</td>
</tr>
<tr>
<td></td>
<td>1,015 bp 3’</td>
<td>CCGTCTGCTACCTCAATCATAAT</td>
<td>TGAAGATGAGGCCCTTCTC</td>
</tr>
<tr>
<td></td>
<td>20 kb 3’</td>
<td>AGTCTTGCACCCACCCTGGA</td>
<td>GGCGTACCTCAGTTTATT</td>
</tr>
<tr>
<td></td>
<td>30 kb 3’</td>
<td>CCGTACTGTCGGTGGATTTG</td>
<td>CACCTTCTACACGAGCACTA</td>
</tr>
<tr>
<td>RPL Peak Genomic Location (mm9)</td>
<td>Distance relative to (CAGAGG)_n repeat region</td>
<td>Forward Primer (5’ - &gt; 3’)</td>
<td>Reverse Primer (5’ - &gt; 3’)</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>-----------------------------------------------</td>
<td>-----------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>I-PpoI site in rDNA sequence</td>
<td>Distance relative to I-PpoI site</td>
<td>Forward Primer (5’ - &gt; 3’)</td>
<td>Reverse Primer (5’ - &gt; 3’)</td>
</tr>
<tr>
<td>11.5 kb 3’</td>
<td>TTAATTAA</td>
<td>CACGCTGGTCTCTTTCCCTATT</td>
<td>GACAGACCGAGGGAGCTAAA</td>
</tr>
<tr>
<td>7 kb 3’</td>
<td>GCTAGAACGCTTACCATC</td>
<td>GCCTGGAAGAGTCTCTGTATTG</td>
<td></td>
</tr>
<tr>
<td>780 bp 3’</td>
<td>AGACCTCTGGCAATTTG</td>
<td>GACCATTTACCCTACGGAAGTTA</td>
<td></td>
</tr>
<tr>
<td>70 bp 3’</td>
<td>CTAGCAGCGGGCTTAAACTG</td>
<td>CAGAAATCAGGGGCGCTCAAC</td>
<td></td>
</tr>
<tr>
<td>60 bp 5’</td>
<td>CTACCTGGTCTCCAGGAAAC</td>
<td>CTAGACCTCTGTCTGCTTACAC</td>
<td></td>
</tr>
<tr>
<td>120 bp 5’</td>
<td>GGGAAAGAGGACCTTTGAG</td>
<td>GGCCTCCCTACTTACTCAG</td>
<td></td>
</tr>
<tr>
<td>750 bp 5’</td>
<td>GAAGCTGGATTGCTTGGTATTAG</td>
<td>CTCTGTACGTAGAGGGATTAC</td>
<td></td>
</tr>
<tr>
<td>20 kb 5’</td>
<td>GAACCCAAAGGGACCTGAAAC</td>
<td>CACGACCTCTTGGTCACTTA</td>
<td></td>
</tr>
</tbody>
</table>
**EXPERIMENTAL MODEL AND SUBJECT DETAILS**

**Cell lines**
Female Atr^frx/-Cre-ERT2^ mouse embryonic fibroblasts (4-3 cells) (Smith et al., 2009); Tim knockdown (shTIM1) 4-3 cells (Smith et al., 2009); I-Pool 4-3 cells were generated by transducing retrovirus expressing the I-Pool restriction enzyme from the pBabe-ddIPpol plasmid (Addgene plasmid #49052, (Goldstein et al., 2013)) into 4-3 cells. 4-3 cells used for BrITL were stably transduced with retrovirus expressing Bcl-xL (Addgene plasmid #8790, 3541 pMIG Bcl-XL) to decrease the amount of apoptotic cells collected. MDA-MB-231 cells (Laboratory of Andy J. Minn, University of Pennsylvania). U-2OS (HTB-98) osteosarcoma cells (female) were purchased from and validated by American Tissue Type Collection.

**Treatment**
4-3 cells were treated with either DMSO or 1 μM ATR-45 (Charrier et al., 2011) and 0.2 μM aphidicolin (Calbiochem, CAS 38966-21-1) for 18 hr; 1 μM ATR-45 and 0.2 μM aphidicolin for 9 hr; 1 μM ATR-45 for 18 hr; and 0.2 μM aphidicolin for 18 hr. I-Pool 4-3 cells were treated with 1 μM Shield-1 (Laboratory of Tom Wandless, Stanford), and 0.5 μM 4-hydroxytamoxifen (4-OHT, EMD Chemicals/Calbiochem) for 14 hr to induce nuclear expression of I-Pool. Parental 4-3 cells were similarly treated with 1 μM Shield-1 and 0.5 μM 4-OHT for 14 hr as a control. MDA-MB-231 cells were treated with either DMSO or 0.5 μM VE-822 (Selleck Chemicals, S7102) and 0.2 μM aphidicolin (Calbiochem, CAS 38966-21-1) for 9 hr.

**Culture**
All cells were grown in Dulbecco’s Modified Eagle’s medium (DMEM, Mediatech) supplemented with 10% fetal bovine serum (FBS, Benchmark, Gemini BioProducts), L-glutamine (2 mM, Mediatech), and streptomycin/penicillin (100 U/ml, Thermo Fisher Scientific) at 37° C/5% CO2. U-2OS cells were cultured in buffered Dulbecco’s Modified Eagle Medium (GIBCO), supplemented with 10% FBS (Hyclone), at 37° C/5% CO2.

**METHOD DETAILS**

**RPA ChIP-Seq**
This assay was performed on 4-3 cells under the following conditions: 1) 1 μM ATR-45 + 0.2 μM aphidicolin for 18 hr; 2) 1 μM ATR-45 + 0.2 μM aphidicolin for 9 hr; 3) 1 μM ATR-45 for 18 hr; 4) 0.2 μM aphidicolin for 18 hr; 5) DMSO; and 6) Tim knockdown 4-3 cells treated with 1 μM ATR-45 for 14 hr. For each condition described, two biological replicates were performed, except for 1), in which three replicates were obtained.

For each immunoprecipitation reaction, 15 × 10^6 cells were trypsinized, collected, spun down and re-suspended in 25 mL PBS. Cells were fixed in 1% formaldehyde for 10 minutes at 37° C and the reaction was stopped by adding glycine to 1% final concentration. The cell pellet was washed in 10 mL PBS and subsequently re-suspended in 1 mL cold PBS. The pellet was then lysed in lysis buffer (50 mM HEPES pH 7.9, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NPZ40, 0.25% Triton X-100) for 10 minutes on ice. The nuclei were recovered by spinning and washing twice (10 mM Tris-Cl pH 8.1, 200 mM NaCl, 1 mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0). The nuclei were re-suspended in 1 mL shearing buffer (0.1% SDS, 1 mM EDTA, 10 mM Tris pH 7.6, and chromatin sheared using Covaris S220 to < 4 kb using parameters according to the company hand book. Buffer (0.1% SDS, 1 mM EDTA, 10 mM Tris pH 7.6, 11% Triton X-100, 1% Na-DOC) was added to 1/10 volume to keep the sample in Radioimmunoprecipitation assay (RIPA) buffer. Dynabeads Protein A beads were pre-bound the night before by mixing 1 mL PBS, 10 μl 100 mg/ml BSA in PBS, 20 μg anti-RPA32 antibody (NA19L, Millipore), and 10 μg bridging antibody, rotating overnight at 4° C. The next day, the beads were washed as follows: 2x with 1 mL of RIPA buffer, 2x with 1 mL of RIPA buffer + 0.3 M NaCl, 2x with 1 mL of LiCl buffer (0.25 M LiCl, 0.5% NP-40, 0.5% NaDOC, stored at 4° C), 1x with 1 mL of TE (pH 8.0) + 0.2% Triton X-100, 1x with 1 mL of TE (pH 8.0). The beads were then incubated with Proteinase K at 65° C to reverse cross-link. DNA was extracted using phenol/chloroform and precipitated with ethanol/sodium acetate. Pellets were re-suspended in TE (pH 8.0) and prepared into libraries for NGS analysis, described below.
**BrITL** –
This assay was performed on 4-3 Bcl-xL cells treated with 1 μM ATR-45 (ATRI) + 0.2 μM aphidicolin, or DMSO (control), for 18 hr and on MDA-MB-231 cells treated with 0.5 μM VE-822 (ATRI) + 0.2 μM aphidicolin, or DMSO (control), for 9 hr. Two biological replicates of each condition were performed and prepared into libraries for NGS analysis, described below. This assay was also performed on I-PpoI-transduced 4-3 cells and parental 4-3 cells treated with 1 μM Shield-1 and 0.5 μM 4-hydroxytamoxifen, or DMSO (control), for 14 hr. Three biological replicates of each condition were performed and processed for qRT-PCR analysis, described below.

For each BrITL reaction, ~2 × 10⁶ cells were trypsinized and collected in an Eppendorf tube. Cells were washed with PBS, permeabilized in 0.1% Triton X-100 in PBS for 5 minutes on ice and subsequently washed with 0.01% Triton X-100 in PBS. Cells were incubated in a reaction containing 20 μM ddNTPs (Affymetrix, 77126) in 1x NEBuffer 2 for 5 minutes at 37°C. The reaction was stopped with 20 mM EDTA. Cells were washed four times with 0.01% Triton X-100 in PBS before resuspending the cell pellet in a reaction mixture containing 2.5 mM CoCl₂ (Roche, 11243306001) and 27 μM biotin-16-ddUTP (Enzo Life Sciences, ENZ-42813) in 1x TdT buffer (Roche, 11243276001). Upon addition of 150 units of TdT (Roche, 03333566001), the end-labeling reaction proceeded for 1 hour at 37°C. Cells were then washed twice with 50 mM EDTA in 0.01% Triton X-100 in PBS. To lyse the cell pellet, TNE buffer (50 mM Tris-HCl pH 7.4, 100 mM NaCl, 0.1 mM EDTA) was added together with 10% SDS and 10 mg/ml Proteinase K for incubation overnight at 37°C. The next day, genomic DNA was extracted using phenol/chloroform followed by ethanol/sodium acetate precipitation. The pellet was re-suspended in TE (pH 8.0). Sonication occurred in the Biorupter (Diagenode) for 2 minutes on medium setting to obtain 0.2-2 kb fragments. After sonication, the samples were purified with Ampure XP beads (Beckman Coulter, A63880), utilizing 0.8x SPRI:DNA ratio. Washed and dried beads were incubated with EB buffer and left at room temperature for up to an hour before placing at 4°C overnight. The next day, the eluate was retrieved from the beads and brought up to 100 μl volume with TE. From this volume, 15 μl was aliquoted into a separate tube containing 85 μl TE and stored at 4°C to serve as the input. The rest of the sample was brought up to 200 μl with TE and proceeded to the next steps for retrieval.

Selection of biotin-labeled DNA fragments was performed with the Dynabeads KiloBaseBinder kit (Life Technologies, 601-01). For this, 25 μl of streptavidin-coated magnetic beads were washed twice with 200 μl Binding buffer containing 5 μg/ml tRNA. The beads were then mixed in 200 μl sample plus 200 μl Binding buffer and left at room temperature on a rotating wheel for 2 hr. The samples were then placed against a magnetic stand and the supernatant discarded. The beads were washed twice with wash buffer (10 mM Tris-HCl pH 7.5, 1 mM EDTA, 2 M NaCl) with 5-minute rotations at room temperature for each wash. The beads were then transferred to a new tube containing wash buffer with 4 μg/ml tRNA and subsequently washed in distilled autoclaved water twice. The washed and dried beads were then re-suspended in TE. To these samples, 20 μg boiled RNase A was added. Samples were incubated at 37°C for 30 minutes to remove RNA contaminants. Elution of biotinylated fragments bound to streptavidin-coated magnetic beads occurred by adding 1% SDS and 1 mg/ml Proteinase K to the samples and incubating at 55°C overnight. The next day, DNA was purified by sequential phenol, phenol/chloroform, and chloroform extraction before subsequent ethanol/sodium acetate precipitation. The DNA pellet was re-suspended in 50 μl of TE.

**qRT-PCR analysis** –
Real-time PCR was performed on the Applied Biosystems 7900HT Sequence Detection System. All PCR reactions were performed in duplicate for each sample with 0.9 μM of the forward and reverse primers in a final volume of 10 μl using SYBR Green PCR Master Mix (Applied Biosystems, 4309155). Primer sets are described in Key Resources Table.

**NGS library preparation** –
Libraries were prepared for both input and IP DNA according to the NEBNext kit. Briefly, DNA was sonicated to ~200 bp. DNA was end-repaired using a combination of T4 DNA polymerase, Klenow DNA polymerase and T4 polynucleotide kinase. The blunt, phosphorylated ends were treated with Klenow fragment (3‘ to 5’ exo minus) and dATP to yield a protruding ‘A’ base for ligation of Illumina’s adapters which have a single ‘T’ base overhang at the 3’ end. After adaptor ligation, DNA fragments of ~200 bp (insert plus adaptor) were band-isolated from a 2% agarose gel. The purified DNA was PCR amplified with Illumina primers for 18 cycles. The purified DNA was captured on an Illumina flow cell for cluster generation and sequenced on Illumina HiSeq 100 bp single-end following the manufacturer’s protocols.

**Transfection of (CAGAGG)₁₀₅ or scrambled sequence vectors** –
HFUGW lentiviral expression vectors containing either the (CAGAGG)₁₀₅ repeat or a scrambled sequence containing the same nucleotide content and length (Key Resources Table) were linearized by digestion with Ndel and transfected into 4-3 cells using lipofectamine transfection (Thermo Fisher Scientific, L3000001). GFP⁺ cells were sorted on a FACSaria machine to select for cells containing integrated HFUGW vector sequences, which express EGFP through the UBC promoter.

**Southern Blot** –
Southern blot performed on HFUGW-transfected cells utilized a biotinylated probe (Biotin DecaLabel DNA Labeling Kit; Thermo Fisher Scientific, K0651; HFUGW probe primer set described in Key Resources Table) between the PstI sites of the HFUGW vector. Briefly, 10 μg of genomic DNA from transfected cells were digested with 60 units of PstI for 2 hours at 37°C. After phenol/chloroform extraction and ethanol/sodium acetate precipitation of digested DNA, DNA was run on an 0.8% TBE gel overnight at 20 V. Gel was
soaked in denaturization buffer (0.5 M NaOH, 1.5 M NaCl, pH 13.0) for 30 minutes twice. The gel was then rinsed with dH₂O before being soaked in neutralization buffer (0.5 M Tris-HCl (pH 7.5), 1.5 M NaCl) for 15 minutes twice. Gel was rinsed with dH₂O before being soaked in 20X SSC buffer for 30 minutes. Transfer to Hybond N+ membrane (Fisher Scientific, 45-000-927) occurred overnight in 20X SSC buffer. The membrane was UV-crosslinked (120,000 μJ). The blot was then pre-hybridized with boiled 100 μl of 10 mg/ml salmon sperm in 15 mL of pre-hybridization buffer (5X SSC, 5X Denhardt’s solution, 0.5% SDS) for 3 hours at 65°C. Probe mix was made by boiling 20 μg of mouse COT-1 DNA with 1 μg of biotinylated probe for 10 minutes, placed on ice for 2 minutes, and incubated with 1 mL pre-hybridization buffer for 1 hour at 37°C. The probe mix was then added to the blot and rotated overnight at 65°C. The next day, the membrane was washed with 0.4X SSC, 0.1% SDS for 15 minutes at 65°C and three more times for 30 minutes each at 65°C. Hybridized biotinylated probe on the membrane was detected using the Biotin Chromogenic Detection kit (Thermo Fisher Scientific, K0661).

Bioinformatics -

**Peak-calling**

ChIP and BrITL libraries from at least 2 biological replicates of each experimental condition and their respective inputs were sequenced through Illumina HiSeq, generating 100 bp single-end sequencing reads. Adaptor contamination in reads were trimmed using trimmomatic (Bolger et al., 2014) and reads were checked for quality control using fastqc (Leggett et al., 2013). Alignment was made to the mm10 reference genome of mouse samples or hg19 reference genome for human samples using STAR-2.5.2a Aligner with at most 3 mismatches (Dobin et al., 2013). Reads that contain non-unique sequences were initially allowed to be placed in up to 100 different genomic regions in order to later measure differences in regional read accumulation between multi-mapping of a single read up to with 100 placements and mapping of a single read to its most likely genomic placement. In the context of these experimental regions, which consist of non-unique repetitive genomic regions, measuring the difference between tracks with reads that are mapped up to 100 different placements and those with reads that are placed in their most likely home will reveal enrichment bias toward low complexity regions (i.e., if it is solely due to the low complexity nature of these regional sequences and not due to the experimental enrichment). Reads were then filtered by mapq score 10 to keep the high-confident read mappings. De-duplication of reads in the aligned tracks took place to increase the complexity of the read population. Additional alignment-specific quality-control metrics were conducted, including strand-cross-correlation (Landt et al., 2012), finger-plots (Ramírez et al., 2016) to gauge mutual back-level of enrichment across samples, Pearson and Spearman correlations of genomic and enriched regions across samples (≥ 0.6), principal component analysis (PCA) for clustering assessment, and a non-arbitrary estimate of ChIP signal over input using an NCIS-generated normalization ratio (Liäng and Keleş, 2012). Black-listed regions in the mm10 and hg19 genome were filtered out prior to peak-calling.

For enrichment analysis, the biological replicates and inputs of each experimental condition underwent an irreproducibility rate (IDR) analysis (Landt et al., 2012) from the ENCODE project with the MACS2 peak-calling program (Zhang et al., 2008) to give the final peak list per condition. IDR thresholds of > 0.05 were used for self-consistency within each biological replicate and for comparison between biological replicates, and > 0.005 for pooled-consistency analysis. Peaks that passed IDR thresholds and that were within a defined region (up to 5 kb) were merged. For RPA ChIP-Seq on 4-3 cells, peaks with p value < 10⁻⁵⁰ and > 4-fold signal enrichment over input were selected. For BrITL on 4-3 cells, peaks with p value < 10⁻⁶ and > 4-fold signal enrichment over input were selected. For BrITL on MDA-MB-231 cells, peaks with p value < 10⁻⁶ and > 4-fold signal enrichment over input were selected. Peaks here were further selected that were greater than or equal to 500 bp, which is the average fragment size for BrITL sonication and retrieval. For all experiments, the final peak list was generated by subtraction of peaks that were also identified in the DMSO-controls. Ratio tracks were generated using deepTools: bigwigCompare, 500 bp bin size.

**Enrichment of complex repeats in RPA ChIP samples**

Trimmed fastq reads from each RPA ChIP-Seq sample that overlapped with different families of complex repeats (LSU_rRNA, SSU_rRNA, tRNA, etc.) were counted for each family of repeats. These numbers were then divided by the total number of reads with at least one reported alignment in each sample. Values from different biological replicates in each condition (ATRi+aph18hrs and DMSO (UT)) were averaged and normalized by the values calculated in the respective input samples. The resulting fold over input values for each family of complex repeats were graphed for each condition.

**REQer**

To understand simple repeat sequences that may be enriched in the experimental conditions relative to input, an assessment of the sequence presence within individual reads was performed. Reads in fastq files were labeled according to how many times a repeat occurs as a single unit (monomer), or as different tandem units based upon the maximum length of the repeat in a 100 bp read, using a python script that incorporated regular expressions. This program was called REQer.

Tandem simple repeat analysis that measured repeats occurring purely in tandem was performed on trimmed and de-duplicated RPA-ChIP Seq reads from combined biological replicates of each condition (ATRi+aph18hrs and DMSO (UT)). The analysis was conducted by counting the number of different tandem lengths of a particular repeat sequence present in the reads and measuring the frequency of each length by dividing by the total count of the repeat sequence present in the reads. The ratio of the calculated frequency in each condition’s ChIP retrieval reads over the frequency calculated in their respective input reads yielded fold over input enrichment values.

Molecular Cell 72, 222–238,e1–e11, October 18, 2018 e7
Total simple repeat analysis of trimmed and de-duplicated RPA-ChIP Seq reads from combined biological replicates of each condition (ATRi+aph1bri and DMSO (UT)) was conducted by categorizing reads by the total unit count of a particular repeat sequence within a read. The frequency was calculated as the fraction of reads that contained the specified count within the total number of reads from the combined replicates. The ratio of the calculated frequency in each condition’s ChIP retrieval reads over the frequency calculated in their respective input reads yielded fold over input enrichment values.

Fork-pausing Experiments – Plasmid constructions

630 bp of CAGAGG tandem repeats were cloned into the BspEI and BssHII site of the pML113 plasmid in opposite orientations (Follonier et al., 2013) for the origin-proximal insertion, and into the BamHI site for the origin-distal insertion. Randomized controls of the same nucleotide composition and length were similarly constructed.

In vitro assay

Templates for polymerase reactions were created by cloning [CAGAGG]15 repeats into the MCS/BamH1 site of the pGEM3Zf(-) vector (Promega, P2261). Inserts in two orientations were isolated in order to purify ssDNA templates of both strands. As controls, randomized sequences of the same nucleotide composition and length were similarly cloned. Subsequently, a C to A mutation at the 5’ BamH1 site (GGATCC) and a G to T mutation at the 3’ BamH1 site (GGATCC) flanking the repeat insert were introduced, in order to disrupt the potential for G-quadruplex formation between the vector and insert sequences. For each construct, single-stranded DNA was isolated after R408 helper phage (Promega, P2291) infection of plasmid-bearing SURE cells (e14-(McrA-); Agilent Technologies, 201052). Log phase plasmid-bearing SURE cells in 2XYT media were infected with 1/50th volume of R408 (titer of phage stock was > 1 x 1011 plaque forming units (pfu)/mL) and incubated in a 37°C shaker for 3 hours to overnight. An overnight incubation was necessary for optimal yields of ssDNA from the [CAGAGG]15 strand. After pelleting the bacterial cells, virus particles in the supernatant were precipitated on ice for 30 min with a polyethylene glycol (Sigma, P5413)/ammonium acetate solution at final concentrations of 4% and 0.75M, respectively. Virus was pelleted and resuspended in an appropriate volume of Phenol Extraction Buffer (PEB; 100mM Tris, pH 8.0, 300mM NaCl, 1mM EDTA, pH 8.0). DNA was extracted one time with two volumes of phenol (Affymetrix/Thermo-Fisher, AAJ5829AN) satu rated with PEB, one time with one volume of phenol, and one time with half volume 24:1 chloroform: isooamyl alcohol. After extraction, DNA was precipitated with ammonium acetate at 2.0M final concentration and 2 volumes of ethanol and resuspended in 10mM Tris and 1mM EDTA, pH 8.0 (Eckert and Gestl, 2010). Small ssDNA preparations from independent clones were sequenced to verify integrity of the insert prior to large scale purification of ssDNA templates. Repeat lengths longer than 15 units precluded our ability to rescue ssDNA of the correct sequence and/or length.

DNA synthesis templates were created by 32P end-labeling (1000Ci/mmol; Perkin-Elmer, BLU002Z001MC) a PAGE-purified 43-mer oligonucleotide (G94-43-mer, Integrated DNA Technologies) using T4 Polynucleotide Kinase (Thermo-Fisher, 18004010) according to the manufacturer’s instructions and hybridizing to ssDNA at a 1:1 molar ratio in 1X SSC buffer (150mM NaCl and 15mM sodium citrate). The G94-43-mer oligonucleotide initiates synthesis 68 nucleotides downstream of the repeat insert. To remove unincorporated radionucleotide, the hybridized primer-template were purified over illustra Microspin G-50 columns (GE Healthcare, 27-5330-01). Primer extension reactions contained 100 fmol of primed ssDNA substrate, 400 fmol human recombinant PCNA (Xu et al., 2001), 1700 fmol yeast RFC (Thompson et al., 2012), and 1700 fmol yeast RFC (Thompson et al., 2012) in 5 mM Tris HCl, pH 7.5, 8 mM MgCl2, 5 mM DTT, 40 μg/ml BSA, 150 mM KCl, 5% glycerol, 0.5 mM ATP, and 250 μM dNTPS, and were preincubated at 37°C for 3 min. Synthesis was initiated upon addition of 100 fmol purified 4-subunit recombinant human Pol δ4 (Zhou et al., 2012). Aliquots were removed at 3, 7, and 15 minutes, quenched in 1 volume STOP dye (Formamide, 5mM EDTA pH 8.0, 0.1% xylene cyanol, 0.1% bromophenol blue) and reaction products were separated on an 8% denaturing polyacrylamide gel and quantitated using a Molecular Dynamics STORM 860 Phosphorimag er. A control for the percent of primers productively hybridized to each primer-template substrate (% Hyb) was performed using excess Exo-Klenow polymerase (Affymetrix/Thermo-Fisher, 70057Z), and a background control for primer impurities (no Pol) was performed by incubating unextended primer-template substrate in reaction buffer without addition of polymerase. Dideoxy sequencing reactions were carried out simultaneously with the Pol δHE reactions, using the same primer-template substrates and Sequenase 2.0 (Affymetrix/Thermo-Fisher, 70775Y). Total percent extension was calculated as the amount of total extended DNA molecules (corrected for percent hybridization and background) divided by the number plus the amount of corrected primer molecules. The number of DNA molecules within four regions were determined from the 15 minute reaction using ImageQuant 5.2 software quantitation: (R1) 68-11 bases 5’ to the insert; (R2) 10-1 bases 5’ to the insert; (R3) the insert; and (R4) all bases 3’ to the insert up to and including the well. After background correction, the termination probability within each region was calculated as the number of molecules within the region divided by the number of molecules within the region plus all longer molecules. To normalize for the different sizes of Regions 1-3, each region’s termination probability was divided by the number of nucleotides under consideration. For example, the termination probability/nt for Region 1 = [molecules in R1 ÷ molecules in R1+R2+R3+R4] ÷ 58 nucleotides.

Ex vivo assay

The SV40-derived pML113, 114 and 115 vectors (Follonier et al., 2013) were gifts from Massimo Lopes (University of Zurich). For the ori-proximal vectors, a 630 bp fragment encoding CAGAGG tandem repeats (Key Resources Table) was cloned into the pML114 and 115 plasmids using the MCS/BspEI and BssHII sites, creating plasmids with the repeats in two orientations. As controls, randomized sequences of the same nucleotide composition and length were created (Key Resources Table) and similarly subcloned into...
pML114/115 vectors. For the ori-distal vectors, the repeats were cloned into the BamHI site of pML113, in two orientations. Subconfluent U-2OS cells (ATCC, HTB-96) were transfected with 5 μg vector DNA using Xtreme Gene XP (ver. 8) transfection reagent (Roche). To induce replication stress, cells were treated with 0.6 μM Aphidicolin (Sigma; DMSO solvent) 24 hours post-transfection, or 0.5 μM VE-822, 21 hours post-transfection, followed by Aphidicolin treatment three hours later. For all experiments, DNA was isolated 48 hours post-transfection using a modified Hirt method, as described (Chandok et al., 2011). Briefly, adherent cells were washed with 1M Tris-buffered saline (50 mM Tris-HCl, pH 7.0 and 150 mM NaCl) followed by lysis in 50 mM Tris–HCl pH 7.0, 20 mM EDTA, 10 mM NaCl, 10% SDS, 0.2 mg/ml proteinase K (Sigma-Aldrich). Chromosomal DNA was precipitated by incubation of lysate overnight in 5M NaCl followed by centrifugation at 27,200 xg for 50 minutes at 4°C. Resulting supernatants were incubated in the presence of ~0.1 mg/mL proteinase K for 2-3 hours at 55°C. DNA was extracted by phenol–chloroform extraction and precipitated in an equal volume of isopropanol with 0.5 μL polyacryl carrier (Molecular Research Center, Cincinnati, OH). To analyze replication intermediates, the purified DNA was digested with DpnI, EcoRI, and EcoN1 (ori-proximal vectors; New England Biolabs) or DpnI, PpuMI, and SacII (ori-distal vectors; New England Biolabs) for 3 hours, followed by ethanol precipitation. DNA products were resuspended in Tris-EDTA and separated first through a 0.4% TBE agarose gel (1V/cm, 14 hr, room temperature, -EtBr) and second through a 1% TBE agarose gel (4V/cm, 6-8hr, 4°C, +EtBr) (Friedman and Brewer, 1995). DNA fragments were transferred by capillary action overnight to a Hybond-N+ membrane (Amersham/GE Healthcare). After UV crosslinking, pre-hybridization of membranes was carried out by incubation with a 0.25M sodium phosphate (pH 7.2)/ 7% SDS/ 100 μg/ml sonicated sperm DNA buffer, and incubation at 65°C, ≥ 2 hr. DNA probes were made by isolating the indicated restriction fragments and using a random labeling method (High Prime; Roche) with a32P dCTP (6000Ci/mmol); Perkin-Elmer) followed by purification using a Microspin G-50 column (GE Healthcare). Membranes were hybridized using 100-200 ng probe at 65°C overnight (with rotation), washed according to manufacturer’s instructions, and imaged using a Molecular Dynamics STORM 860 Phosphoimager.

Biophysical Characterization of DNA Secondary Structure - DNA and buffers for structural studies

All DNA sequences used for biophysical characterization are summarized in Table S3. DNA samples were ordered from IDT (Texas, USA) as HPLC purified samples, dissolved in water at 1 mM final concentration and stored at~ -80°C. Samples were diluted to the desired concentration into final 10 mM lithium cacodylate buffer pH 7.2 supplemented with 100 mM KCl and 2 mM MgCl2 (100K2Mg buffer). Samples were annealed at 90°C for 5 minutes, cooled slowly to room temperature over the course of 3-4 hr and equilibrated overnight at 4°C. All samples were examined with circular dichroism (CD) for consistency in folding.

Circular Dichroism wavelength scans

All experiments were performed on an Aviv 410, Aviv 435, or a Jasco 815 spectropolarimeter with a Peltier heating unit using 1 cm quartz cuvettes. The accuracy of the external temperature probe was ± 0.3 K. Each CD trace was an average of 3 - 5 scans collected from 220 to 330 nm with 1-2 nm bandwidth, 1 nm step, 1 s averaging time at 4°C. CD data were treated as described elsewhere (Niculids et al., 2012).

Thermal stability of DNA via CD melting

Thermal denaturation experiments were collected at the maxima of CD wavelength scans with 2 nm bandwidth, 5 s equilibration time, 1°C step, and 15 or 20 s averaging time. The samples were heated from 4°C to 95°C, maintained at 95°C for less than 5 minutes, and cooled to 4°C at the same rate. The cooling step was included to determine the reversibility of folding/unfolding process. Superposition of melting and cooling data suggested that the melting process was reversible. Melting data were analyzed assuming a two-state model with constant enthalpy, ΔH (heat capacity, Cp = 0) (Ramsay and Eftink, 1994). This model suggests that at any point during melting or cooling only folded and unfolded DNA is present (no intermediates, i.e., two-state system). Starting and final baseline (assuming to be linear), melting temperature (Tm) and enthalpy of unfolding were adjusted to get the best fits. The melting curves were also analyzed by the derivative method. The Tm was determined as the minima or maxima on the first derivative curves. These values were read by eye and are associated with an error of ±0.5°C. When the melting transition was poorly defined, the derivative method was deemed more reliable. Melting and cooling data were processed separately and the hysteresis (difference in Tm between the heating and cooling curves) was determined.

Thermal stability of DNA via UV-vis melting

Studies were performed on a Cary 300 spectrophotometer thermostated with an external waterbath (temperature accuracy of ± 0.3°C). We monitored 260, 295, and 330 nm wavelengths. The latter wavelength was used as a reference to monitor instrument performance; the extinction coefficient of DNA at 330 nm is negligible. Melting studies were completed in 1 cm pathlength quartz cuvettes, data were typically collected from 4°C to 95°C and back to 4°C with 2 nm bandwidth, 1°C temperature step, 2 s averaging time, and 0.75 or 1°C/min rate. Melting and cooling curves were collected to determine the reversibility of the melting process. The temperature was measured with the temperature sensor inserted into the cuvette filled with water and placed next to DNA samples. The signal at 330 nm was subtracted from each dataset. The resulting data were analyzed as described above. The results from UV-vis study are in great agreement with the results from CD melting study.

Molecularity of DNA structures via UV-vis melting

Samples were annealed in 100K2Mg buffer at concentrations ranging from 1.4 to 23.6 μM for CA5 and from 0.3 to 6.9 μM for CA10; targeted concentration ratio between the most dilute and the most concentrated sample was ~20. Samples were placed in quartz...
cuvettes with 1.0 or 0.2 cm pathlength depending on strand concentration and equilibrated at 4°C for at least 20 minutes. Melting experiments and data analysis was performed as described above.

**Polyacrylamide gel electrophoresis**

Native PAGE gels were typically prepared at 12% polyacrylamide in 1xTAC (50 mM Tris Acetate pH 7.3) buffer supplemented with 3 mM MgCl₂. Running buffer consisted of 1xTAC with 3 mM MgCl₂. Gels were cooled with a water bath and premigrated for 30 minutes at 140 V. Each DNA sample of 10 μL contained ~3 μg of annealed oligo in 100K2Mg buffer to which 3 μL of 50% w/v sucrose was added immediately prior to loading. Oligothymidylate markers S’ dTₙ (where n = 15, 24, 30, 57 or 60, and 90) as well as a 76-nt tRNA were used as internal migration standards. Typically, gels were run for 3-4 hr at 140-300 V; gel temperature did not exceed 16°C. Gels were stained using Stains-All and de-colored in water under visible light. Gels were visualized on ETNA-NS ChemiBis 3.2 gel visualization device (using lower light, 580 nm filter) or with an iPhone 5 camera.

**QUANTIFICATION AND STATISTICAL ANALYSIS**

Statistical details can be found in figure legends and in Method Details.

**RPA ChiP-Seq**

For REQer analysis, ChiP-Seq reads from two biological replicates of ATRi+aph18hrs were combined and quantified for short tandem repeat presence. This was compared against similarly quantified ChiP-Seq reads from the inputs, allowing for calculation of the fold enrichment of short tandem repeats in ChiP reads over input reads. For these values graphed at the minimum and maximum range of repeat units within a sequencing read (Figures S3B–S3D), p values were obtained at 95% confidence interval using the Kolmogorov-Smirnov test between the distributions of ATRi+aph18hrs and its input.

To assess the statistical significance of the overlap of RPL or BrITL peaks with specific genomic features (Table 1 and Figure 2A), permutation tests were performed in which RPL or BrITL peaks were randomly shuffled across the mouse or human reference genome 100,000 times. The number of overlapping peaks between each set of shuffled RPL or BrITL sites and the associated dataset was determined, thereby generating an empirical distribution of the overlap under the null hypothesis that the RPL or BrITL peaks were randomly distributed throughout the genome. The p value of each comparison is the number of times the overlap between the two datasets occurred more than observed. A small p value indicates that the observed overlap between RPL or BrITL sites and the associated dataset is statistically significant and is not due to random chance. For box-plot analysis of short tandem repeat length within RPL peaks identified from ATRi+aph18hrs versus the background genome (Figure 2G), a two-sided Wilcoxon test was performed.

**BrITL**

For qPCR graphs, experiments were performed in triplicate and the data are represented as mean ± SEM. For the significance of breakage associated with transfected sequences (scrambled or (CAGAGG)₁₀₅) by BrITL qRT-PCR, dot plots were made and a Student’s t test was performed between the averages of each condition (n = 3 for Scrambled, DMSO; n = 5 for Scrambled, ATRi+aph; CAGAGG, DMSO; CAGAGG, ATRi+aph), where n is the number of replicates from each condition. Outliers were not included in the calculation of the averages.

For bar graph analysis of repeat count within human BrITL peaks identified from ATRi+aph9hrs versus the count in the background genome (Figure 7C), the following calculations were performed to obtain statistical values: bedtools shuffle and bedtools getfasta generated randomly selected genomic regions (hg19) (excludes unmappable regions) with equivalent footprint size as the human ATRi+aph9hrs BrITL peaks (5,268,109 iterations). The MISA- and HOMER2-identified motif repeats were then counted throughout each randomly generated peak list. The occurrences of the repeat count being <= or ≤ than the repeat count in the ATRi+aph9hrs BrITL peaks were recorded. The empirical p value was then calculated by dividing these occurrences by the number of randomly generated peak sets (5,268,109).

**Biophysical Characterization of DNA Secondary Structure** –

Each biophysical experiment was done a minimum of three times for each oligo using at least two independently synthesized DNA stocks. Peak molar ellipticity values, melting temperatures, and enthalpies of unfolding were determined as the arithmetic mean of these trials. Differences in the normalized termination probabilities among template sequences and various regions were analyzed statistically using a two-way ANOVA and significance was defined as producing a p value of < 0.05. The columns in Figure 4C indicate the mean of the three replicate reactions and error bars indicate the SEM. The statistical details can be found in the Results section and in the legend to Figure 4.

**Fork-pausing Experiments** –

**In vitro assay**

Three independent PolⅢHE reactions were performed for each primer-template examined. Differences in the normalized termination probabilities among template sequences and various regions were analyzed statistically using a two-way ANOVA and significance was defined as producing a p value of < 0.05. The columns in Figure 4C indicate the mean of the three replicate reactions and error bars indicate the SEM. The statistical details can be found in the Results section and in the legend to Figure 4.
**Ex vivo assay**
ImageQuant 5.2 software was used to determine the amount of radioactivity in the descending arm of the simple Y arc, relative to radioactivity in the double Y arc extending above the simple Y apex. Background radioactivity was calculated and subtracted for each blot prior to calculating the Replication Fork Barrier (RFB) index. RFB values shown are the average of two or three biological replicates (independent U2OS cell transfections).

**DATA AND SOFTWARE AVAILABILITY**

**Mendeley Deposed Data**
https://doi.org/10.17632/9gwyx4vxt8.1

**Accession Numbers**
The data discussed in this publication have been deposited in NCBI’s Gene Expression Omnibus and are accessible through GEO SuperSeries accession number GSE115625.

Subseries linked to GSE115625 are as follows:
- RPA-ChIP Seq data of MEFs are deposited under accession number GSE100790.
- BrfTL data of MEFs are deposited under accession number GSE115624.
- BrfTL data of MDA-MB-231 cells are deposited under accession number GSE115623.

**ENCODE Mouse Datasets –**

**DNase hypersensitivity**
- ENCODE Project Consortium, 2012; wgEncodeEM001936; Stamatoyannopoulos Laboratory, ENCODE/UW
- H3K4me3 ChIP-Seq
- ENCODE Project Consortium, 2012; wgEncodeEM001455; Ren Laboratory, ENCODE/LICR
- H3K27me3 ChIP-Seq
- ENCODE Project Consortium, 2012; wgEncodeEM002501; Ren Laboratory, ENCODE/LICR

**Replication timing (Repli-Chip):**
- ENCODE Project Consortium, 2012; wgEncodeEM002974; Gilbert Laboratory, ENCODE/FSU
- CTCF ChIP-Seq
- ENCODE Project Consortium, 2012; wgEncodeEM001698; Ren Laboratory, ENCODE/LICR

**ENCODE Human Datasets –**

**H3K9me3 ChIP-Seq**
- ENCODE Project Consortium, 2012; wgEncodeEH002875; Farnham Laboratory, ENCODE/SYDH
- H3K27me3 ChIP-Seq
- ENCODE Project Consortium, 2012; wgEncodeEH002872; Farnham Laboratory, ENCODE/SYDH
- H3K27me3 ChIP-Seq
- ENCODE Project Consortium, 2012; wgEncodeEH002922; Farnham Laboratory, ENCODE/SYDH
- H3K36me3 ChIP-Seq
- ENCODE Project Consortium, 2012; wgEncodeEH002923; Farnham Laboratory, ENCODE/SYDH
- H3K4me3 ChIP-Seq
- ENCODE Project Consortium, 2012; wgEncodeEH000967; Stamatoyannopoulos Laboratory, ENCODE/UW
- CTCF ChIP-Seq
- ENCODE Project Consortium, 2012; wgEncodeEH002057; Stamatoyannopoulos Laboratory, ENCODE/UW