

Swarthmore College

## Works

---

Senior Theses, Projects, and Awards

Student Scholarship


---

Spring 2023

# Biophysical and Structural Characterization of Non-Canonical DNA structures

Paul C. Seth , '23

Follow this and additional works at: <https://works.swarthmore.edu/theses>

 Part of the [Chemistry Commons](#)

---

### Recommended Citation

Seth, Paul C. , '23, "Biophysical and Structural Characterization of Non-Canonical DNA structures" (2023). *Senior Theses, Projects, and Awards*. 268.  
<https://works.swarthmore.edu/theses/268>

Please note: the theses in this collection are undergraduate senior theses completed by senior undergraduate students who have received a bachelor's degree.

This work is brought to you for free by Swarthmore College Libraries' Works. It has been accepted for inclusion in Senior Theses, Projects, and Awards by an authorized administrator of Works. For more information, please contact [myworks@swarthmore.edu](mailto:myworks@swarthmore.edu).

# Biophysical and Structural Characterization of Non-Canonical DNA structures

**Paul Seth**

**April 17, 2023**

**Swarthmore College**

**Advisor: Liliya Yatsunyk**

# Acknowledgements

I would not have been able to get to where I am today without a great deal of help. I would like to first thank my advisor, Liliya Yatsunyk, for welcoming me into her lab and being a constant source of guidance and inspiration over the past year. I would also like to thank my wonderful labmates Kailey Martin, David Ye '23, Erin Chen '23, and Harrison Kim '24, who also welcomed me into the lab and made sure it was never a dull place to work. I would like to thank Kevin Li '22, who started the LHGQ project and passed it to me when he graduated. I would like to thank Eric Xing '26, who has helped collect data for the LHGQ projects.

I would also like to thank Dr. Daniela Fera, in whose lab I learned so much about what being a scientist at Swarthmore meant. I would like to thank the Hannay Chemistry Fellowship and Swarthmore college for supporting my summer research in 2022. I must also thank the wonderful professors and faculty of the Swarthmore College Chemistry and Biology departments, who have helped foster my love for science in my time at this school. And finally, I thank my friends and family for being behind me every step of the way.

## Abstract

G-quadruplex (GQ) and i-motif (iM) non-canonical DNA structures can be formed by G- and C-rich sequences within the genome, respectively. GQ and iM DNA play important roles in different biological processes and are promising therapeutic targets. Recently, the first left-handed GQ (LHGQ) was discovered and much remains to be learned about the structure and function of LHGQ DNA.

In this thesis, we investigate the biophysical and structural behavior of two LHGQ-forming sequences from the human genome, from the Human Solute Carrier 1 (SLC) and nuclear receptor binding SET domain (NSD) genes. We found that flanking bases play an important role in determining if a sequence forms LHGQ or right-handed GQ (RHGQ). We also investigated the effect of GQ ligand N-methyl mesoporphyrin IX (NMM) on LHGQ DNA. NMM can partially convert LHGQ structures to RHGQ. We report the structure of the SLC GQ in complex with NMM, which forms a GQ with both left and right-handed subunits. Finally, we present preliminary efforts towards crystallizing an iM from the human platelet derived growth factor gene.

The work in this thesis helps further our understanding of the rules that govern the formation of LHGQ DNA. The SLC+NMM structure represents the second reported right-left hybrid GQ and the first in complex with a ligand. These results are an important step towards the understanding of LHGQ DNA and potential development of drugs to harness the biological power of non-canonical DNA.

# **Table of Contents**

**Acknowledgements 2**

**Abstract 3**

**Table of Contents 4**

**List of Figures 6**

**List of Tables 6**

**List of Abbreviations 6**

**Ch 1: Introduction 8**

**Introduction to non-canonical DNA structures 8**

**GQs are non-canonical DNA structures. 8**

**iMs are non-canonical DNA structures 10**

**Biological relevance of GQ and iM DNA 11**

**Scope of work 14**

**Ch.2: Materials and methods 16**

**DNA, ligand, and buffers 16**

**UV-vis spectroscopy 16**

**Thermal difference spectra (TDS) 17**

**UV-vis melting studies 17**

**Circular dichroism (CD) scans 18**

**CD melting studies 18**

**Native polyacrylamide gel electrophoresis (PAGE) 19**

**Crystallization of SLC+NMM 19**

**Chapter 3: Biophysical and structural characterization of the SLC G-quadruplex in complex with NMM**

**21**

**Biophysical characterization of SLC constructs 22**

**Effect of NMM on right-left hybrid folding 25**

**SLC+NMM complex contains a right-left hybrid G-quadruplex 28**

**Conclusion 29**

**Chapter 4: Handedness of Homo sapiens nuclear receptor binding SET domain (NSD) G-quadruplexes. 31**

**Biophysical Characterization of Constructs 32**

**NMM partially induces right-handed folding in NSD-TT constructs. 35**

**Certain flanking sequences permit left-handed folding. 39**

**Crystallography of NSD-TT and NSD-TT1L+NMM 42**

**Conclusion 43**

**Chapter 5: Characterization of C-rich sequence from Homo sapiens platelet derived growth factor subunit**

**B (PDGF-B) 44**

**Biophysical characterization reveals that iM folding depends solely on the C-rich part of the sequence.**

**46**

**Conclusion 48**

**Chapter 6: Conclusions and future directions. 49**

**Appendix 55**

## List of Figures

Figure 1. G-quadruplex and i-motif structures.....	9
Figure 2. LHGQ structures.....	10
Figure 3. Sites of potential non-canonical DNA formation. ....	12
Figure 4. Structure of N-methyl mesoporphyrin IX (NMM).....	13
Figure 5. Biophysical characterization of SLC constructs.....	23
Figure 6. NMM Titration of SLC and SLC-1R.....	27
Figure 7. Crystal structure of SLC+NMM complex.....	28
Figure 8. Biophysical characterization of NSD constructs.....	33
Figure 9. Effect of NMM on NSD-TT construct folding. ....	36
Figure 10. NMM Titration of NSD-TT and NSD-TT1L.....	38
Figure 11. Biophysical characterization of NSD-TT overhang constructs .....	41
Figure 12. Biophysical characterization of PDGF-B. ....	47

## List of Tables

Table 1. SLC construct sequences and CD signatures. ....	22
Table 2. NSD construct sequences and CD signatures. ....	32
Table 3. NSD-TT constructs .....	40
Table 4. i-Motif DNA sequences .....	44

## List of Abbreviations

GQ - G-quadruplex.

iM - i-motif.

LHGQ - Left-handed G-quadruplex.

RHGQ - Right-handed G-quadruplex.

SLC - Solute Carrier 1.

NSD - Nuclear Set Domain.

**NMM - N-methyl mesoporphyrin.**

**PDGF- Platelet-derived growth factor.**

**TDS - Thermal difference spectra.**

**CD - Circular dichroism.**

**PAGE - Polyacrylamide gel electrophoresis.**

**TBE - Tris-borate-EDTA.**

**NaCaCo - Sodium Cacodylate.**



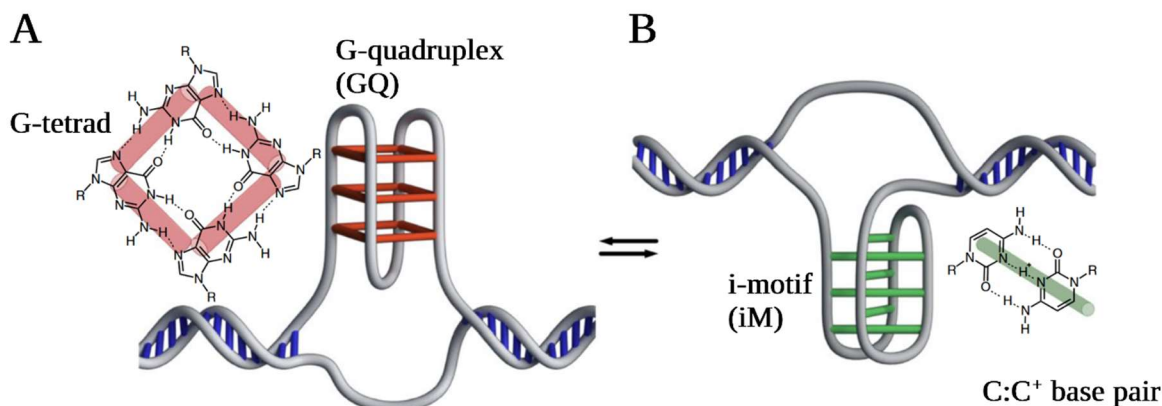
## Ch 1: Introduction

### Introduction to non-canonical DNA structures

The best-known form of DNA is the double helix. In this canonical conformation, two strands of DNA are bound together by complementary Watson-Crick base pairing. However, DNA can fold into many non-canonical secondary structures. In transcription and replication, the double helix must be transiently separated for enzymes to bind, creating single-stranded DNA. Single stranded DNA is also present in the telomeric overhangs of chromosomes. In these contexts, single stranded DNA can fold into G-quadruplex (GQ) and i-motif (iM) structures, which are the focus of this work.

### GQs are non-canonical DNA structures.

The GQ is formed by repetitive guanine-rich sequences. Four guanines can bond to each other to form a G-tetrad (**Figure 1A**)<sup>1</sup>. Each tetrad is stabilized by hydrogen bonds between the hoogsteen faces of the guanines, as well as the presence of a monovalent cation such as K<sup>+</sup> or Na<sup>+</sup>, which helps to neutralize the electrostatic repulsion from the electronegative C6 oxygens. Between two and four G-tetrads  $\pi$ – $\pi$  stack with a slight right-handed offset to form a helical GQ.

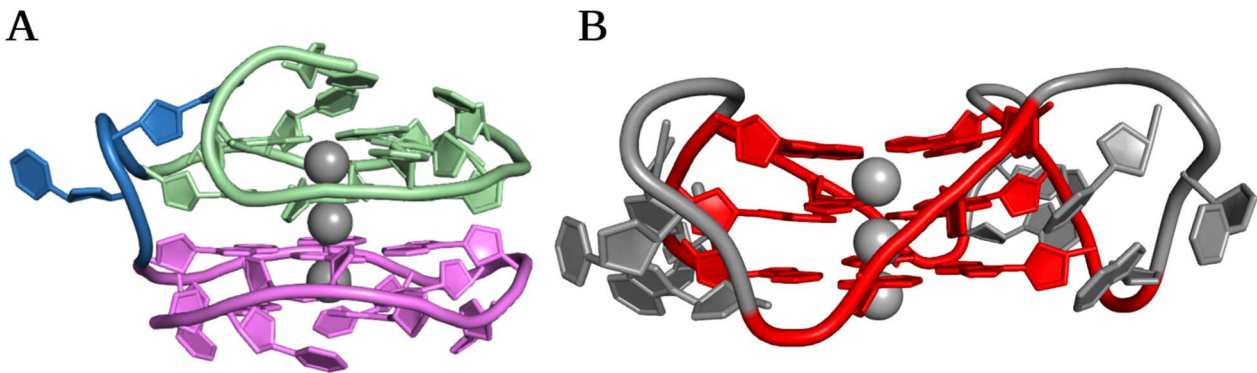


**Figure 1. G-quadruplex and i-motif structures.** Each structure is displayed alongside its respective constituent: G-tetrad for G-quadruplex, and C:C<sup>+</sup> base pair for i-motif. Figure adapted from <sup>2</sup>.

Similar to the right- and left-handed helices of B- and Z-DNA<sup>3</sup>, the backbone progression of GQs can take right- and left-handed forms<sup>4</sup>. While the right-handed GQ (RHGQ) has been extensively studied, the first left-handed GQ (LHGQ) was discovered only in 2015<sup>4</sup>. Like Z-DNA, LHGQ formation is generally less favorable than RHGQ formation, but there are two known minimal motifs for LHGQ formation: *Motif1*, with the sequence G(TGG)<sub>3</sub>TG and *Motif2*, with the sequence (GGT)<sub>3</sub>(TG)<sub>2</sub><sup>5,6</sup>. These minimal motifs alone form LHGQ DNA, and can drive adjacent sequences to fold left-handed as well<sup>5,6</sup>. This left-handed folding is sequence sensitive; single base variations in the sequence of LHGQ DNA can abrogate LH folding<sup>5-7</sup>.

All observed LHGQ structures consist of two sets of two G-tetrads stacked on top of each other to make a bilayered GQ structure (**Figure 2**)<sup>4-7</sup>. This differs from RHGQ structures, which have a great deal of structural diversity. The binding of ligands can also affect LHGQ formation;

an LHGQ-forming sequence has been found to fold into a RHGQ in the presence of a RHGQ-stabilizing ligand<sup>8</sup>.



**Figure 2. LHGQ structure.** (A). A typical bilayered LHGQ structure (PDB ID: 4U5M<sup>4</sup>). One subunit is colored green, and the other pink. The linker is blue. (B) A contrasting RHGQ structure which is not divided into two layers (PDB ID: 4FXM<sup>9</sup>). The G-tetrads are in red and the loops are grey.

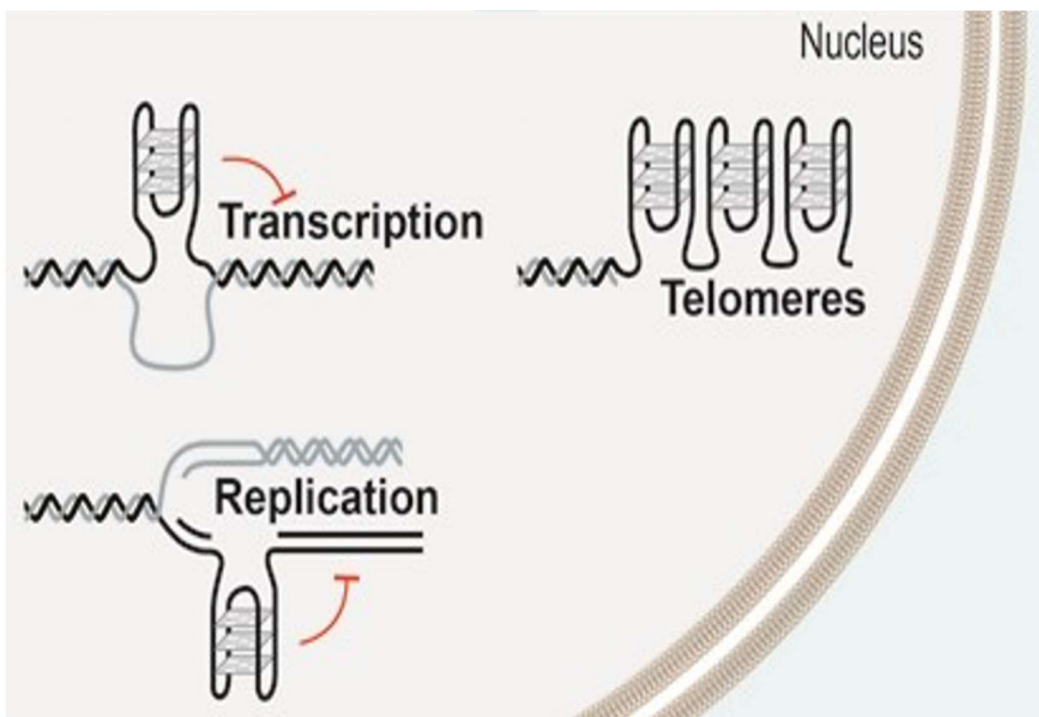
## iMs are non-canonical DNA structures

G-rich regions of DNA which form GQs have complementary C-rich regions which can fold into another type of non-canonical secondary structure known as the i-Motif (iM). In the iM, hemiprotonated C<sup>+</sup>:C base pairs intercalate to form a four-stranded stack (**Figure 1B**). These C<sup>+</sup>:C base pairs are stabilized by acidic conditions, so it was believed that iM DNA could not form under physiological conditions and was therefore biologically irrelevant. This may explain why iMs have been researched less extensively than GQs – there are only 1,450 iM related publications in the web of science core collection, compared to almost 10,000 GQ-related

publications. However, iM DNA can be stabilized by other means including molecular crowding<sup>10</sup> and negative superhelicity<sup>11</sup>, and the presence of iM DNA has recently been demonstrated in the nuclei of human cells<sup>2</sup>.

## Biological relevance of GQ and iM DNA

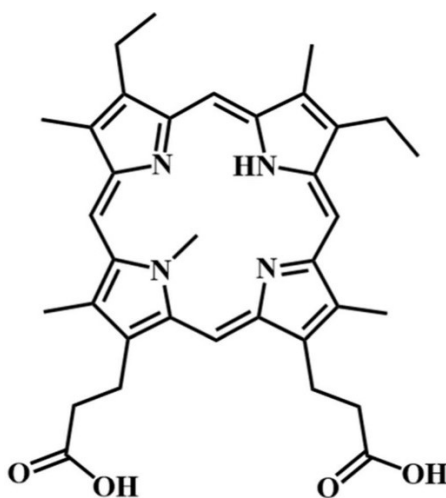
Transcription, replication, and telomere extension all involve single stranded DNA which can fold into non-canonical secondary structures (**Figure 3**). The complementary nature of G-rich and C-rich sequences means that GQs and iMs can accompany each other in the genome, namely in regulatory regions of the genome, including promoters and telomeres<sup>2,12</sup>. Cancer cells divide rapidly and typically employ high levels telomerase to support this division<sup>13–15</sup>, and GQ-stabilizing compounds may be useful for inhibiting telomerase for anticancer purposes<sup>16,17</sup>. Furthermore, non-canonical DNA structures play a role in the regulation of the genome. Both GQ and iM DNA have been shown to positively or negatively regulate transcription<sup>18,19</sup> in a context-dependent manner, which represents another motivation for investigation: manipulation of GQ and iM structures could be used to regulate gene expression.



**Figure 3. Sites of potential non-canonical DNA formation.** Single stranded DNA (black) is present in various biological contexts such as in transcription, replication, or telomeres. This single stranded DNA can form non-canonical DNA structures such as G-quadruplexes (pictured) and i-motifs. Figure adapted from <sup>20</sup>.

While the biological relevance of right-handed GQ DNA has been established, the relevance of LHGQ DNA remains unclear. Sequences have been shown to form LHGQ DNA *in vitro*, but whether those sequences retain left-handed folding *in vivo* is unclear. If LHGQ DNA is present *in vivo*, it may play regulatory roles similarly to RHGQ DNA. If LHGQ DNA is *not* present *in vivo*, it could be relevant for precisely that reason – induced LHGQ DNA could be used as a therapeutic tool which could be clearly delineated from natural RHGQ DNA, in addition to possible applications in DNA structure engineering.

GQ/iM DNA are promising therapeutic targets. A ligand that can selectively stabilize or destabilize a non-canonical DNA structure can be used to have a desired effect on a gene. Understanding the interactions between non-canonical DNA and small molecule ligands is essential to inform drug design. Towards this purpose, our lab works with N-methyl mesoporphyrin (NMM). NMM is a water-soluble porphyrin which is similar in structure to heme but has a distinctive central methyl group instead of an iron molecule (**Figure 4**). The methyl group enables NMM to selectively bind to and stabilize GQ DNA – it protrudes out of the plane and fits neatly into the GQ central potassium channel<sup>9,21,22</sup>. NMM does not stabilize duplex DNA, RNA, or other non-canonical structures such as iM DNA, triplex DNA and Z-DNA, and our lab has shown that NMM confers impressive increases in stability and favors the formation of right-handed GQ structures<sup>9,21–24</sup>. Investigating the interactions between NMM and GQ structures helps further our understanding of GQ-ligand interactions.



**Figure 4. Structure of N-methyl mesoporphyrin IX (NMM).** NMM is a highly selective GQ-targeting ligand.

## Scope of work

In this work, I characterize non-canonical DNA forming sequences using biophysical and structural methods. A large part of the project is focused on LHGQ DNA, which is relatively new. To learn more about the potential biological relevance of LHGQ DNA, I investigated the effect of sequence on LHGQ folding to help determine if any of the LHGQ-forming sequences present in the human genome can fold into LHGQs *in vivo*. I also investigated the effects of GQ ligand NMM on LHGQ folding.

I worked with two LHGQ-forming sequences, both of which originated from promoter regions of human genes associated with disease states. The first sequence is from the promoter region of the human solute carrier family 2 member 1 gene, hereby referred to as SLC. The solute carrier protein is a glucose transporter implicated in fatty liver disease, and this sequence is found to be upregulated in liver cancer cells<sup>25–27</sup>. I investigated the sequence dependence of the left-handed folding of SLC. I also solved the structure of SLC in complex with NMM, which forms a right-left hybrid G-quadruplex.

The second sequence I studied is from the *Homo sapiens* nuclear receptor binding SET domain protein 1 gene (NSD), which encodes a methyltransferase. Abnormal function of the NSD gene has been associated with tumorigenesis<sup>28–30</sup>. I performed further investigations on the effect of sequence on left-handed folding, and also investigated the effect of NMM on left-handed folding. I also report a preliminary structure solution of a novel LHGQ, NSD-TT.

Our lab works to identify iM-forming DNA sequences from the human genome and subsequently solve their structures. I worked with an iM-forming sequence from the *Homo*

*sapiens* platelet derived growth factor (PDGF) gene, which encodes an important mitogen which affects a large number of cell types<sup>31</sup>. I determined the iM folding of the PDGF sequence to be robust to sequence changes and attempted crystallization of the constructs that I created.

We approach these questions from two main angles. After designing short nucleotide sequences to study, we characterize their folding behavior using a variety of biophysical techniques. Informed by biophysical results, we choose certain sequences to study with X-ray crystallography towards resolving atomic-level details. My work will help to inform our understanding of the relationship between non-canonical DNA structure, function, and biological purpose.



## Ch.2: Materials and methods

### DNA, ligand, and buffers

Lyophilized oligonucleotides were purchased from Integrated DNA Technologies (IDT; Coralville, IA) with standard desalting purification. DNA was hydrated in doubly-distilled water to 1-2 mM and stored at -20°C. Extinction coefficients for all sequences were obtained using IDT's OligoAnalyzer 3.1 and DNA concentration was determined from UV-vis spectra collected at 90 °C. The full list of DNA sequences used in this work, along with their extinction coefficients and molecular weights, can be found in **Table A1-3**. Extinction coefficients were calculated using the OligoAnalyzer tool by Integrated DNA Technologies.

NMM stock was prepared in doubly-distilled water and its concentration was determined via UV-vis scans using an extinction coefficient of  $1.45 \times 10^5 \text{ M}^{-1}\text{cm}^{-1}$  at 379 nm<sup>23</sup>. GQ experiments were performed in **100K** buffer consisting of 10 mM lithium cacodylate pH 7.2 and 100 mM Potassium Chloride (KCl). iM experiments were performed in 10 mM pH 6.0 MES buffer. Prior to experimental work DNA was diluted into a desired buffer at appropriate concentration. To induce GQ/iM formation, DNA was heated at 90°C for 5 minutes, cooled slowly to room temperature over 4 hours, and equilibrated at 4 °C overnight.

### UV-vis spectroscopy

An Agilent Cary 3500 UV-vis spectrophotometer equipped with a Peltier block probe temperature controller was used to collect all UV-vis data ( $\pm 0.5$  °C error) from 220-330 nm for

DNA at 1.0 nm intervals with 0.02 s averaging time, 2 nm spectral bandwidth, and automatic baseline correction.

## Thermal difference spectra (TDS)

DNA was prepared at 4  $\mu\text{M}$  in buffer for all TDS studies. TDS was performed according to the procedure described in Mergny et. al<sup>32</sup>. TDS data was obtained by subtracting UV-vis scans taken at 20 °C after 5 minutes of equilibration from scans taken at 90 °C after 10 minutes of equilibration. In principle, the low and high temperature limits are defined by the temperatures at which the DNA is (mostly) folded and unfolded. The resulting signal is characteristic of the secondary DNA structure.

## UV-vis melting studies

UV-vis melting experiments were conducted from 20-95 °C with 1 °C step, 1.0 °C/min temperature rate for GQ and 0.5 °C/min temperature rate for iM, 2.00 s averaging time, and 2.00 nm spectral bandwidth. DNA was prepared at 4  $\mu\text{M}$ . Absorbance was monitored at fixed wavelengths of 260, 295, and 335 nm. No absorbance was expected from the DNA sample at 335 nm, therefore any signal at 335 nm was due to instrumental fluctuation. The 260 and 295 nm wavelengths were corrected for the absorbance at 335 nm. Reported  $T_m$  are from the inflection points of the melt curve, determined by visually inspecting the first derivative graph ( $\pm 0.5$  °C error). Data from reversible melts were fit with a two-state model to find the enthalpy of melting<sup>33</sup>.

## Circular dichroism (CD) scans

Circular dichroism (CD) provides information on chiral structures by measuring the difference in their interactions with right- and left-handed plane polarized light. Achiral samples generate no CD signal. In addition to the chirality of the DNA bases, the GQ secondary structure is itself chiral and therefore can be detected by CD. CD experiments were performed on an Aviv-435 or a Jasco J-1500 circular dichroism spectrophotometer equipped with a Peltier thermal controller ( $\pm 0.3$  °C error Aviv;  $\pm 0.5$  °C error Jasco) in 1 cm quartz cuvettes with Teflon caps at 20 °C. CD scans were collected from 330-220 nm with 1 s averaging time, 200 nm/min speed, 2 nm bandwidth, and 1.00 nm step. Five scans were collected for each sample in the corresponding cuvettes. DNA samples for CD studies were prepared as described above at 4  $\mu$ M. All data processing was done with Origin 9.1 software unless otherwise noted.

## CD melting studies

CD melting experiments were conducted either from 4°C or 20°C to 95 °C and in reverse with 1 °C step, 1.0 °C/min temperature rate, 8 s data integration time, and 2.00 nm spectral bandwidth. Absorbance was monitored at fixed wavelengths corresponding to the location of the relevant CD peaks; always 330 nm, then 280 nm for LHGQ and 262 nm for RHGQ. DNA was prepared at 4  $\mu$ M. Baseline measurements of each cuvette were taken and used to correct for non-DNA CD signal for every sample. The difference between  $T_m$  from the melting and cooling curves was used to determine hysteresis values.  $T_m$  values were obtained by taking the first derivative of each corrected signal and visually locating the center of the peak corresponding to the inflection point ( $\pm 0.5$  °C error). These values signify the point at which half the DNA is unfolded and are

reproducible only under the same experimental conditions. Data from reversible melts were fit with a two-state model as with UV-vis melting studies. The reported data are the average of 1-3 trials for each sequence. Representative raw data are shown in Appendix, **Figure A1**.

## Native polyacrylamide gel electrophoresis (PAGE)

PAGE samples consisted of 10  $\mu$ L of 50  $\mu$ M DNA in a buffer and were weighted down with 7% w/v sucrose prior to loading. Fifteen percent native polyacrylamide gels were made with 5 mM KCl and 1 $\times$ Tris-Borate-EDTA (TBE) for GQ gels, and 10 mM KCl and 50 mM MES pH 6.0 for iM gels. Gels were pre-migrated at 150 V for 30 min, loaded with 10  $\mu$ L of each sample, and allowed to run for 150 min at 150 V at room temperature. A tracking dye was used to monitor gel progress and an oligothymidylate ladder consisting of dT<sub>15</sub>, dT<sub>24</sub>, dT<sub>30</sub>, and dT<sub>60</sub> was used as a length marker. DNA bands were visualized using Stains-All and the resulting gel was captured using a smartphone camera.

## Crystallization of SLC+NMM

Crystallization samples were prepared by annealing 0.5mM HPLC-purified SLC NMM with a 1:1 molar ratio of NMM in 10mM lithium cacodylate 100mM KCl pH 7.2. For the crystallization of SLC+NMM, I used the commercial crystal screens Helix (Molecular Dimensions) and NATRIX (Hampton Research) and got initial crystallization hits with NATRIX conditions 2-19 and 2-32. Both conditions had 0.08 M NaCl, 0.04 M sodium cacodylate (NaCaCo), 12 mM spermine, and a pH of 6.0, differing only in their MPD content, at 45% and 30% respectively. The crystals produced by these conditions were inconsistent, so with extensive

optimization I was able to identify 0.08M NaCl, 0.04M NaCaCo, 2mM Spermine, 45% MPD, pH 6.0 as the best condition.

In addition to optimizing the crystallization condition, I discovered that HPLC-purified DNA at 0.5mM concentration produced the most reproducible crystals. Notably, the HPLC purification proved to be essential – SLC+NMM crystals prepared with identical conditions, but unpurified DNA crystallized less consistently and diffracted relatively poorly. I prepared crystallization drops by hand using 1  $\mu$ L well condition and 1  $\mu$ L sample, suspended over 400  $\mu$ L crystallization condition in the reservoir. The resulting crystals grew quickly – crystals were often observed to grow overnight. They were significantly larger than the initial hits and somewhat amorphous in shape but had ordered diffraction. We collected six data sets from three different crystals, ranging from 1.45-1.5 Å resolution.

Data on the attempts at crystallography for other sequences is present in the appendix, **Tables A6-A9.**

## Chapter 3: Biophysical and structural characterization of the SLC G-quadruplex in complex with NMM

Our lab performed a BLAST search of the human genome for the G(TGG)<sub>3</sub>TG LHGQ *Motif1* and (GGT)<sub>3</sub>(TG)<sub>2</sub> LHGQ *Motif2* and found at least ten different matches across eight chromosomes. We chose three sequences for further characterization, one of which was SLC. We wanted to investigate the factors influencing left-handed G-quadruplex (LHGQ) folding, the effect of small molecule ligand, particularly NMM on LHGQ folding, as well as help determine if LHGQ DNA could be found *in vivo*. Our lab approaches these questions from two main angles. After designing short nucleotide sequences to study, we characterize their folding behavior using a variety of biophysical techniques. Informed by biophysical results, we choose certain sequences to study with X-ray crystallography towards resolving atomic-level details.

The base SLC sequence consists of LHGQ *Motif1* connected to an RHGQ-forming motif by an AT linker. To approach biological setting, we designed native constructs by extending the sequence in 5' and 3' directions using nucleotides from genetic setting. The constructs are named according to their extensions; the construct with an additional base on the 3'/Right hand side is named SLC-1R, and the construct with an additional base on the 5'/Left hand side is named SLC-1L (**Table 1**). Creating constructs also provides more options for successful crystallization.

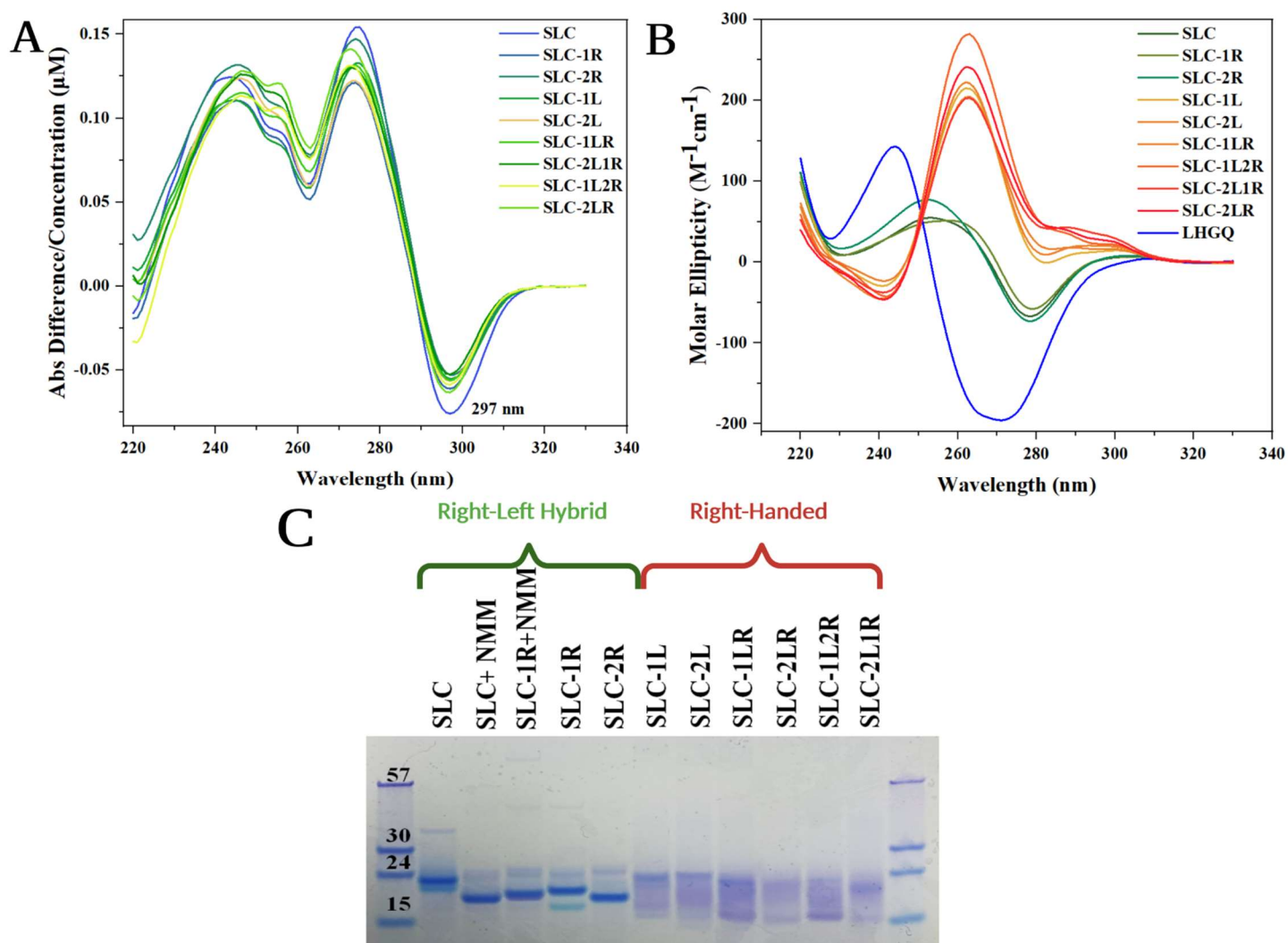
**Table 1.** SLC construct sequences. LHGQ *Motif1* is colored blue, the AT linker is black, and the RHGQ motif is colored red. 5'- and 3'-overnags are green. Note conformation was determined based on CD signature and PAGE (**Figure 5**).

Name	Sequence	Conformation
SLC	GTGGTGGTGGT <del>AT</del> GGTGGTGGTGG	R/L Hybrid
SLC-1R	GTGGTGGTGGT <del>AT</del> GGTGGTGGTGGC	R/L Hybrid
SLC-2R	GTGGTGGTGGT <del>AT</del> GGTGGTGGTGGCA	R/L Hybrid
SLC-1L	GGTGGTGGTGGT <del>AT</del> GGTGGTGGTGG	RH
SLC-2L	GGGTGGTGGTGGT <del>AT</del> GGTGGTGGTGG	RH
SLC-1LR	GGTGGTGGTGGT <del>AT</del> GGTGGTGGTGGC	RH
SLC-2LR	GGGTGGTGGTGGT <del>AT</del> GGTGGTGGTGGCA	RH
SLC-2L1R	GGGTGGTGGTGGT <del>AT</del> GGTGGTGGTGGC	RH
SLC-1L2R	GGTGGTGGTGGT <del>AT</del> GGTGGTGGTGGCA	RH

## Biophysical characterization of SLC constructs

We used various biophysical techniques to characterize the effect of sequence changes on LHGQ folding, as well as identify suitable constructs to proceed with structural determination through crystallization. To determine if the constructs form GQs, we use thermal difference spectra (TDS). TDS is generated by recording the ultraviolet absorbance spectra of the folded DNA at low temperature, then increasing temperature to melt the sequences and recording a second spectra of the unfolded constructs<sup>32</sup>. Taking the difference of these two spectra generates a TDS, which contains only the contributions from the DNA secondary structure. The TDS of all

nine constructs have peaks and troughs at 273 and 297 nm, respectively, indicating GQ folding<sup>32</sup> (Figure 5A).



**Figure 5. Biophysical characterization of SLC constructs.** (A) TDS spectra. All nine constructs have a trough at 297 nm indicative of GQ folding. (B) CD spectra. Right-left hybrid GQs are in shades of green, right-handed GQs are in shades of red. A previously characterized LHGQ (PDB ID: 4U5M<sup>4</sup>) is included as a reference (navy blue). Samples with NMM have one molar equivalent of NMM. (C) 15% Native PAGE gel of constructs. DNA was annealed in 100 mM potassium chloride, 10 mM lithium cacodylate pH 7.2 (100K buffer).

Next, we use circular dichroism (CD), which provides information on chiral DNA secondary structures. A CD peak at 260 nm is indicative of parallel RHGQ DNA, whereas



LHGQ DNA has a nearly mirrored trough of similar intensity at 271 nm<sup>4-6,9,21,34</sup>. The CD signals are corrected for the actual concentration of each sample, measured by UV-Vis, yielding molar ellipticity. The intensity of CD peaks and troughs can be used as a qualitative measure for how well folded a sample is.

The SLC constructs fell into two groups of CD behavior (**Figure 5B**). The first group of SLC, SLC-1R, and SLC-2R all had CD spectra with peaks and troughs characteristic of both RHGQ and LHGQ. The second group, consisting of all -L constructs, had typical RHGQ peaks. The first group leaves the 5' LHGQ *Motif1* unmodified, whereas the second group has one or two 5' guanine bases extending *Motif1*. Appending a G to LHGQ *Motif1* changes it to fit the generic RHGQ formula (GGT)<sub>4</sub><sup>5</sup>, detected here via CD (**Figure 5B**).

There is one published right-left hybrid GQ structure<sup>7</sup>. The CD spectra of this structure closely resembles the CD spectra of the first group of SLC constructs. However, the first group of SLC constructs cannot be presumed to form right-left hybrid GQs from CD alone. A heterogeneous solution containing equal parts of RHGQ and LHGQ DNA would produce a similar CD signature. Native polyacrylamide gel electrophoresis (PAGE) must be used to determine the homogeneity of the solution.

DNA structures migrate through the native gel in their folded state. Different conformations of the same DNA construct are shaped differently and therefore can be separated by native PAGE. Thus PAGE allows determination both the conformational and oligomeric heterogeneity of DNA constructs. The PAGE (**Figure 5C**) revealed a striking pattern; the first group maintained dominant single bands, indicating homogenous samples. In contrast, the second group displayed smeared bands likely indicating multiple species with similar

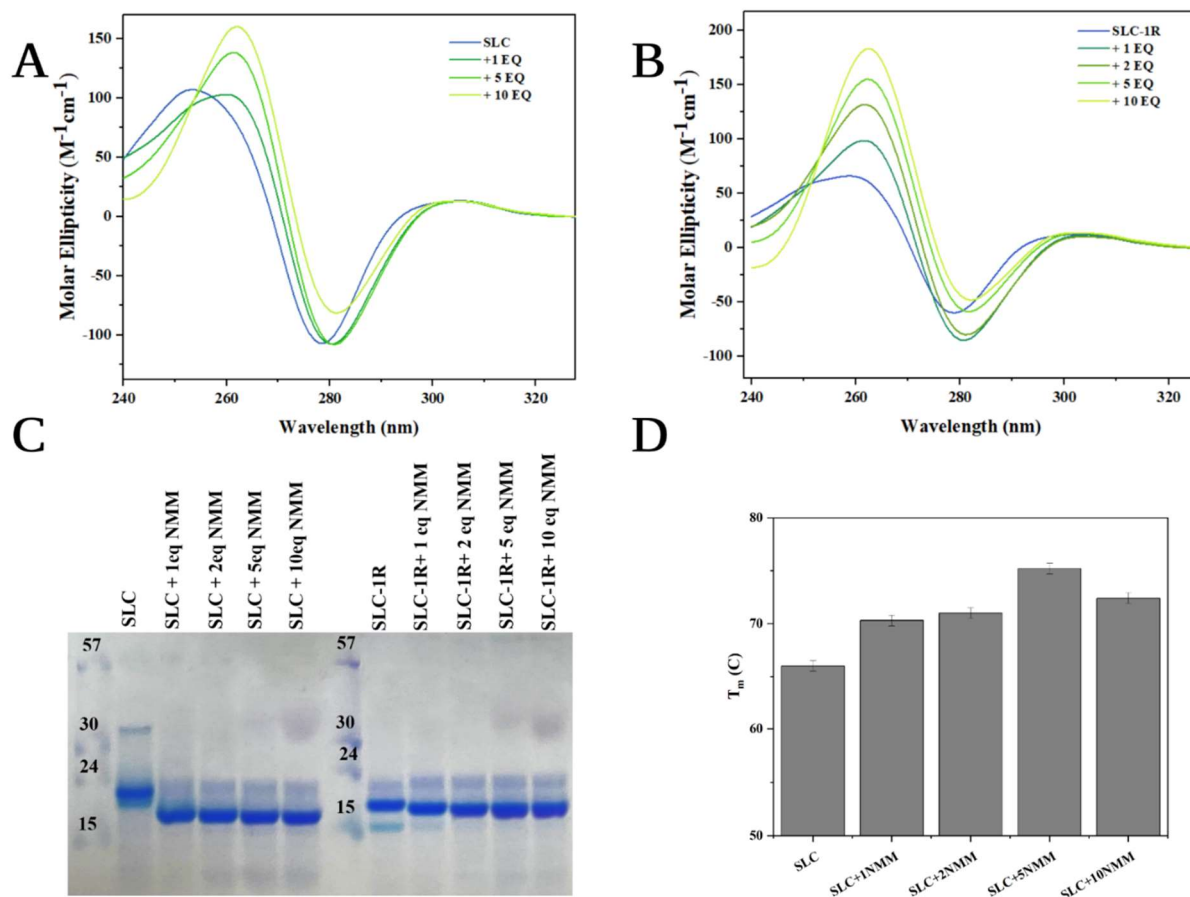
conformations. The second group of constructs have more than the 16 Gs required to form exactly four G-tetrads. Additional Gs create flexibility in which Gs are included in which tetrad, leading to the observed sliding conformations. Combined with the CD spectra, the PAGE data enable us to conclude that the first group of constructs all form right-left hybrid GQ structures in solution, whereas the second group of constructs form a mixture of different RHGQ structures in solution.

## Effect of NMM on right-left hybrid folding

NMM has been shown to bind to and stabilize right-handed parallel GQs, and the stabilization conferred can shift the equilibrium from hybrid or antiparallel GQs towards folding parallel<sup>9,22–24</sup>. The interactions between NMM and LHGQ DNA have not been investigated. It is not known if NMM will stabilize LHGQ DNA, coerce LHGQ DNA to fold right-handed, or fail to interact entirely. I therefore proceeded to investigate the effects of NMM on the folding, stability, and oligomeric state of SLC and SLC-1R. Group 1 constructs were chosen partly because of their interesting CD signatures and partly because PAGE indicates that they have single dominant conformations in solution, which simplifies analysis and bodes well for chances of crystallization. The choice of SLC and SLC-1R rather than SLC-2R was made arbitrarily.

I performed a batch titration by annealing the sequences with 1-10 molar equivalents of NMM. TDS showed that GQ folding was not disrupted by NMM (Appendix, Figure A1). With increasing concentrations of NMM, the RHGQ peak on the CD spectra of SLC became increasingly intense without the loss of the LHGQ trough, though some shifting of the trough was observed (**Figure 6A**). A similar pattern was observed with SLC-1R (**Figure 6B**). On PAGE, SLC displayed a complete shift with the first equivalent of NMM and retained a single

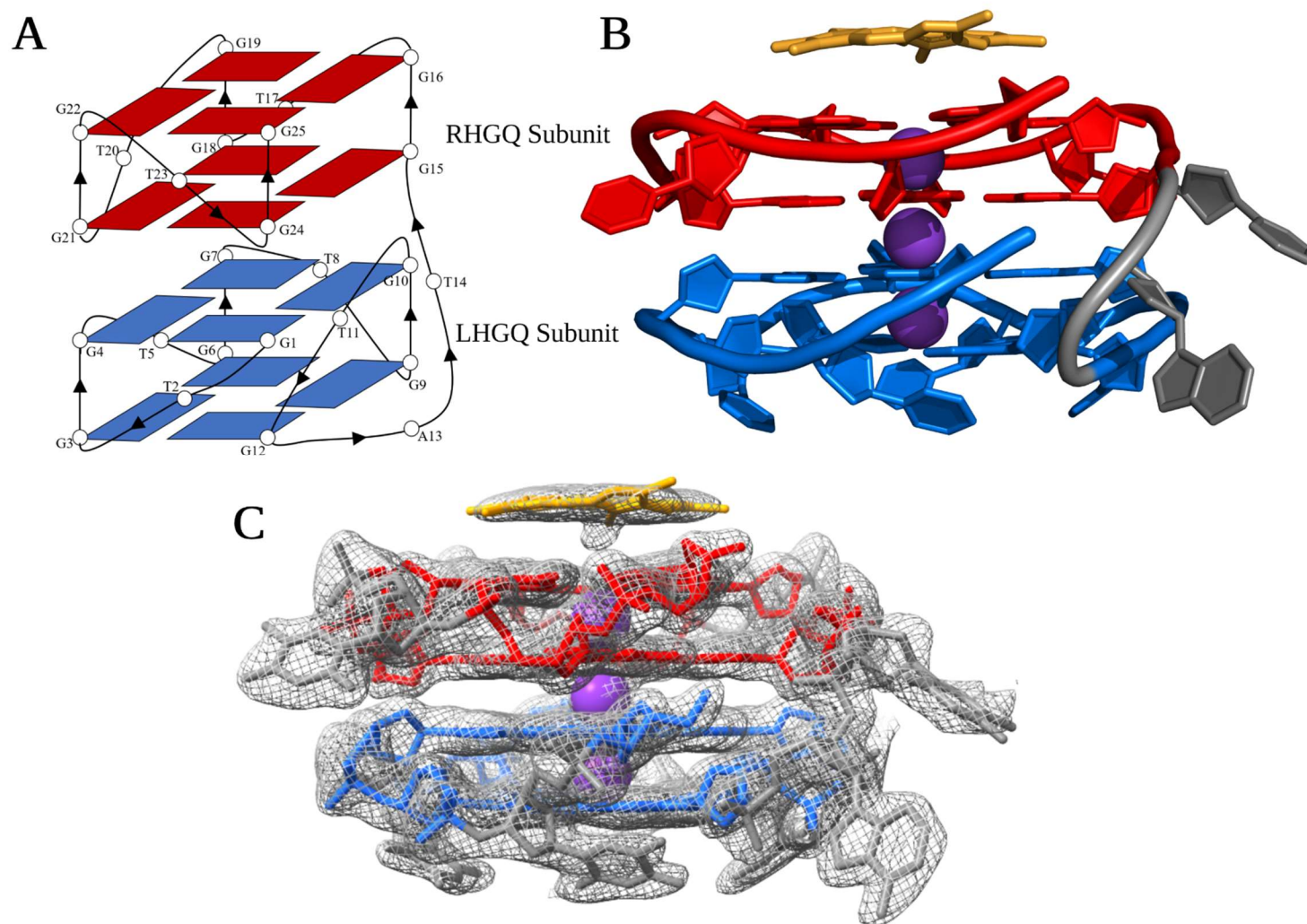
band. SLC-1R also retained its single band, but saw the band continue to shift downstream and broaden increasingly up to five equivalents of NMM (**Figure 6C**). We performed a CD melting experiment to determine the stabilization of SLC by NMM (**Figure 6D, Table A3**). In CD melting, we track the intensity of the characteristic CD peak as the DNA is melted by heating to 95°C, and subsequently cooled to 4°C. This allows us to determine the thermodynamic stability of DNA expressed as  $T_m$ . The base SLC sequence has a  $T_m$  of  $66.0 \pm 0.5^\circ\text{C}$ , and which was increased by 9°C to  $75.0 \pm 0.5^\circ\text{C}$  with five equivalents of NMM. This is a modest stabilization compared to other GQ-NMM complexes, which can exceed twenty degrees<sup>24</sup>. I expect the relative decrease in stabilization from five to ten equivalents to be attributable to experimental error; additional trials are needed. The CD and PAGE indicate that SLC and SLC-1R maintain right-left hybrid structures with the addition of NMM. This conclusion indicates that NMM likely binds to and stabilizes the right-handed subunit and not the left-handed subunit.



**Figure 6. NMM Titration of SLC and SLC-1R.** CD spectra of (A) SLC annealed with 1, 5 and 10 equivalents of NMM, (B) SLC-1R annealed with 1, 2, 5, and 10 equivalents of NMM. Both experiments were conducted in 100K buffer. (C) PAGE gel of SLC and SLC-1R with increasing amounts of NMM. Gel was run by Eric Xing. (D) Stabilization of SLC by NMM. CD melting study was conducted by Eric Xing.

## SLC+NMM complex contains a right-left hybrid G-quadruplex

I proceeded with crystallization of SLC and SLC-1R with and without NMM because these sequences produced homogenous samples. I was successful in crystallizing both constructs in complex with NMM and produced diffraction-quality crystals of the SLC+NMM complex whose structure I subsequently solved via molecular replacement.



**Figure 7. Crystal structure of SLC+NMM complex.** (A) Schematic representation of the structure. All bases are numbered, and the arrowed lines indicate backbone progression. (B)

Cartoon representation of the SLC+NMM biological unit. Bases are shown as filled rings, with the LHGQ subunit in blue, the RHGQ subunit in red, and the linker in grey. Potassium ions are purple spheres. NMM is orange. (C) SLC+NMM with electron density shown at 1.0 I/ $\sigma$ .

The SLC+NMM structure was solved in the  $P 4_1 2_1 2$  space group to a resolution of 1.45 Å. The asymmetric unit contains a single DNA chain, which is folded into a right-left hybrid GQ in agreement with our expectations. All previously published LHGQ structures are four layered structures consisting of a pair of two-tetrad subunits<sup>4-7</sup>. SLC+NMM fits this pattern, with a two-tetrad LHGQ subunit connected to a two-tetrad RHGQ subunit by an AT linker (**Figure 7A**). The left-handed subunit is formed by the G(TGG)<sub>3</sub>TG LHGQ *Motif1*, and the right-handed subunit is formed by the remaining (TGG)<sub>4</sub>.

The two subunits are lined up such that a shared column of potassium ions runs through the center of both GQs (**Figure 7B**). NMM caps the 3' end of the right-handed subunit with its out-of-plane methyl group pointing into the center of the GQ, as previously observed by our lab<sup>9,21,22</sup>.

## Conclusion

In this chapter, I presented biophysical and structural characterization of the SLC DNA, which we found to form the second reported left-right hybrid GQ structure. NMM bound to and stabilized the right-handed subunit without disrupting left-handed folding. We also determined that additions of guanines to the half of the SLC sequence containing the LHGQ *Motif1* disrupted left-handed folding, but additions to the RHGQ forming section were tolerated.

There remains a great deal of work to do for this project. Our lab replicates biophysical results at least three times, using at least two separate sources of DNA. Several of the

experiments presented in this section have been performed only once, see **Table A10**.

Furthermore, the CD melting stability data for this section has not been fully processed. The refinement of the SLC+NMM crystal structure is nearly complete – it has yet to be deposited into the PDB. Once the structure is finalized, final analysis of structural characteristics must be performed.

## Chapter 4: Handedness of *Homo sapiens* nuclear receptor binding SET domain (NSD) G-quadruplexes.

Using a BLAST search, our lab found the NSD sequence consisting of both the G(TGG)<sub>3</sub>TG LHGQ *Motif1* and (GGT)<sub>3</sub>(TG)<sub>2</sub> LHGQ *Motif2*, connected by a GT linker. The addition of the G from this linker means that *Motif1* fits the general RHGQ formula<sup>35</sup>. Because this would lead to right-handed folding, we designed the NSD-TT construct by substituting the GT linker for TT so the sequence would match LHGQ *Motif1*<sup>5</sup>. This substitution in the gene leads to a single point mutation. Further TT constructs were designed by expanding the TT sequence according to genetic context (**Table 2**). Overall, we designed nine native constructs and three TT mutants, for a total of twelve different constructs.

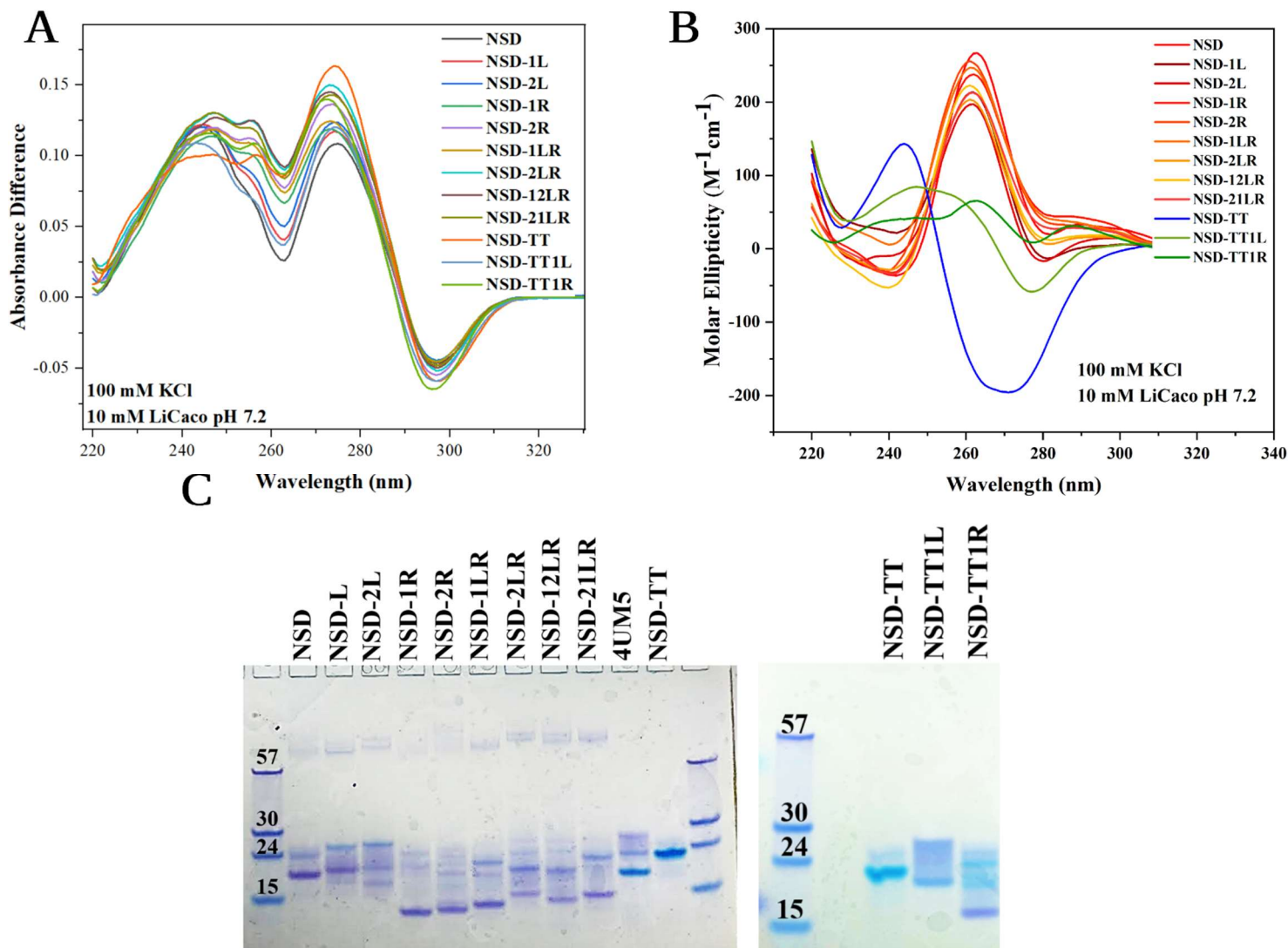


**Table 2.** NSD construct sequences and CD signatures. In the sequences, LHGQ *Motif1* is colored dark blue, the two-nucleotide linker is black, and the LHGQ *Motif2* is colored teal. The sequence additions that differentiate the constructs are colored red. Note conformation was determined based on CD signature and PAGE (**Figure 8**).

[illegible]

## Biophysical Characterization of Constructs

We performed biophysical characterization of the constructs to determine how changes in sequence affect left-handed folding and to identify promising constructs to proceed with structural characterization. The TDS spectra of twelve constructs have a peak at 275 and a trough at 295 nm, characteristic of GQ folding<sup>32</sup> (**Figure 8A**). All nine native NSD constructs form RHGQ, and NSD-TT forms LHGQ. NSD-TT1L and NSD-TT1R, however, have intermediate CD signatures with features indicative of both RHGQ and LHGQ (**Figure 8B**).



**Figure 8. Biophysical characterization of NSD constructs.** (A) TDS and (B) CD scans. RHGQ sequences are in red, whereas LHGQ NSD-TT is blue. The intermediate NSD-TT1L and NSD-TT1R are in shades of green (C) Native PAGE gel of constructs. 4UM5 is a previously characterized LHGQ included as a reference.

We used PAGE to assess the homogeneity and oligomeric state of the constructs (Figure 8C). All nine NSD constructs are heterogeneous and show higher order species. In contrast,

NSD-TT is monomeric and homogenous. NSD-TT1L is mainly monomeric with some smearing, whereas NSD-TT1R has multiple species.

Despite containing two LHGQ minimal motifs, none of the native NSD construct CD spectra showed observable left-handed folding. Combined with the PAGE, all nine native NSD constructs can be concluded to form a heterogeneous mixture of RHGQs in solution. The likely culprit for the disruption of left-handed folding is the GT linker between the minimal motifs. The additional G from the linker changes the G(TGG)<sub>3</sub>TG *Motif1* to G(TGG)<sub>4</sub>. The (TGG)<sub>4</sub> section is expected to fold right-handed.

Without structural data, we cannot be certain of how this change to *Motif1* disrupts left-handed folding for *Motif2* as well. A possible explanation lies with the linker - if the additional G from the linker is participating in a G-tetrad, then the linker is effectively reduced to just T. All observed LHGQ structures are composed of two stacked subunits connected by a minimum of two bases<sup>4-7</sup>. The single base may be too short to allow this architecture to form, leading to the observed RHGQ signatures.

The NSD-TT constructs did show left-handed behavior. Interestingly, only NSD-TT was fully left-handed. Because NSD-TT also had a single band with minimal smearing on PAGE, it was determined to be a promising construct to proceed with crystallography.

NSD-TT1R and -TT1L are neither full RHGQ or LHGQ. The spectrum for NSD-TT1R had a weak RHGQ peak and minimal LHGQ features, and PAGE demonstrated that NSD-TT1R had multiple conformations in solution. The NSD-TT1R construct has an additional G on the 5' end – it is likely that the presence of this G disrupts left-handed folding, as we have previously observed with the native NSD and SLC-L constructs. NSD-TT1L has an additional T on the 3'

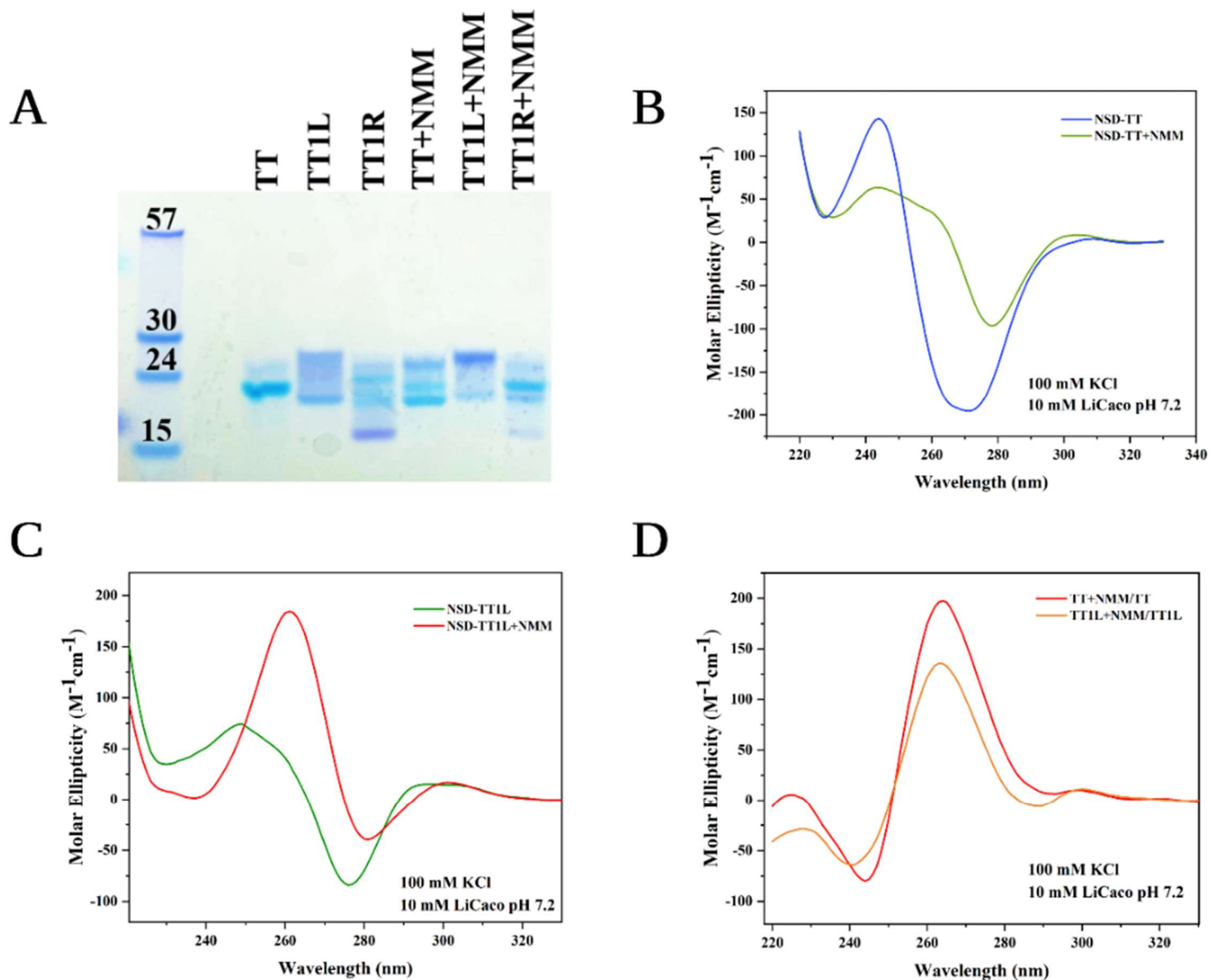
end and had a CD signature very similar to the right-left hybrid GQ SLC+NMM, with well-defined RHGQ and LHGQ peaks. However, the PAGE shows smearing so without structural data NSD-TT1L cannot be confirmed to fold into a right-left hybrid GQ.

Adding guanines to the LHGQ minimal motifs disrupts left-handed folding, as in the case of the GT linker or NSD-TT1R. This is expected because the guanines can participate in G-tetrads. Furthermore, left-handed folding is disrupted by the addition of thymine to LHGQ *Motif1* as in NSD-TT1L. The mechanism of this disruption is not immediately apparent and requires further investigation. Together, this implies that any putative LHGQ sequences in the genome have strict requirements of the surrounding sequence if they are to fold left-handed *in vivo*.

## NMM partially induces right-handed folding in NSD-TT constructs.

Because of their interesting CD spectra and promising PAGE results, we chose to proceed with investigating the effects of NMM on the three NSD-TT constructs. The TDS spectra confirmed maintained GQ secondary structure when the constructs were annealed with NMM (**Figure 8A**). Native PAGE of the three NSD-TT constructs annealed with and without NMM showed NMM-induced conformational shifts (**Figure 9A**). In the case of NSD-TT, the addition of NMM led to multiple bands representing different conformations. The addition of NMM to NSD-TT1L led to a reduction in the number of different conformations – where the original sample had two moderately intense bands with smearing between them, the NMM sample has a single strong band, with a faint smear and band following it. Notably, the strong and weak band from NSD-TT1L+NMM align with the two moderate bands of NSD-TT1L alone.

This could imply that NSD-TT1L is a sample with two conformations in equilibrium, and NMM shifts the equilibrium in favor of one of the two.



**Figure 9. Effect of NMM on NSD-TT construct folding.** (A) Native PAGE of NSD-TT constructs with and without two equivalents of NMM (B) CD spectra of NSD-TT and NSD-TT+NMM (C) CD spectra of NSD-TT1L and NSD-TT1L+NMM (D) CD difference spectra for

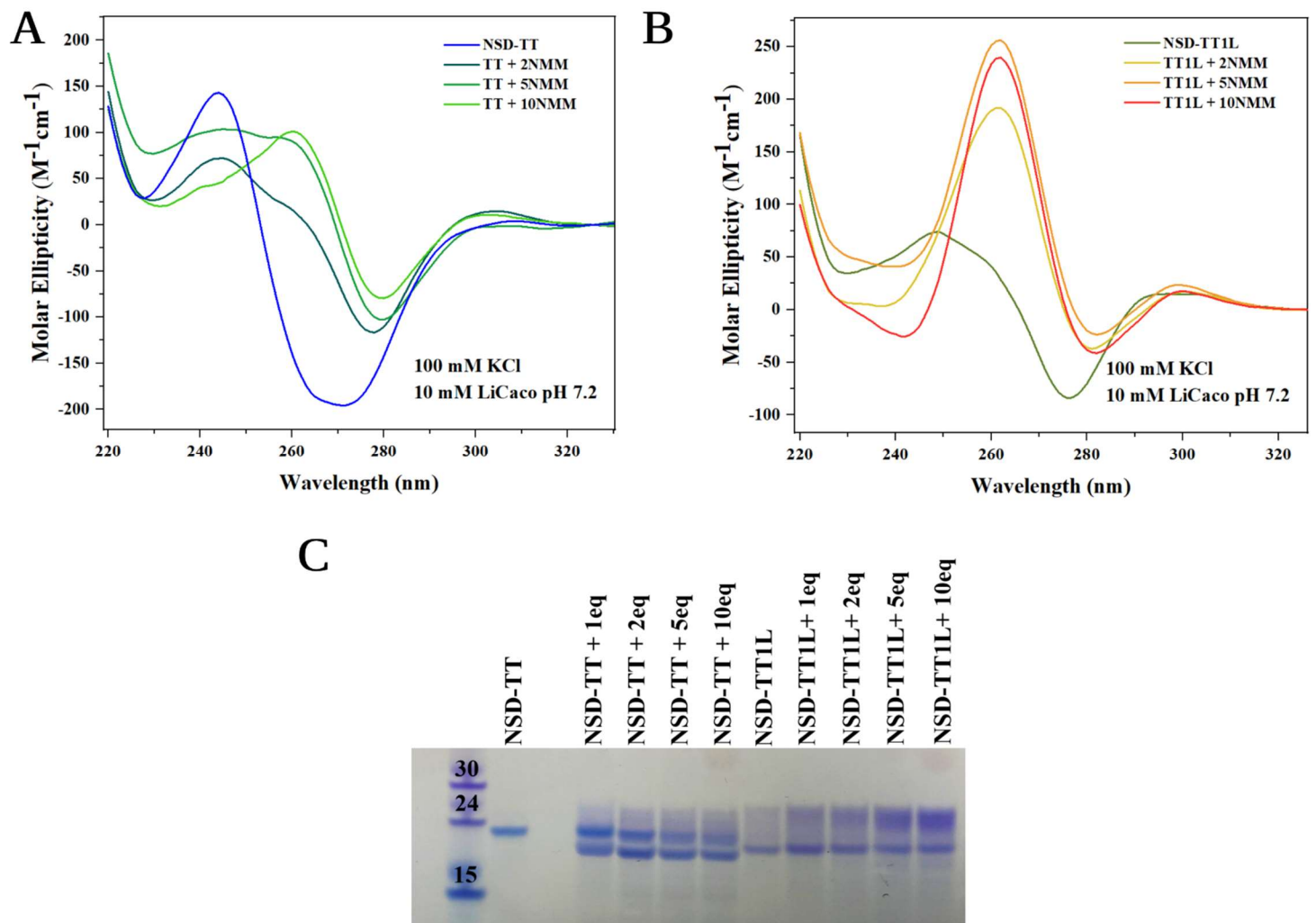
NSD-TT and -TT1L with NMM. In both cases, CD spectra of the NMM-free sample was subtracted from the spectra of the sample with NMM.

NSD-TT1R had a great deal of conformational heterogeneity with and without NMM. Because of this heterogeneity as well as the ill-defined CD spectra of NSD-TT1R, I chose to proceed with only NSD-TT and -TT1L for further time-intensive biophysical characterization. The CD spectra of NSD-TT and -TT1L with and without NMM confirm that NMM causes conformational shifts (**Figure 9B-C**). With the addition of NMM, the LHGQ signature of NSD-TT converts to a right-left hybrid and the right-left hybrid signature of NSD-TT1L converts to RHGQ.

Considering the conformational heterogeneity shown by PAGE, it is possible that the right-left hybrid signatures of NSD-TT+NMM and NSD-TT1L are the result of a mixture of discrete RHGQ and LHGQ rather than true right-left hybrids. In both cases, the addition of NMM led to an increase in RHGQ character, in agreement with the observed ability of NMM to stabilize and favor the formation of RHGQ DNA<sup>9,34</sup>. This is well illustrated by subtracting the CD spectra of the NMM-free samples from that of the samples containing NMM (**Figure 9D**). For both NSD-TT and -TT1L this CD difference spectra shows that the addition of NMM leads to increase in the peak at approximately 260 nm, indicative of the RH parallel GQ.

The CD results suggest that NMM is capable of coercing LHG4-forming sequences into right-handed conformations, reminiscent of the ability of NMM to induce parallel GQ formation<sup>23</sup>. However, full conversion was not observed – I proceeded with an NMM titration of NSD-TT and NSD-TT1L to determine if saturating concentrations of NMM can fully convert

LHGQ DNA to RHGQ DNA. Each construct was annealed with two, five, and ten equivalents of NMM. The shape and intensity of the main trough at 279 nm in the CD signature of NSD-TT varies little between five and ten equivalents of NMM, indication that 5 equivalents of NMM is sufficient to saturate NSD-TT (**Figure 10A**). NSD-TT1L similarly stabilized at a RHGQ signature with the addition of five equivalents of NMM (**Figure 10B**). The PAGE shifts stabilize at a similar amount of NMM (**Figure 10C**).



**Figure 10. NMM Titration of NSD-TT and NSD-TT1L.** CD spectra of (A) NSD-TT and (B) NSD-TT1L annealed with two, five, and ten equivalents of NMM. (C) Native PAGE of both constructs with increasing NMM.

Even with saturating amounts of NMM, NSD-TT remains a right-left hybrid. I hypothesize that this is because LHGQ *Motif1* cannot be coerced by NMM to fold right-handed. If this is correct, these results can be interpreted as NMM converting NSD-TT's *Motif2* to RHGQ while leaving NSD-TT's *Motif1* as LHGQ. In NSD-TT1L, *Motif1* has been disrupted by sequence modification. Presumably, this causes it to form RHGQ, leading to the observed right-left hybrid CD signature. With the addition of NMM, NSD-TT1L+NMM's *Motif2* is coerced to form RHGQ, leading to full RHGQ signature. This theory is consistent with SLC+NMM, where *Motif1* is confirmed to remain LHGQ even with NMM in excess. Further investigation, detailed in Chapter 6, is required to validate this hypothesis.

## Certain flanking sequences permit left-handed folding.

So far, we have observed that the extension of our LHGQ sequences according to genomic context has led to at least partial disruption of left-handed folding. LHGQ DNA is unlikely to be relevant *in vivo* if the presence of any flanking base causes LHGQ to become RHGQ. To determine if this is the case, we decided to design constructs to test all possible flanking bases (**Table 3**).



**Table 3.** NSD-TT constructs and their folding topology based on the CD signatures. Constructs are named to describe the identity and location of the additional flanking nucleotide – NSD-TTLT has a 5'/Left side T. Color scheme is consistent with **Table 2**.

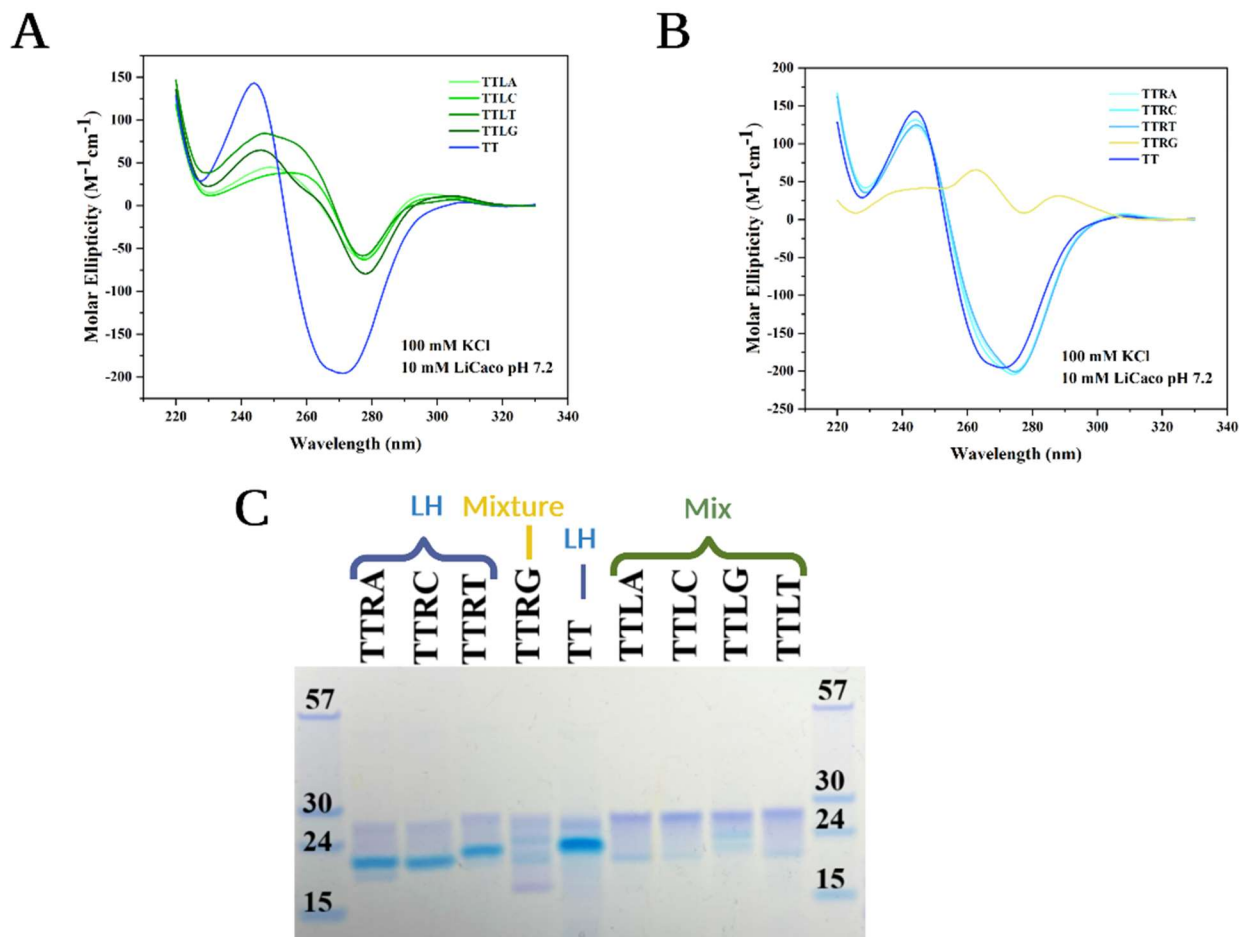
Name	Sequence	CD Signature
NSD-TT	GTGGTGGTGGTGTGTTGGTGGTGGTGTG	LH
NSD-TTLT	TGTGGTGGTGGTGTGTTGGTGGTGGTGTG	R/L Hybrid
NSD-TTLA	AGTGGTGGTGGTGTGTTGGTGGTGGTGTG	R/L Hybrid
NSD-TTLC	CGTGGTGGTGGTGTGTTGGTGGTGGTGTG	R/L Hybrid
NSD-TTLG	GGTGGTGGTGGTGTGTTGGTGGTGGTGTG	R/L Hybrid
NSD-TTRT	GTGGTGGTGGTGTGTTGGTGGTGGTGTGT	LH
NSD-TTRA	GTGGTGGTGGTGTGTTGGTGGTGGTGTGA	LH
NSD-TTRC	GTGGTGGTGGTGTGTTGGTGGTGGTGTGC	LH
NSD-TTRG	GTGGTGGTGGTGTGTTGGTGGTGGTGTGG	Inconclusive
<i>Motif1</i>	GTGGTGGTGGTG	LH
<i>Motif2</i>	GGTGGTGGTGTG	LH

Starting with the base NSD-TT sequence we created non-native constructs with the remaining four possible flanking bases on either side. For clarity, NSD-TT1L and -TT1R are hereby referred to as NSD-TTLT and -TTRG, respectively. The four constructs adding to the left side of the sequence, modify LHGQ *Motif1*, whereas the right-side constructs modify LHGQ *Motif2*.

TDS confirmed GQ folding in all constructs (**Figure A2**). The PAGE bands of NSD-TTLG, -TTLA, and -TTLC appear very similar to each other, with an upper and lower band bounding a smeared section (**Figure 11A**). NSD-TTLG has a similar upper band but has additional weaker bands following it. In an analogous pattern, NSD-TTRT, -TTRA, and -TTRC

have PAGE bands that are similar to each other and also to NSD-TT, with a single dominant band and some minor smearing. NSD-TTRG shows many conformations.

The four L constructs all have moderate-intensity LHGQ negative peaks at around 280 nm (**Figure 11B**). Three of the R constructs, NSD-TTRA, -TTRC, and -TTRT have full intensity LHGQ peaks and no right-handed behavior, whereas NSD-TTRG has a CD spectra with weak peaks (**Figure 11C**).



**Figure 11.** Biophysical characterization of NSD-TT overhang constructs. CD spectra of L (**A**) and R (**B**) constructs. NSD-TT included as a reference. In CD spectra, LHGQ sequences are

blue, right-left hybrid sequences are green, and inconclusive sequences are yellow. (C) 15% Native PAGE of constructs.

NSD-TTLG and -TTRG both lose the CD features attributed to the full LHGQ and several new conformations appear on PAGE. Therefore, 5' or 3'-G is disruptive to LHGQ beyond the *Motif1* or *Motif2* sequence. Interestingly, NSD-TTRT, -TTRA, and -TTRC remain fully left-handed and maintain homogeneity according to PAGE, which represents the first example we have observed of additions to a LHGQ motif not disrupting left-handed folding. Finally, all L variants, not only LG display disrupted LHGQ folding. Evidently, the 5' end of LHGQ *Motif1* is less tolerant to base addition than the 3' end of LHGQ *Motif2*. The natural next steps are to determine if the other ends of each motif can tolerate base additions to see if a combination of motifs could conceivably fold left-handed within a larger genetic context. More detail is provided in Chapter 6.

## Crystallography of NSD-TT and NSD-TT1L+NMM

I was able to produce diffraction-quality crystals of NSD-TT, solve the phase problem, and proceed to building and refinement (R-free: 0.3680). The structure is being solved in the C 2 2<sub>1</sub> space group with 2.40 Å data. There are five DNA chains in the ASU, and all five of the chains display a bilayered LHGQ. Further conclusions cannot be made without additional refinement. I also produced and optimized crystals of NSD-TT1L+NMM. The crystals diffracted to high resolution in the P 2 2<sub>1</sub> 2<sub>1</sub> space group, but I have so far been unsuccessful with molecular replacement. Details on crystallization can be found in the appendix.

## Conclusion

Chapter 4 focused on characterization of the left-handed folding of the NSD sequence with and without the addition of NMM. I found that the NSD-TT construct, consisting of both LHGQ minimal motifs, formed a LHGQ. Adding a G to either LHGQ motif disrupted the LHGQ folding of that motif, but I found that LHGQ *Motif2* permitted 3' additions of A, C, and T. We also observed that NMM is capable of coercing LHGQ *Motif2* to fold right-handed but does not appear to be able to do so for LHGQ *Motif1*. I also reported the partially built crystal structure of NSD-TT, which I have so far confirmed to be a two-subunit LHGQ but have not resolved the loops.

In addition to the proposed future directions described in chapter 6, significant work remains to be done to validate the results described in this chapter. As with chapter 3, additional biophysical replicates with new DNA stocks must be performed, detailed in **Table A10**. Stability data has yet to be fully processed. The NSD-TT structure also needs additional refinement to resolve the loops.

## Chapter 5: Characterization of C-rich sequence from *Homo sapiens* platelet derived growth factor subunit B (PDGF-B)

This section focuses on the characterization of an iM forming sequence in the promoter region of *Homo sapiens* platelet derived growth factor (PDGF), an important mitogen for smooth muscle cells, fibroblasts, and platelets<sup>36</sup>. PDGF is associated with wound healing, cell differentiation, and tumorigenesis<sup>37,38</sup>. Using a BLAST search, our lab identified three overlapping sets of C repeats and designated them PDGF-A, -B, and -C . I worked with the (CCCCG)<sub>3</sub>C<sub>4</sub> PDGF-B sequence. We designed thirteen different constructs (labeled B1-B13) by extending or truncating the ends of the PDGF-B reference sequence according to genetic context (**Table 4**). The reason for creating these native constructs was twofold: first, to determine if and how extending the PDGF sequence would affect iM folding, and second to give us more opportunities to crystallize and solve iM structures. Crystallization depends strongly on the contact surface, thus 5'- and 3'-overhang play important roles in the crystallization process.

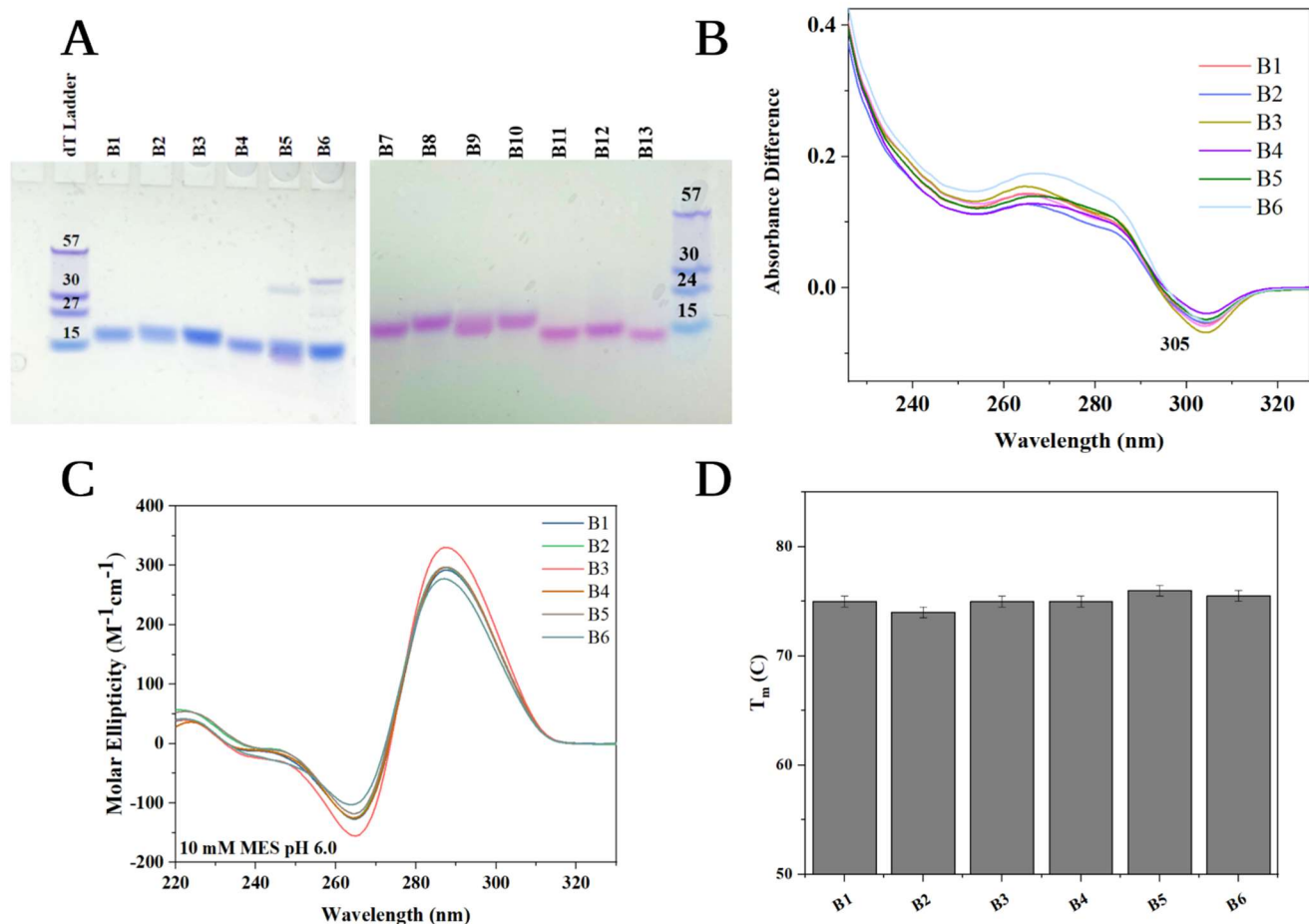
**Table 4.** i-Motif DNA sequences studied in this work. Nucleotides in red denote additions to the reference sequence. PDGF-A and PDGF-C sequences are included for reference and are aligned according to their overlap with PDGF-B. The molar extinction coefficient  $\epsilon$  and molecular weight were calculated using the Integrated DNA Technology OligoAnalyzer tool.

Name	Sequence	Length	$\epsilon$ , mM <sup>-1</sup> cm <sup>-1</sup>	MW, g/mol	T <sub>m</sub> Derivative, K; error $\pm 0.5$
PDGF-A	GCCCCTCCCCCGCCCCGCCCCCG	-	-	-	-
PDGF-B	CCCCCGCCCCCGCCCCCGCCCCC	23	172.7	6709.3	-
PDGF-C	CCGCCCCCCCCCCCCCGCCTCCCCCGG				-
PDGF-B1	TCCCCCGCCCCCGCCCCCGCCCCC	24	181.5	7013.5	348
PDGF-B2	CTCCCCCGCCCCCGCCCCCGCCCCC	25	188.0	7302.7	347
PDGF-B3	CCTCCCCCGCCCCCGCCCCCGCCCCC	26	195.2	7591.9	348
PDGF-B4	CCCCCGCCCCCGCCCCCGCCCCC	24	179.9	6998.5	348
PDGF-B5	CCCCCGCCCCCGCCCCCGCCCCC	25	187.1	7287.7	349
PDGF-B6	TCCCCCGCCCCCGCCCCCGCCCCC	25	188.7	7302.7	348.5
PDGF-B7	CTCCCCCGCCCCCGCCCCCGCCCCC	26	195.2	7591.9	-
PDGF-B8	CCTCCCCCGCCCCCGCCCCCGCCCCC	27	202.4	7881.1	-
PDGF-B9	CCCTCCCCCGCCCCCGCCCCCGCCCCC	27	202.4	7881.1	-
PDGF-B10	TCCCCCGCCCCCGCCCCCGCCCCC	26	195.9	7591.9	-
PDGF-B11	CCCCGCCCCCGCCCCCGCCCCC	21	158.3	6131.0	-
PDGF-B12	CCCTCCCCCGCCCCCGCCCCCG	22	167.5	6435.2	-
PDGF-B13	CCCTCCCCCGCCCCCGCCCCC	21	156.9	6106.0	-

## Biophysical characterization reveals that iM folding depends solely on the C-rich part of the sequence.

I first determined the homogeneity and oligomeric state of the constructs using PAGE (**Figure 12A**). All constructs except B5 and B6 are homogenous and without higher order species. I screened all thirteen constructs for their ability to form 3D crystals and made an arbitrary selection of the first six constructs for further time-intensive biophysical characterization.

I determined the TDS and CD signatures of six constructs to confirm iM folding. The TDS spectra display an iM trough at 305 nm<sup>32</sup> for each construct (**Figure 12B**). This position is different from the 295 nm typically observed in the cacodylate buffers and is likely due to the choice of MES buffer for this project. The CD signatures for all six constructs have a trough and peak at 264 nm and 287 nm respectively, also suggesting i-motif folding (**Figure 12C**). Subsequently, I assessed the stability of B1-B6 with UV-vis melting (**Figure 12D**). UV melts work similarly to CD melts but monitor UV absorbance rather than CD signal. The hysteresis did not exceed 3 °C for any construct, indicating reversible melting transitions.



**Figure 12. Biophysical characterization of PDGF-B.** (A) Fifteen percent native PAGE gels for all thirteen constructs. (B) TDS, (C) CD scans, and (D)  $T_m$  did not differ significantly between the first six constructs.

The iM folding of PDGF-B is robust. Native constructs created by extending the sequence in the 5' and 3' directions have similar folding, and the biophysical behavior of the variants is determined by the iM core rather than the overhangs. This bodes well for using native



constructs to crystallize and solve iM DNA structures. I have had only limited success crystallizing iM sequences; I have found weakly diffracting crystals for a single construct, PDGF-B7, but I have so far been unable to reproduce these.

## Conclusion

Chapter 5 described biophysical characterization of PDGF-B iM forming constructs. I found that the PDGF sequence has robust iM formation and variations to the sequence did not disrupt iM formation or stability. I attempted to crystallize the thirteen constructs I designed but was not successful in producing diffraction quality crystals. Future work for this project would lie with continuing crystallization attempts, perhaps with an additive screen.

## Chapter 6: Conclusions and future directions.

My thesis focused on characterizing non-canonical DNA structures biophysically and structurally. The GQ and the iM are formed by repetitive G- and C-rich sequences which are enriched in telomeres and promoter regions of oncogenes, among other parts of human genome. GQs and iMs play regulatory roles *in vivo* and are therefore relevant to a variety of disease states. I specifically investigated the relatively recently discovered phenomenon of LHGQ, whose *in vivo* existence and biological relevance has not yet been established.

In Chapters 3 and 4, we investigated two different sequences from the human genome which contain LHGQ minimal motifs. We investigated the sequence dependence of LHGQ folding and found that sequence overhangs play an important role in determining if LHGQ or RHGQ is formed. In interrogating the interaction of NMM with LHGQ DNA, we found that NMM does not bind LHGQ DNA and instead favors the formation of RHGQ DNA. We also reported the second crystal structure of a right-left hybrid GQ, as well as a preliminary solution of a novel LHGQ.

The research in Chapters 3 and 4 presents multiple options for future investigation. It may be the case that NMM is only capable of coercing LHGQ *Motif2* to fold right-handed, whereas LHGQ *Motif1* is not affected. This could also simply be due to the placement of the motifs within the GQ structure – in both SLC and NSD-TT, *Motif1* was present in the 5' half of the sequence.

I would need new constructs to test this. Our lab has found that isolated LHGQ motifs fold left-handed but are not thermodynamically stable, so working with a 2-motif GQ structure is preferable. *2xMotif1*<sup>5</sup> and *2xMotif2*<sup>6</sup> are previously characterized LHGQs consisting of two

instances of their respective LHGQ motif connected by TT linkers. As a test to my hypothesis, it would be valuable to anneal these sequences with NMM and check their CD spectra. If *Motif1* does not interact with NMM and *Motif2* does, then  $2x\textit{Motif1}$ +NMM is expected to be fully left-handed and  $2x\textit{Motif2}$ +NMM is expected to be fully right-handed. Alternatively, if the interaction is determined by the placement of the motif within the structure, both  $2x\textit{Motif1}$ +NMM and  $2x\textit{Motif2}$ +NMM should be right-left GQ hybrids.

The exploration of the NSD-TT overhang constructs also merits further investigation. In NSD-TTRA, -TTRC, and -TTRT, we observed that when LHGQ *Motif2* made up the 3' half of an LHGQ structure, 3' additions of A, C, T were tolerated. Conversely, 5' additions of A, T, or C to LHGQ *Motif1* and 5' and 3' additions of G were not tolerated. To determine if LHGQ formation is possible within a larger sequence, new constructs testing the other possible combinations of overhangs are necessary. Starting with  $2x\textit{Motif1}$  and  $2x\textit{Motif2}$ , I would create sixteen new constructs: each of the four possible flanking bases on both the 5' and 3' sides, for both base sequences. CD, TDS, and PAGE of the new constructs would reveal if the observed patterns persisted. Specifically, I would be seeking to determine if any LHGQ sequence could tolerate both 3' and 5' additions. If no such construct exists, LHGQ DNA is unlikely to form *in vivo*.

The work in this thesis adds to our understanding of non-canonical DNA, specifically bringing us closer to determining the biological relevance of LHGQ DNA. The crystal structures presented similarly contribute to knowledge of LHGQ structural features and behavior. Our work with NMM adds to knowledge of GQ-ligand interactions towards the eventual goal of development of therapeutics.

## References

- (1) Sen, D.; Gilbert, W. Formation of Parallel Four-Stranded Complexes by Guanine-Rich Motifs in DNA and Its Implications for Meiosis. *Nature* **1988**, *334* (6180), 364–366. <https://doi.org/10.1038/334364a0>.
- (2) Zeraati, M.; Langley, D. B.; Schofield, P.; Moye, A. L.; Rouet, R.; Hughes, W. E.; Bryan, T. M.; Dinger, M. E.; Christ, D. I-Motif DNA Structures Are Formed in the Nuclei of Human Cells. *Nat. Chem.* **2018**, *10* (6), 631–637. <https://doi.org/10.1038/s41557-018-0046-3>.
- (3) Wang, A. H.-J.; Quigley, G. J.; Kolpak, F. J.; Crawford, J. L.; van Boom, J. H.; van der Marel, G.; Rich, A. Molecular Structure of a Left-Handed Double Helical DNA Fragment at Atomic Resolution. *Nature* **1979**, *282* (5740), 680–686. <https://doi.org/10.1038/282680a0>.
- (4) Chung, W. J.; Heddi, B.; Schmitt, E.; Lim, K. W.; Mechulam, Y.; Phan, A. T. Structure of a Left-Handed DNA G-Quadruplex. *Proc. Natl. Acad. Sci.* **2015**, *112* (9), 2729–2733. <https://doi.org/10.1073/pnas.1418718112>.
- (5) Bakalar, B.; Heddi, B.; Schmitt, E.; Mechulam, Y.; Phan, A. T. A Minimal Sequence for Left-Handed G-Quadruplex Formation. *Angew. Chem. Int. Ed Engl.* **2019**, *58* (8), 2331–2335. <https://doi.org/10.1002/anie.201812628>.
- (6) Das, P.; Winnerdy, F. R.; Maity, A.; Mechulam, Y.; Phan, A. T. A Novel Minimal Motif for Left-Handed G-Quadruplex Formation. *Chem. Commun.* **2021**, *57* (20), 2527–2530. <https://doi.org/10.1039/D0CC08146A>.
- (7) Winnerdy, F. R.; Bakalar, B.; Maity, A.; Vandana, J. J.; Mechulam, Y.; Schmitt, E.; Phan, A. T. NMR Solution and X-Ray Crystal Structures of a DNA Molecule Containing Both Right- and Left-Handed Parallel-Stranded G-Quadruplexes. *Nucleic Acids Res.* **2019**, *47* (15), 8272–8281. <https://doi.org/10.1093/nar/gkz349>.
- (8) Fu, B.; Huang, J.; Chen, Y.; Wang, Y.; Xue, T.; Xu, G.; Wang, S.; Zhou, X. Right-Handed and Left-Handed G-Quadruplexes Have the Same DNA Sequence: Distinct Conformations Induced by an Organic Small Molecule and Potassium. *Chem. Commun.* **2016**, *52* (65), 10052–10055. <https://doi.org/10.1039/C6CC04866H>.
- (9) Nicoludis, J. M.; Miller, S. T.; Jeffrey, P. D.; Barrett, S. P.; Rablen, P. R.; Lawton, T. J.; Yatsunyk, L. A. Optimized End-Stacking Provides Specificity of *N*-Methyl Mesoporphyrin IX for Human Telomeric G-Quadruplex DNA. *J. Am. Chem. Soc.* **2012**, *134* (50), 20446–20456. <https://doi.org/10.1021/ja3088746>.
- (10) Cui, J.; Waltman, P.; Le, V. H. & Lewis, E. A. The Effect of Molecular Crowding on the Stability of Human C-MYC Promoter Sequence I-Motif at Neutral PH. *Molecules* **2013**, *18*, 12751–12767.
- (11) Sun, D.; Hurley, L. H. The Importance of Negative Superhelicity in Inducing the Formation of G-Quadruplex and i-Motif Structures in the c-Myc Promoter: Implications for Drug Targeting and Control of Gene Expression. *J. Med. Chem.* **2009**, *52* (9), 2863–2874. <https://doi.org/10.1021/jm900055s>.
- (12) Tassinari, M.; Richter, S. N.; Gandellini, P. Biological Relevance and Therapeutic Potential of G-Quadruplex Structures in the Human Noncoding Transcriptome. *Nucleic Acids Res.* **2021**, *49* (7), 3617–3633. <https://doi.org/10.1093/nar/gkab127>.
- (13) Broccoli, D.; Young, J. W.; de Lange, T. Telomerase Activity in Normal and Malignant Hematopoietic Cells. *Proc. Natl. Acad. Sci.* **1995**, *92* (20), 9082–9086. <https://doi.org/10.1073/pnas.92.20.9082>.

- (14) Hiyama, K.; Hiyama, E.; Ishioka, S.; Yamakido, M.; Inai, K.; Gazdar, A. F.; Piatyszek, M. A.; Shay, J. W. Telomerase Activity in Small-Cell and Non-Small-Cell Lung Cancers. *JNCI J. Natl. Cancer Inst.* **1995**, *87* (12), 895–902. <https://doi.org/10.1093/jnci/87.12.895>.
- (15) Hiyama, K.; Ishioka, S.; Shay, J. W.; Taooka, Y.; Maeda, A.; Isobe, T.; Hiyama, E.; Maeda, H.; Yamakido, M. Telomerase Activity as a Novel Marker of Lung Cancer and Immune-Associated Lung Diseases. *Int. J. Mol. Med.* **1998**, *1* (3), 545–554. <https://doi.org/10.3892/ijmm.1.3.545>.
- (16) Sun, D.; Thompson, B.; Cathers, B. E.; Salazar, M.; Kerwin, S. M.; Trent, J. O.; Jenkins, T. C.; Neidle, S.; Hurley, L. H. Inhibition of Human Telomerase by a G-Quadruplex-Interactive Compound. *J. Med. Chem.* **1997**, *40* (14), 2113–2116. <https://doi.org/10.1021/jm970199z>.
- (17) Seimiya, H.; Nagasawa, K.; Shin-ya, K. Chemical Targeting of G-Quadruplexes in Telomeres and beyond for Molecular Cancer Therapeutics. *J. Antibiot. (Tokyo)* **2021**, *74* (10), 617–628. <https://doi.org/10.1038/s41429-021-00454-x>.
- (18) Johnson, J. E.; Cao, K.; Ryvkin, P.; Wang, L.-S.; Johnson, F. B. Altered Gene Expression in the Werner and Bloom Syndromes Is Associated with Sequences Having G-Quadruplex Forming Potential. *Nucleic Acids Res.* **2010**, *38* (4), 1114–1122. <https://doi.org/10.1093/nar/gkp1103>.
- (19) Niu, K.; Zhang, X.; Deng, H.; Wu, F.; Ren, Y.; Xiang, H.; Zheng, S.; Liu, L.; Huang, L.; Zeng, B.; Li, S.; Xia, Q.; Song, Q.; Palli, S. R.; Feng, Q. BmILF and I-Motif Structure Are Involved in Transcriptional Regulation of BmPOUM2 in Bombyx Mori. *Nucleic Acids Res.* **2018**, *46* (4), 1710–1723. <https://doi.org/10.1093/nar/gkx1207>.
- (20) Rhodes, D.; Lipps, H. J. G-Quadruplexes and Their Regulatory Roles in Biology. *Nucleic Acids Res.* **2015**, *43* (18), 8627–8637. <https://doi.org/10.1093/nar/gkv862>.
- (21) Lin, L. Y.; McCarthy, S.; Powell, B. M.; Manurung, Y.; Xiang, I. M.; Dean, W. L.; Chaires, B.; Yatsunyk, L. A. Biophysical and X-Ray Structural Studies of the (GGGTT)3GGG G-Quadruplex in Complex with N-Methyl Mesoporphyrin IX. *PLOS ONE* **2020**, *15* (11), e0241513. <https://doi.org/10.1371/journal.pone.0241513>.
- (22) Ye, M.; Chen, E. V.; Pfeil, S. H.; Martin, K. N.; Atrafi, T.; Yun, S.; Martinez, Z.; Yatsunyk, L. A. Homopurine Guanine-Rich Sequences in Complex with N-Methyl Mesoporphyrin IX Form Parallel G-Quadruplex Dimers and Display a Unique Symmetry Tetrad. *Bioorg. Med. Chem.* **2023**, *77*, 117112. <https://doi.org/10.1016/j.bmc.2022.117112>.
- (23) Nicoludis, J. M.; Barrett, S. P.; Mergny, J.; Yatsunyk, L. Interaction of Human Telomeric DNA with N-Methyl Mesoporphyrin IX. *Nucleic Acids Res.* **2012**. <https://doi.org/10.1093/nar/gks152>.
- (24) Yett, A.; Lin, L. Y.; Beseiso, D.; Miao, J.; Yatsunyk, L. A. N-Methyl Mesoporphyrin IX as a Highly Selective Light-up Probe for G-Quadruplex DNA. *J. Porphyr. Phthalocyanines* **2019**, *23* (11n12), 1195–1215. <https://doi.org/10.1142/s1088424619300179>.
- (25) Vazquez-Chantada, M.; Gonzalez-Lahera, A.; Martinez-Arranz, I.; Garcia-Monzon, C.; Regueiro, M. M.; Garcia-Rodriguez, J. L.; Schlangen, K. A.; Mendibil, I.; Rodriguez-Ezpeleta, N.; Lozano, J. J.; Banasik, K.; Justesen, J. M.; Joergensen, T.; Witte, D. R.; Lauritzen, T.; Hansen, T.; Pedersen, O.; Veyrie, N.; Clement, K.; Tordjman, J.; Tran, A.; Le Marchand-Brustel, Y.; Buque, X.; Aspichueta, P.; Echevarria-Uraga, J. J.; Martin-Duce, A.; Caballeria, J.; Gual, P.; Castro, A.; Mato, J. M.; Martinez-Chantar, M. L.; Aransay, A. M. Solute Carrier Family 2 Member 1 Is Involved in the Development of Nonalcoholic Fatty

- Liver Disease. *Hepatol. Baltim. Md* **2013**, 57 (2), 505–514. <https://doi.org/10.1002/hep.26052>.
- (26) Peng, Q.; Hao, L.-Y.; Guo, Y.-L.; Zhang, Z.-Q.; Ji, J.-M.; Xue, Y.; Liu, Y.-W.; Lu, J.-L.; Li, C.-G.; Shi, X.-L. Solute Carrier Family 2 Members 1 and 2 as Prognostic Biomarkers in Hepatocellular Carcinoma Associated with Immune Infiltration. *World J. Clin. Cases* **2022**, 10 (13), 3989–4019. <https://doi.org/10.12998/wjcc.v10.i13.3989>.
  - (27) He, L.; Vasiliou, K.; Nebert, D. W. Analysis and Update of the Human Solute Carrier (SLC) Gene Superfamily. *Hum. Genomics* **2009**, 3 (2), 195. <https://doi.org/10.1186/1479-7364-3-2-195>.
  - (28) Chen, Y.; Li, X.; Xu, J.; Xiao, H.; Tang, C.; Liang, W.; Zhu, X.; Fang, Y.; Wang, H.; Shi, J. Knockdown of Nuclear Receptor Binding SET Domain-Containing Protein 1 (NSD1) Inhibits Proliferation and Facilitates Apoptosis in Paclitaxel-Resistant Breast Cancer Cells via Inactivating the Wnt/ $\beta$ -Catenin Signaling Pathway. *Bioengineered* **2022**, 13 (2), 3526–3536. <https://doi.org/10.1080/21655979.2021.2018973>.
  - (29) Chen, Y.; Tang, W.; Zhu, X.; Zhang, L.; Zhu, Y.; Xiao, H.; Xu, J.; Fang, Y.; Li, X.; Tang, C.; Shi, J. Nuclear Receptor Binding SET Domain Protein 1 Promotes Epithelial-Mesenchymal Transition in Paclitaxel-Resistant Breast Cancer Cells via Regulating Nuclear Factor Kappa B and F-Box and Leucine-Rich Repeat Protein 11. *Bioengineered* **2021**, 12 (2), 11506–11519. <https://doi.org/10.1080/21655979.2021.2009963>.
  - (30) Bianco-Miotto, T.; Chiam, K.; Buchanan, G.; Jindal, S.; Day, T. K.; Thomas, M.; Pickering, M. A.; O’Loughlin, M. A.; Ryan, N. K.; Raymond, W. A.; Horvath, L. G.; Kench, J. G.; Stricker, P. D.; Marshall, V. R.; Sutherland, R. L.; Henshall, S. M.; Gerald, W. L.; Scher, H. I.; Risbridger, G. P.; Clements, J. A.; Butler, L. M.; Tilley, W. D.; Horsfall, D. J.; Ricciardelli, C. Global Levels of Specific Histone Modifications and an Epigenetic Gene Signature Predict Prostate Cancer Progression and Development. *Cancer Epidemiol. Biomarkers Prev.* **2010**, 19 (10), 2611–2622. <https://doi.org/10.1158/1055-9965.EPI-10-0555>.
  - (31) Ross, R.; Raines, E. W.; Bowen-Pope, D. F. The Biology of Platelet-Derived Growth Factor. *Cell* **1986**, 46 (2), 155–169. [https://doi.org/10.1016/0092-8674\(86\)90733-6](https://doi.org/10.1016/0092-8674(86)90733-6).
  - (32) Mergny, J.-L.; Li, J.; Lacroix, L.; Amrane, S.; Chaires, J. B. Thermal Difference Spectra: A Specific Signature for Nucleic Acid Structures. *Nucleic Acids Res.* **2005**, 33 (16), e138. <https://doi.org/10.1093/nar/gni134>.
  - (33) Ramsay, G. D.; Eftink, M. R. [27] Analysis of Multidimensional Spectroscopic Data to Monitor Unfolding of Proteins. In *Methods in Enzymology*; Part B: Numerical Computer Methods; Academic Press, 1994; Vol. 240, pp 615–645. [https://doi.org/10.1016/S0076-6879\(94\)40066-0](https://doi.org/10.1016/S0076-6879(94)40066-0).
  - (34) Nicoludis, J. M.; Barrett, S. P.; Mergny, J.-L.; Yatsunyk, L. A. Interaction of Human Telomeric DNA with N-Methyl Mesoporphyrin IX. *Nucleic Acids Res.* **2012**, 40 (12), 5432–5447. <https://doi.org/10.1093/nar/gks152>.
  - (35) Kikin, O.; D’Antonio, L.; Bagga, P. S. QGRS Mapper: A Web-Based Server for Predicting G-Quadruplexes in Nucleotide Sequences. *Nucleic Acids Res.* **2006**, 34 (Web Server issue), W676–W682. <https://doi.org/10.1093/nar/gkl253>.
  - (36) Antoniades, H. N.; Williams, L. T. Human Platelet-Derived Growth Factor: Structure and Function. *Fed. Proc.* **1983**, 42 (9), 2630–2634.

- (37) Shih, A. H.; Holland, E. C. Platelet-Derived Growth Factor (PDGF) and Glial Tumorigenesis. *Cancer Lett.* **2006**, 232 (2), 139–147.  
<https://doi.org/10.1016/j.canlet.2005.02.002>.
- (38) Soriano, P. Abnormal Kidney Development and Hematological Disorders in PDGF Beta-Receptor Mutant Mice. *Genes Dev.* **1994**, 8 (16), 1888–1896.  
<https://doi.org/10.1101/gad.8.16.1888>.

## Appendix

**Table A1.** SLC GQ DNA sequences studied in this work. Bases in red denote additions to the reference sequence.

<i>Name</i>	<i>Sequence</i>	<i>Length</i>	$\epsilon, \text{mM}^{-1}\text{cm}^{-1}$	<i>MW, g/mol</i>
<i>SLC</i>	GTGGTGGTGGTGGTGGTGGTGG	25	245.2	7952.1
<i>SLC-1R</i>	GTGGTGGTGGTGGTGGTGGTGGC	26	251.3	8241.3
<i>SLC-1L</i>	GGTGGTGGTGGTGGTGGTGGTGG	26	255.3	8281.3
<i>SLC-2R</i>	GTGGTGGTGGTGGTGGTGGTGGCA	27	265.1	8554.5
<i>SLC-2L</i>	GGGTGGTGGTGGTGGTGGTGGTGG	27	265.4	8610.6
<i>SLC-1LR</i>	GGTGGTGGTGGTGGTGGTGGTGGC	27	261.4	8570.5
<i>SLC-2LR</i>	GGGTGGTGGTGGTGGTGGTGGTGGCA	29	285.3	9213.0
<i>SLC-2L1R</i>	GGGTGGTGGTGGTGGTGGTGGTGGC	28	271.5	8899.7
<i>SLC-1L2R</i>	GGTGGTGGTGGTGGTGGTGGTGGCA	28	275.2	8883.7
<i>4U5M</i>	TGGTGGTGGTGGTGTGGTGGTGTGTT	28	251.4	8855.7



**Table A2.** NSD GQ DNA sequences studied in this work. Bases in red denote additions to the reference sequence.

<i>Name</i>	<i>Sequence</i>	<i>Length</i>	$\varepsilon, mM^{-1}cm^{-1}$	<i>MW, g/mol</i>
<i>NSD</i>	GTGGTGGTGGTGGTGGTGGTGGTGTG	26	251.4	8272.3
<i>NSD-1L</i>	TGTGGTGGTGGTGGTGGTGGTGGTGTG	27	258.9	8576.5
<i>NSD-2L</i>	CTGTGGTGGTGGTGGTGGTGGTGGTGTG	28	265.4	8865.7
<i>NSD-1R</i>	GTGGTGGTGGTGGTGGTGGTGGTGTG <b>G</b>	27	261.5	8601.5
<i>NSD-2R</i>	GTGGTGGTGGTGGTGGTGGTGGTGTG <b>GG</b>	28	271.6	8930.8
<i>NSD-1LR</i>	TGTGGTGGTGGTGGTGGTGGTGGTGTG <b>G</b>	28	269.0	8905.7
<i>NSD-2LR</i>	CTGTGGTGGTGGTGGTGGTGGTGGTGTG <b>GG</b>	30	275.5	9524.1
<i>NSD-12LR</i>	TGTGGTGGTGGTGGTGGTGGTGGTGTG <b>GG</b>	29	279.1	9234.9
<i>NSD-21LR</i>	CTGTGGTGGTGGTGGTGGTGGTGGTGTG <b>G</b>	29	275.5	9194.9
<i>NSD-TT</i>	GTGGTGGTGGTGT <b>T</b> GGTGGTGGTGTG	26	249.4	8247.3
<i>NSD-TT1L</i>	TGTGGTGGTGGTGT <b>T</b> GGTGGTGGTGTG	27	256.9	8551.5
<i>NSD-TT1R</i>	GTGGTGGTGGTGT <b>T</b> GGTGGTGGTGTG <b>G</b>	27	259.5	8576.5

**Table A3.** CD melt Stability Data for SLC constructs. Data calculated as described in the methods. Melt and table were both produced by Eric Xing.

	Concentration of DNA ( $\mu\text{M}$ )	Heating			Cooling			
Seq		T <sub>m</sub> -Der (°C)	T <sub>m</sub> -fit (°C)	$\Delta\text{H}$ fit (kJ/mol)	T <sub>m</sub> -Der (°C)	T <sub>m</sub> -fit (°C)	$\Delta\text{H}$ fit (kJ/mol)	Hysteresis (°C)
SLC	4.10	74.0 $\pm$ 0.5	74.4 $\pm$ 0.3	383 $\pm$ 22	59.0 $\pm$ 0.5	58.8 $\pm$ 0.07	259 $\pm$ 4	15.0
SLC-1L	4.44	61.0 $\pm$ 0.5	62.62 $\pm$ 0.06	344 $\pm$ 6	60.0 $\pm$ 0.5	61.4 $\pm$ 0.1	335 $\pm$ 9	1.0
SLC-1R	3.82	N/A*	N/A*	N/A*	51.0 $\pm$ 0.5	55.7 $\pm$ 1.3	116 $\pm$ 14	N/A*
SLC-2R (1)	3.77	56.0 $\pm$ 1.0	55.5 $\pm$ 0.2	194 $\pm$ 7	54.0 $\pm$ 1.0	52.5 $\pm$ 0.2	215 $\pm$ 7	2.0
SLC-2L	3.98	59.0 $\pm$ 0.5	60.4 $\pm$ 0.9	326 $\pm$ 5	59.0 $\pm$ 0.5	59.67 $\pm$ 0.05	310 $\pm$ 4	0
SLC-1LR	4.14	60.0 $\pm$ 0.5	59.38 $\pm$ 0.06	309 $\pm$ 5	57.0 $\pm$ 1.0	57.8 $\pm$ 0.1	301 $\pm$ 9	3.0
SLC-2LR	4.38	56.0 $\pm$ 0.5	56.87 $\pm$ 0.03	294 $\pm$ 2	56.0 $\pm$ 0.5	56.32 $\pm$ 0.02	273 $\pm$ 2	0
SLC-2L1R	1.34	56.0 $\pm$ 0.5	57.2 $\pm$ 0.06	302 $\pm$ 4	56.0 $\pm$ 0.5	56.41 $\pm$ 0.05	279 $\pm$ 4	0
SLC-1L2R	4.11	59.0 $\pm$ 0.5	59.56 $\pm$ 0.07	308 $\pm$ 5	59.0 $\pm$ 0.5	59.13 $\pm$ 0.04	290 $\pm$ 3	0

**Table A3. Stabilization of SLC by NMM, CD melt**

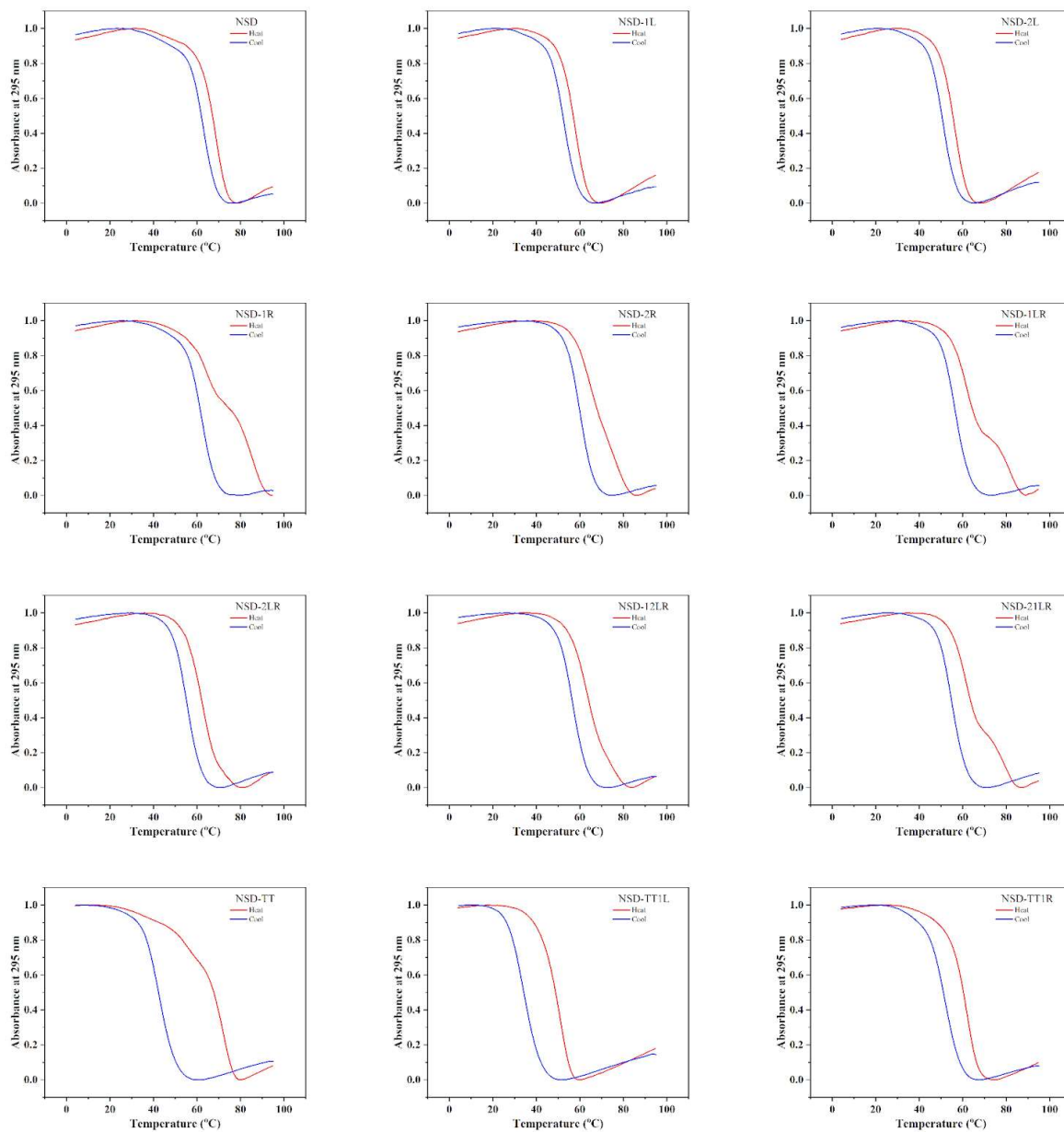
	<b>HEATING</b>		<b>COOLING</b>	
Sequence	<b>T<sub>m</sub> Derivative, K; error ± 0.5</b>	<b>T<sub>m</sub> Derivative, C; error ± 0.5</b>	T <sub>m</sub> Derivative, K; error ± 0.5	T <sub>m</sub> Derivative, C; error ± 0.5
SLC	339	66.0	327.3	54.3
SLC+1NMM	309.3	70.3	341	68.0
SLC+2NMM	344	71.0	340.3	67.3
SLC+5NMM	348.2	75.2	343.1	70.1
SLC+10NMM	345.4	72.4	342.2	69.2

**Table A5.** Stability of NSD-TT overhang constructs in 100k buffer.

Sequence	Heating		Cooling	
	Tm Derivative (K) Error C; error $\pm 0.5$	Tm Derivative (C) C; error $\pm 0.5$	Tm Derivative (K) C; error $\pm 0.5$	Tm Derivative (C) C; error $\pm 0.5$
NSD-TTLT	324	51	307	34
NSD-TTLA	334	61	319	46
NSD-TTLC	317	44	307.5	34.5
NSD-TTLG	326.5	53.5	323	50
NSD-TTRG	335	62	325	52
NSD-TTRA	335	62	314.5	41.5
NSD-TTRC	335	62	314	41
NSD-TTRT	334.5	61.5	315	42

**Table A6.** i-Motif DNA sequences studied in this work. Nucleotides in red denote additions to the reference sequence. PDGF-A and PDGF-C sequences are included for reference and are aligned according to their overlap with PDGF-B.

<i>Name</i>	<i>Sequence</i>	<i>Length</i>	$\epsilon, \text{mM}^{-1}\text{cm}^{-1}$	<i>MW, g/mol</i>
<i>PDGF-A</i>	GCCCCTCCCCCGCCCCGCCCCCG	-	-	-
<i>PDGF-B</i>	CCCCCGCCCCCGCCCCGCCCCC	23	172.7	6709.3
<i>PDGF-C</i>	CCGCCCCCCCCCCCCCGCCTCCCCCGG			
<i>PDGF-B1</i>	TCCCCCGCCCCCGCCCCGCCCCC	24	181.5	7013.5
<i>PDGF-B2</i>	CTCCCCCGCCCCCGCCCCGCCCCC	25	188	7302.7
<i>PDGF-B3</i>	CCTCCCCCGCCCCCGCCCCGCCCCC	26	195.2	7591.9
<i>PDGF-B4</i>	CCCCCGCCCCCGCCCCGCCCCC	24	179.9	6998.5
<i>PDGF-B5</i>	CCCCCGCCCCCGCCCCGCCCCC	25	187.1	7287.7
<i>PDGF-B6</i>	TCCCCCGCCCCCGCCCCGCCCCC	25	188.7	7302.7
<i>PDGF-B7</i>	CTCCCCCGCCCCCGCCCCGCCCCC	26	195.2	7591.9
<i>PDGF-B8</i>	CCTCCCCCGCCCCCGCCCCGCCCCC	27	202.4	7881.1
<i>PDGF-B9</i>	CCCTCCCCCGCCCCCGCCCCGCCCCC	27	202.4	7881.1
<i>PDGF-B10</i>	TCCCCCGCCCCCGCCCCGCCCCC	26	195.9	7591.9
<i>PDGF-B11</i>	CCCCGCCCCCGCCCCGCCCCC	21	158.3	6131.0
<i>PDGF-B12</i>	CCCTCCCCCGCCCCCGCCCCG	22	167.5	6435.2
<i>PDGF-B13</i>	CCCTCCCCCGCCCCCGCCCCC	21	156.9	6106.0



**Figure A1.** Representative melting experiment data of the NSD constructs. Above graphs can be used to visually assess the hysteresis of melting. A large gap between the melting and cooling curves represents high hysteresis.

**Table A6.** Details on datasets used for solving the SLC+NMM crystal structure.

Dataset name	PS25_2_A2_1_run1_1	PS25_2_A5_3_run1_1	PS25_2_A2_1_run2_1	PS25_2_A1_3_run1_1
Resolution (Å)	1.74	1.79	1.74	1.76
I/ $\sigma$ Overall	28	30.9	35.9	24.4
Rmerge Overall	0.033	0.036	0.047	0.050
ISa	46.04	41.37	35.35	32.35
Tray	PS1-25-2 (24 well)	PS1-25-2 (24 well)	PS1-25-2 (24 well)	PS1-25-2 (24 well)
DNA	0.5 mM HPLC SLC	0.5 mM HPLC SLC	0.5 mM HPLC SLC	0.5 mM HPLC SLC
NMM	1 eq	1 eq	1 eq	1 eq
Crystallization condition	2mM Spermine, 45% MPD, 0.08M NaCl, 0.04M NaCaCo pH 6.0 (PS1-23-1-1)	6mM Spermine, 45% MPD, 0.08M NaCl, 0.04M NaCaCo pH 6.0 (PS1-23-1-2)	2mM Spermine, 45% MPD, 0.08M NaCl, 0.04M NaCaCo pH 6.0 (PS1-23-1-1)	2mM Spermine, 45% MPD, 0.08M NaCl, 0.04M NaCaCo pH 6.0 (PS1-23-1-1)

**Table A7.** Details on optimization of SLC-1R crystals.

condition	crystal type	Salt 1	Salt 2	Buffer	MPD (%)	spermine (mM)
natrrix 2-33	little square	0.04M LiCl	0.08M SrCl <sub>2</sub> 6H <sub>2</sub> O	0.04M NaCaCo pH 7	30	12
natrrix 2-17	little square	0.08M KCl	0.02M BaCl <sub>2</sub> 2H <sub>2</sub> O	0.04M NaCaCo pH 6	40	12
natrrix 2-20	little square	0.08M NaCl	0.02M BaCl <sub>2</sub> 2H <sub>2</sub> O	0.04M NaCaCo pH 6	45	12
helix 2-38	hexagon	0.25M NaCl	-	0.05M Bis-tris pH 7.0	20% Peg	5
PS1-32-1-1	very small square	0.04M LiCl	0.08M SrCl <sub>2</sub> 6H <sub>2</sub> O	0.04M NaCaCo pH 6.5	30	12
PS1-32-1-2	very small square	0.04M LiCl	0.08M SrCl <sub>2</sub> 6H <sub>2</sub> O	0.04M NaCaCo pH 6.5	35	12
PS1-32-1-3	extremely small square	0.04M LiCl	0.08M SrCl <sub>2</sub> 6H <sub>2</sub> O	0.04M NaCaCo pH 6.5	40	12
PS1-32-1-4	extremely small square	0.04M LiCl	0.08M SrCl <sub>2</sub> 6H <sub>2</sub> O	0.04M NaCaCo pH 6.5	45	12

**Table A8.** Details on crystal optimization attempts for NSD-TT1L. All conditions had 2 equivalents of NMM.

<b>condition</b>	<b>crystal type</b>	<b>crystal quantity/consistency</b>	<b>Salt 1</b>	<b>Salt 2</b>	<b>Buffer</b>	<b>MPD (%)</b>	<b>spermine (mM)</b>
matrix 2-15	Urchin	bad (did not regrow)	0.08M NaCl	0.02M MgCl <sub>2</sub> 6H <sub>2</sub> O	0.04M NaCaCo pH 6	35	12
matrix 2-18	Urchin	bad (did not regrow)	0.08M KCl	0.02M MgCl <sub>2</sub> 6H <sub>2</sub> O	0.04M NaCaCo pH 6	45	12
PS1-28-2-1	Urchin	medium (1/2 drops)	0.08M NaCl	0.02M MgCl <sub>2</sub> 6H <sub>2</sub> O	0.04M NaCaCo pH 6	35	12
PS1-28-2-2	Urchin	good	0.08M KCl	0.02M MgCl <sub>2</sub> 6H <sub>2</sub> O	0.04M NaCaCo pH 6	45	12
PS1-28-2-3	n/a	bad (did not regrow)	0.08M NaCl	0.02M MgCl <sub>2</sub> 6H <sub>2</sub> O	0.04M NaCaCo pH 6	35	<b>2</b>
PS1-28-2-4	Urchin	medium (1/2 drops)	0.08M NaCl	0.02M MgCl <sub>2</sub> 6H <sub>2</sub> O	0.04M NaCaCo pH 6	35	<b>4</b>
PS1-28-2-5	n/a	bad (did not regrow)	0.08M NaCl	0.02M MgCl <sub>2</sub> 6H <sub>2</sub> O	0.04M NaCaCo pH 6	35	<b>1</b>



**Table A9.** Details on crystallization of NSD-TT.

Well	condition	crystal type	Time for initial hit	Synchrotron notes	space group	Salt 1	Salt 2	Additive 1	Buffer	Precipitant
PS1-17-1-A1	Helix 1-1	Large rectangle	>3 months	no confirmation		.05M KCl	0.1 M LiCl	12 mM Spermine	0.05M Mes pH 6.5	25% PEG400
PS1-17-1-B4	Helix 1-16	good parallelograms	>3 months	DNA; 2.38A	C 2 2 2 <sub>1</sub>	0.2M KCl		5mM Hexamine cobalt chloride	0.05M Mes pH 6.5	17% w/v PEG 4000
PS1-17-1-B5	Helix 1-17	gorgeous small xtals	>3 months	DNA; 3.16A	C 2 2 2 <sub>1</sub>	0.1M LiCl		0.01M Manganese(II) chloride tetrahydrate	0.05M Mes pH 6.5	17% w/v PEG 4000
PS1-17-1-A4	Helix 1-4	large splintered rectangle	>3 months	no confirmation		0.2M KCl		5mM Hexamine cobalt chloride	0.05M Mes pH 6.5	25% PEG400
PS1-17-1-H1	Helix 2-37	soccer balls	>3 months	DNA; 2.87A	P 63 2 2	.05M NaCl	.1M KCl	.01M CaCl <sub>2</sub> dihydrate	0.05M Bis-tris pH 7	25% PEG 2000
PS1-17-1-H6	Helix 2-42	Long rectangle	>3 months	DNA; 2.8 A	C 2 2 2 <sub>1</sub>	0.1M KCl		2mM Spermine	0.05M Bis-tris pH 7	15% PEG MME
PS1-17-2-D12	Natrix 1-48	diamonds	>3 months	no confirmation		0.2M Ammonium chloride		0.01M CaCl <sub>2</sub> dihydrate	0.05M Tris HCl pH 8.5	30% w/v PEG 4000

**Table A10.** Tracking of number of trials of biophysical data.

Description	Constructs	Assay	Trial 1	Trial1 Details	Trial 2
NSD Biophysical	9 base constructs, 3 TT constructs	Biophysical	6/22/22, 6/23/22, 7/1/22	Old CD, UV melt, TDS	
NSD Gel	9 base constructs, NSD-TT construct	Gel	6/23/22, 6/29/22	Done across 2 gels	
NSD-TT NMM	TT1L+NMM, TT+NMM	Biophysical	07/01/22	2 Eq of NMM, Old CD	7/8/22 (through nmm titration, same stocks though)
NSD-TT NMM	TT1L+NMM, TT+NMM, TT1R+NMM	Gel	06/29/22		
NSD-TT NMM titration	TT, TT1L - 2, 5, 10 eq NMM	Biophysical	7/7/22	Old CD, UV melt, TDS	
NSD-TT NMM titration	TT, TT1L - 2, 5, 10 eq NMM	Gel	7/13/22	Single gel - not publication quality	02/24/23
SLC&constructs	9 base constructs. Done in 2 batches	Biophysical	9/7/22, 9/13/22	New CD, TDS, <u><b>first batch melt done incorrectly</b></u>	04/05/06
SLC&constructs	9 base constructs, SLC and SLC1R with NMM	Gel	07/01/22	Gel looks good.	
SLC NMM titration	SLC - 1, 2, 5 eq NMM, SLC1R with 2eq NMM	Biophysical	10/27/22	New CD, TDS, melt data kinda messy	
SLC-1R NMM titration	SLC-1R - 1, 2, 5, 10 eq NMM, SLC with 10 Eq NMM	Biophysical	01/26/23	New CD, TDS, melt	
SLC NMM titration	SLC, SLC-1R with NMM	Gel	<b>not done yet</b>		
Overhang gel	TTRA/C/G/T, TTLA/C/G/T	Gel	44872	good gel, reorder slightly for next run	
Overhang biophysical	TTRA/C/T, TTLA/C/G	Biophysical	44852	New CD, TDS, melt	
PDGF-B1-6 Gel	B1-B6	Gel	05/25/22	good gel, only shows 6 variants though	
PDGF B4-13 Gel	B4-B13	Gel	07/12/22	questionable - no higher order species show up	
PDGF-B1-6 Biophysical	B1-B6	Biophysical	05/27/22	Old CD, <u><b>bad melt</b></u> , TDS	6/9/22