Table of Contents

Acknowledgements iv

Abstract vi

List of Figures viii

Chapter 1 1

1. RNA Polymerase I Transcription 2

2. Core Factor and Selectivity Factor I 4

3. Principles of DNA Recognition by Proteins 8

3.1 Protein-DNA Recognition Overview 8

3.2 Overview of Sequence-Based Recognition 9

3.3 The *Hind*III Restriction Endonuclease as an Example of Sequence-Based DNA Recognition 11

3.4 Overview of Structure-Based Recognition 12

3.5 E2 Protein as an Example of Structure-Based Recognition of Pre-Bent DNA 17

3.6 HU and IHF Proteins as Examples of Structure-Based Recognition of DNA after DNA Bending 18

3.7 Hox Proteins and RNA Pol I as Examples of Structure-Based Recognition of DNA by Cofactors 20

3.8 Chromatin as an Example of Structure-Based DNA Recognition 22

Disease Relevance 23

Thesis Goal 26

Chapter 2 32

Abstract 34

Introduction 35

Results 38

3.01 Effect of Point Mutations on CF DNA Binding 38

3.02 Correlation between Competition and DNA Structural Features 40

3.03 Identification of Novel CEs by In Vitro SELEX 42

3.04 SELEX enriched Structural Features 44

3.05 Two classes of SELEX sequences 49

3.06 Importance of Structural Features of Top Repeat 53

3.07 SELEX Validation 56

3.08 Identification of Novel CEs by In Vivo Selection 58

3.09 Enriched Structural Preferences In Vivo 60

3.10 Two classes of In Vivo Sequences 64
Discussion 68
Material and Methods 75
Chapter 3 88
1. Conclusions 89
2. In vitro Analysis of Single Base Pair CE Mutants 90
2.1 Direct Binding EMSA 90
2.2 In Vitro Transcription Assay 91
3. Analysis of Physical Characteristics of DNA 92
3.1 NMR 92
3.2 Circular Dichroism 93
3.3 Atomic Force Microscopy 93
3.4 Fluorescence resonance energy transfer(FRET) 95
3.5 Single Molecule Manipulation Techniques 96
4. Altered Specificity Assay 98
5. Coevolution Analysis 100
6. Mapping Genome Binding of CF with CHEC-seq 103
7. Alternative methods of SELEX 104
8. Summary 105
High Priority 105
Low Priority 107

Acknowledgements

I want to thank my advisor Bruce for everything he has done for me during my time here at Upstate. He took a chance on me when I reached out while still at undergrad, offering research experience and training where I previously had none. Bruce has been a very patient and kind mentor, and this has benefited me greatly, allowing me to gain confidence in myself and my abilities. He has worked hard to create a welcoming lab environment and I've always felt at home here because of that. I've greatly appreciated all the time he has put into me and my project and his hands on mentoring. He's always been available for me to ask questions, whether it be in person, via email, call, or text. Bruce also has a disturbingly good memory so any questions I've had, he has usually known the answer even if I've forgotten. Bruce has been a great mentor to have.

I'd also like to thank all the members of the Knutson lab whom I've had the pleasure of working alongside. We've made many great memories and gone on so many fun lab outings. I have you all to thank for the expanding of my palette.

I would like to give a special thanks to Wayne Decatur. Wayne has been instrumental in all the data analysis of my research. Anytime I ever had a problem I couldn't solve or needed help with anything bioinformatics, Wayne was always ready at the drop of a hat to help. He saved me countless hours manually analyzing my data, creating all manner of automated solutions for me.

I want to thank my family for their support. They've always encouraged me in everything I do want me to succeed and most importantly, be happy. I have enjoyed my

father's endless questions and confusion about what I do. "Are you still trying to make polymerases?"

Lastly, I'd like to thank the College of graduate studies and the Department of Biochemistry and Molecular Biology. All the staff, my cohort, faculty, and professors have made my achievements possible. I'd specifically like to thank my thesis defense and advisory committee for their time and support: Dr. Wenyi Feng and Dr. Gary Chan.

Abstract

Title: Specific Structural Features of the RNA Polymerase I Core Promoter Element Targeted by Core Factor

Author's Name: Nathan J. Munoff

Sponsor's Name: Bruce A. Knutson

In yeast, Core Factor (CF) is a critical and essential RNA Polymerase I (Pol I) transcription factor that plays fundamental roles in the transcription process by recruiting Pol I and opening Pol I promoter DNA before initiation. CF binds to a ~24 bp region in the rDNA promoter called the Core Element (CE) prior to Pol I recruitment. Pol I transcribes the rDNA gene into the 35S precursor rRNA (pre-rRNA) which serves both catalytic and structural roles in the ribosome. Up-regulation of Pol I transcription has been linked to a variety of human cancers, as increased protein production can facilitate the rapid growth of cancer cells. Thus, Pol I transcription is a promising target for therapeutic development. Previous studies from our lab suggest that CF and its human orthologue, Selectivity Factor 1 (SL1), use an evolutionarily conserved mechanism to target DNA, governed by the structural features of their respective promoters. Eukaryotic rDNA promoters also exhibit conserved structural features, such as intrinsic curvature and kinks but show a distinct lack of sequence conservation. These sequence independent structurally conserved features of rDNA promoters might explain how they are being recognized by CF and its orthologues. Our findings here revealed that CF is capable of tolerating mutations at some positions of the CE while mutation in the rigid "A" patch being particularly sensitive to mutations changing structural properties. Along

with conditional tolerance for sequence mutations, our results show that CF prefers a variety of structural features such as overall increased bendability and decreased curvature as well as specific profiles of bendability. Furthermore, we describe the preferences of CF for the parameters of helix twist, propeller twist, roll, and minor groove width.

List of Figures

Chapter 1

Figure 1. RNA Polymerase I Pre-Initiation Complex……..……...3

Figure 2. Domain Organization and Structure of Core Factor…….……...5

Figure 3. Sequence and Structural Conservation between Eukaryotic Species……..7

Figure 4. DNA Base Readout……...10

Figure 5. The Restriction Endonuclease *Hin*dIII…….….12

Figure 6. DNA Structural Parameters……...14

Figure 7. Purine and Pyrimidine Base Stacking Areas…….15

Figure 8. Pre-bent Physical Conformation of the E2 Binding Site of HPV-18……18

Figure 9. Significant Conformation Change of DNA upon IHF Binding…….19

Figure 10. Latent Specificity of Motif Preferences by Hox-Exd Complexes……...21

Chapter 2

Figure 1. Effects of DNA Structural Properties on CF Binding……...41

Figure 2. In Vitro SELEX Round Summary…….43

Figure 3. Bend-it Analysis of In Vitro SELEX Sequences Enriched by Yeast RNA Polymerase I CF…….....45

Figure 4. GB Shape Analysis of In Vitro SELEX Sequences Enriched by Yeast RNA Polymerase I CF…….48

Figure 5. Bend-it Analysis of In Vitro SELEX Sequences Found in Round 7 Classified into Rigid or Flexible Based on Bendability…….50

Figure 6. GB Shape Analysis of In Vitro SELEX Sequences Found in Round 7 Classified into Rigid or Flexible Based on Bendability…….52

Figure 7. Top Repeat Binding Profile Analysis……55

Figure 8. Validation of the In Vitro SELEX Round 7 Consensus Sequences Enriched by CF……...57

Figure 9. In Vivo Selection Summary……..59

Figure 10. Bendit Analysis of In Vivo Selection Sequences Enriched by Yeast RNA polymerase I
CF……...……61

Figure 11. GB shape analysis of in vivo selection sequences enriched by yeast RNA polymerase I
CF……...……63

Figure 12. Bendit Analysis of In Vivo Selection Sequences Enriched by Yeast RNA Polymerase I CF Grouped into Rigid and Flexible Categories by Bendability…….65

Figure 13. GB Shape Analysis of In Vivo Selection Sequences Enriched by Yeast RNA Polymerase I CF Grouped into Rigid and Flexible Categories by Bendability…….67

Chapter 3

Figure 1. Atomic Force Microscopy Setup in Solution……..……..95

Figure 2. DNA Pulley Setup……..……...98

Figure 3. Example of Altered Specificity Assay Using TBP……..100

Figure 4. Example of Coevolution between a Protein and DNA……102

Chapter 1

Introduction to RNA Polymerase I Transcription, Core Factor, and DNA Binding

Nathan J. Munoff

Department of Biochemistry and Molecular Biology, SUNY Upstate Medical University,

Syracuse, NY 13210.

1. RNA Polymerase I Transcription

Transcription of DNA to RNA is a crucial and essential operation for every living species and is carried out by one of three essential DNA dependent RNA polymerases in eukaryotes[1]. Each polymerase is responsible for producing different classes of RNA, with the help of different sets of transcription factors that recruit Pols to their respective promoters. Pol II is mainly responsible for producing mRNA and Pol III is mainly responsible for producing tRNA[1]. Pol I has a more specialized function than both Pol II and III as it is only responsible for producing ribosomal RNA (rRNA) with few exceptions[1]. In the budding yeast *Saccharomyces cerevisiae*, Pol I transcribes the rDNA gene into the polycistronic 35S precursor rRNA (pre-rRNA)[2-4]. This is then post-transcriptionally processed into the 18S, 5.8S and 25S rRNAs which serve both catalytic and structural roles in the ribosome[2-4]. The ribosome produces all proteins in the cell and thus the synthesis of rRNA is vital[5, 6]. In addition, Pol I transcription makes up more than 60% of total cellular transcription activity and as such requires high transcription rates[3]. As a result, a significant portion of a cell's energy is devoted to Pol I transcription. Logically, any dysregulation of Pol I transcription often has severe consequences as can be seen in many human cancers and ribosomopathies[7-14].

In order for Pol I transcription to take place, a variety of transcription factors are needed. In yeast, Pol I transcription requires four transcription factors, which include the Upstream activation factor (UAF), Core Factor (CF), Rrn3, and TATA-box binding protein (TBP)[3, 4, 15-23]. These transcription factors assemble to form the pre-initiation complex (PIC) at the rDNA promoter[24-26]. Across eukaryotes, the rDNA promoter has a structurally conserved, bipartite module comprising the Upstream Control Element and

the Core Promoter Element[24-29]. In yeast, these two elements are called the Upstream Activation Sequence (UAS) and the Core Element (CE) respectively[27-29]. UAS can be found from positions -150 to -60, and CE from positions -38 to -15, relative to the +1 start site of transcription[27-29]. UAF binds to the UAS, and CF binds to the CE. UAF helps to recruit CF, TBP helps to stimulate transcription, and Rrn3 induces and stabilizes the monomeric form of Pol I[3, 4, 15-23, 30-33]. Minimally, only Pol I, Rrn3, and CF are needed for basal transcription; however, transcription is stimulated 40-fold by assembly of the complete PIC[34-36].

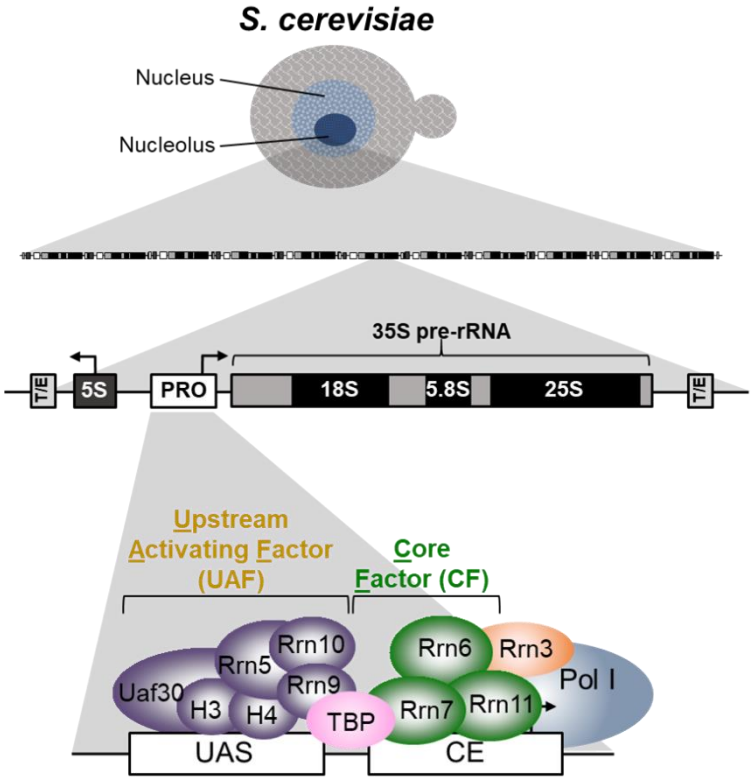

Figure 1. RNA Polymerase I Pre-Initiation Complex.

Schematic of the Pol I PIC. Depicted is a representation of a yeast nucleus and within it, nucleolus where an expanded view reveals the rDNA locus and rDNA repeats. An expanded view of an individual rDNA repeat is shown with a further expansion of the Pol I promoter region within. A cartoon representation of the bimodular nature of the promoter with the UAS and CE regions of the DNA are shown along with their respective transcription factors UAF and CF bound to these regions. Also shown are the positions and locations of TBP, Rrn3, and Pol I within the PIC. Adapted from the Knutson Lab.

2. Core Factor and Selectivity Factor I

The Pol I transcription factor CF is a heterotrimeric protein complex comprised of subunits Rrn6, Rrn7, and Rrn11, which have molecular weights of 102, 60, and 59 kD, respectively[34, 35, 37-39]. Rrn6 is composed of a WD40/β-propeller, helical bundle, and C-terminal domain (**Fig 2A, Blue**). Rrn7 has an N-terminal Zn ribbon, β-reader, β-linker, N-terminal cyclin domain, C-terminal cyclin domain separated by an insertion, and a C-terminal domain[16, 17, 21, 37, 40-42] (**Fig. 2A, Green**). Rrn11 contains an N-terminal domain followed by a tetratricopeptide repeat (TPR) domain and a C-terminal domain[17, 21, 37, 40-42] (**Fig 2A, Yellow**). In forming the CF complex, Rrn6 acts as a scaffold for Rrn7 and 11[17, 21, 37, 40-42]. The C-terminal helical bundle domain of Rrn6 forms extensive contacts with Rrn7, wrapping around the cyclin folds[17, 21, 37, 40-42]. The WD40/β-propeller domain of Rrn6 also forms extensive contacts with the TPR domain of Rrn11, and the N-terminal domain of Rrn11 is buried within with its deletion proving lethal[17, 21, 37, 40-42]. Rrn7 and Rrn11 are the two subunits that contact CE in the Pol I promoter, primarily by making contacts with the phosphate backbone[40, 42] (**Fig 2**). Overall, the structure of CF can be divided into two modules[17]. Module I contains Rrn11 and the WD40/β-propeller of Rrn6[17]. Module II contains the Rrn7 and the C-terminal domain of Rrn6[17]. Module II is mobile and is able to adopt multiple conformations and positions during transcription initiation[17]. The insertion within the C-terminal cyclin domain of Rrn7 connects the two modules, and is known as the flexible linker or hinge[17] (**Fig 2B)**.

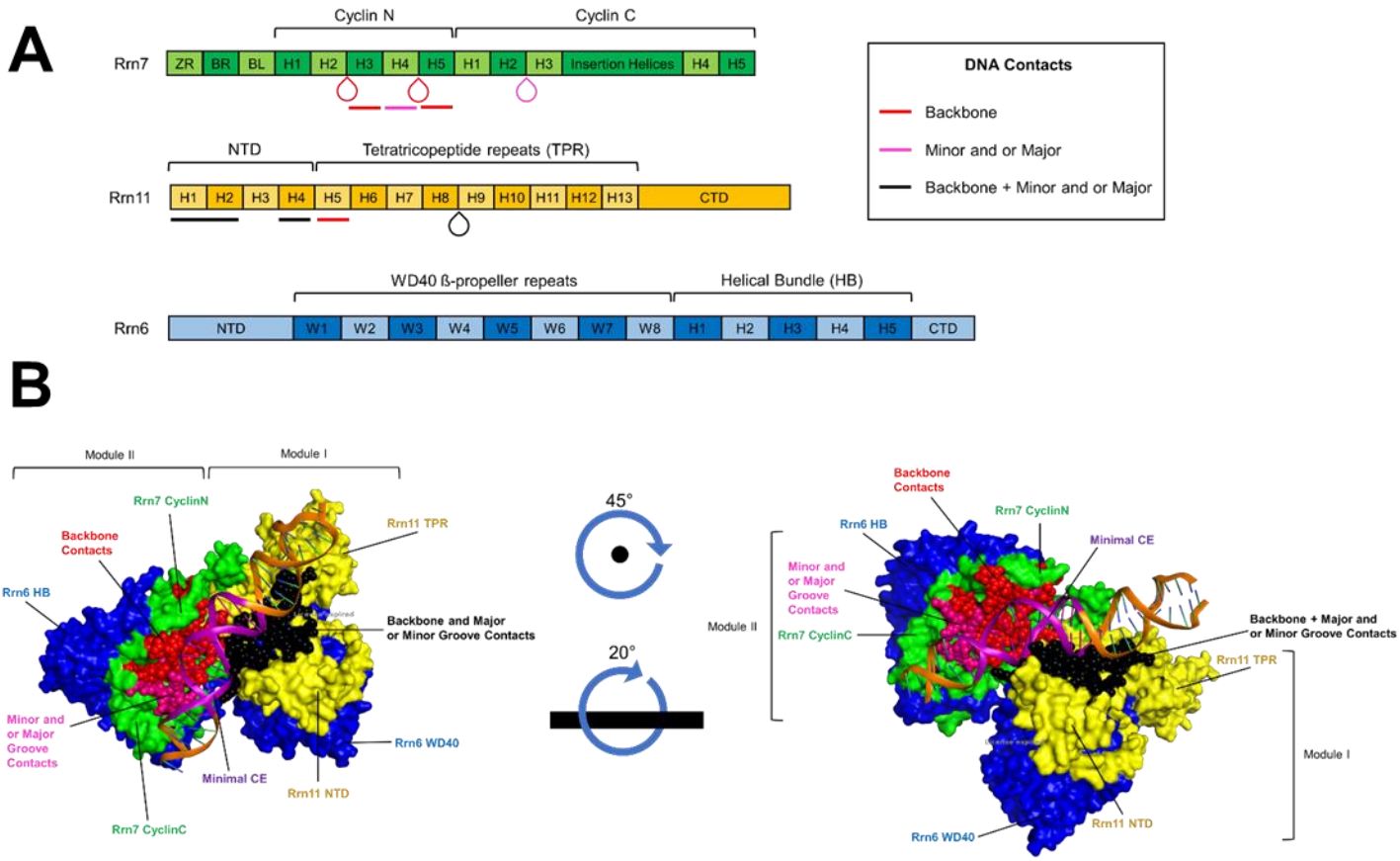

Figure 2. Domain Organization and Structure of Core Factor

A. Domain organization of the CF subunits. Red lines denote regions of Rrn7 and Rrn11 predicted to contact the backbone of DNA, pink lines contacts with the minor and or major groove, and black lines contacts with both the backbone and the minor and or major groove. Teardrops beneath Rrn7 and Rrn11 denote loops between respective helices. ZR, Z-reader; BR, B-reader; BL, B-linker; H, Helix; NTD, N-terminal; CTD, C-terminal domain; W, WD40 repeat. Adapted from Knutson et. Al. 2014. **B**. Structure of CF in complex with the CE. Subunit colors correspond to those in panel A. The minimal CE region, defined previously by Jackobel et al, 2019, denoted in purple. Residues in red are predicted to interact with the backbone of DNA, pink with the minor and or major groove, and black with both the backbone and the minor and or major groove. Structures modeled from PDB 5W5Y.

CF in yeast is orthologous to Selectivity Factor 1 (SL1) in humans[43]. Each subunit of CF has an ortholog in SL1: yeast Rrn6, Rrn7, and Rrn11 are orthologous to the human TBP associated factors (TAFs) TAF1C, TAF1B, and TAF1A, respectively[43]. In addition to the orthologous TAFs, SL1 contains three additional components: TBP (which is lacking in yeast CF), and the metazoan-specific subunits TAF1D and TAF12[44-46]. While CF and SL1 perform the same function of recruiting Pol I to the Pol I promoter, their constituents are poorly conserved. The three subunits of CF share only ~8-16% sequence identity with their orthologs in SL1[47] (**Fig 3**). Nevertheless, there are several domains in Rrn7 can be exchanged with the homologous domains in TAF1B to yield a wild-type growth phenotype in yeast[47]. Most notably, this applies to the DNA-binding domain of Rrn7[47]. The exchangeability of the DNA-binding domain of Rrn7 with the equivalent DNA-binding domain of TAF1B (despite their lack of sequence conservation and divergence of their target promoters) suggests that DNA binding occurs through a conserved, structural mechanism. Furthermore, yeast CF has been shown to bind the human Core Promoter Element, which functions in yeast in a positionally dependent manner[48]. Like the poor sequence conservation between CF and SL1, there is poor sequence conservation across eukaryotic rDNA promoters. However, they do share conserved structural features, such as a bend approximately 25 bp upstream of the transcription start site[49-54] (**Fig 3**). For example, DNA binding studies have shown that yeast CF has a preference for the GC-rich minor groove, a structural feature the orthologous TIF-IB protein in *Acanthamoeba castellani* also uses for DNA recognition and binding[48, 52]. These observations support the idea that CF is using a structure-based mechanism to recognize and bind to the CE. More importantly, these

findings highlight the similarities between CF and SL1 in binding their respective promoters which indicates that any discoveries made regarding CF will have relevant applications to SL1.

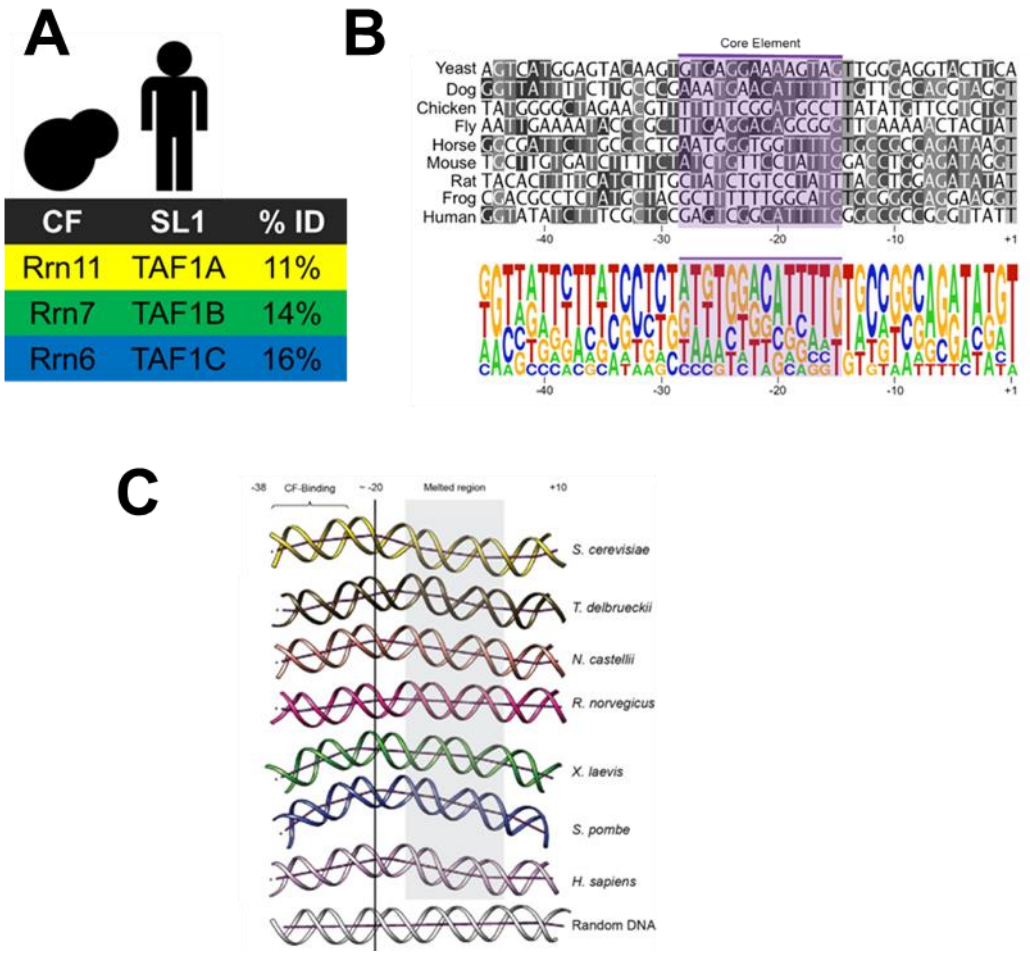

Figure 3. Sequence and Structural Conservation between Eukaryotic Species

A. Comparison of sequence identity between the subunits of yeast CF and their orthologues in human SL1. **B**. Sequence alignment of rDNA promoters from various eukaryotic species with the corresponding weblogo showing nucleotide frequency at each position featured below. Positions below both elements are marked relative to the TSS. Bases from the sequence alignment that match the most frequent nucleotide at that position are shaded with a darker to lighter shading indicating lower to higher total frequency respectively. Highlighted in purple is the minimal CE region needed for CF binding in vitro as defined by Jackobel et al 2019. **C**. Modeled DNA structures of aligned rDNA promoters from various eukaryotic species and random DNA with positions marked relative to TSS at the top. Also marked at the top is the region where CF binds and the region highlighted in grey is where the DNA melts at the start of transcription. The vertical line at -20 marks the general location of the conserved bend in the DNA structures. (Adapted from Knutson et al. 2012 and Engel et al 2017)

3. Principles of DNA Recognition by Proteins

3.1 Protein-DNA Recognition Overview

One aspect critical to all protein-DNA interactions is the specificity of DNA recognition. For example, a transcription factor such as CF must recognize and bind the correct promoter (i.e., CE in the rDNA promoter) to transcribe the correct DNA into RNA (i.e., the rRNA of the ribosome). Additional examples of the importance of the specificity of DNA recognition include DNA repair (wherein repair proteins must recognize and bind to damaged DNA) and DNA replication (wherein proteins must identify and bind to origins of replication). This concept also applies to chromatin regulation. Histone proteins need to bind DNA at the correct locations and recognize the state of the DNA (i.e., if it is methylated or acetylated), lest the incorrect region become hetero or euchromatin. Proteins simply cannot carry out specialized and targeted functions if they interact with the wrong locus or if the interactions are nonspecific and transient.

DNA recognition by DNA-binding proteins occurs via two main mechanisms. The first mechanism of DNA recognition is sequence-based, wherein a DNA-binding protein recognizes and binds to a specific order of nucleotides[55, 56]. The second mechanism of DNA recognition is structure-based, wherein a DNA-binding protein recognizes the physical conformation of the DNA, and binds utilizing the structural properties of the DNA[55, 56]. In both mechanisms, DNA-binding is affected not only by the DNA-binding protein and the sequence/structure of the DNA, but is also affected by other elements, such as cofactors, cooperativity of multiple DNA-binding proteins, and the state of chromatin. The complexity of these caveats makes determining to what extent

either or both mechanisms govern any given interaction a challenge. Typically, it is simpler to demonstrate sequence-based recognition of DNA by a DNA-binding protein than structure-based recognition. Structure-based recognition requires a suite of complementary methods to demonstrate. Consensus sequence analysis, coevolution analysis, amino acid-DNA binding preferences, predictive programs of structural features, and altered specificity assays together can demonstrate that a DNA-binding protein recognizes and binds to DNA by a structure-based mechanism.

3.2 Overview of Sequence-Based Recognition

One of the ways in which DNA-binding proteins recognize DNA is by a sequence-based mechanism. Sequence-based recognition (otherwise known as direct or base readout) refers to the interactions between amino acid side chains of the DNA-binding protein and the hydrogen bond acceptors and donors of the DNA base functional groups[55, 56]. Consequently, these interactions are dominated by hydrogen and water-mediated hydrogen bonding. Hydrophobic interactions occur as well, albeit to a lesser extent[56]. Hydrogen bonding can be further divided into bidentate, bifurcate, and monodentate classifications, listed in order of highest to lowest specificity[56]. Bidentate hydrogen bonds are two hydrogen bonds that have different acceptor and donor atoms[56]. Bidentate bonds can be made with a (i) single base, (ii) both bases in a pair, or (iii) two diagonal bases belonging to different base pairs and strands[56]. It is this range of different combinations of bonding that give bidentate bonds the highest specificity. Bifurcate hydrogen bonds are two hydrogen bonds that share an electron donor[56].

Hydrogen- and water-mediated hydrogen bonding can occur in both the major and minor groove. The major groove possesses more specificity due to the four unique

patterns of hydrogen bond donors and acceptors presented by the four bases[56] (**Fig 4**). In the minor groove there is no difference between in the pattern of donor and acceptors between an AT or TA base pair nor between a GC or CG base pair. Thus, there only exist two unique patterns[56] (**Fig 4**). The physical constraints of both the major and minor grooves can also affect whether an amino acid can even gain access to the functional groups of a DNA base.

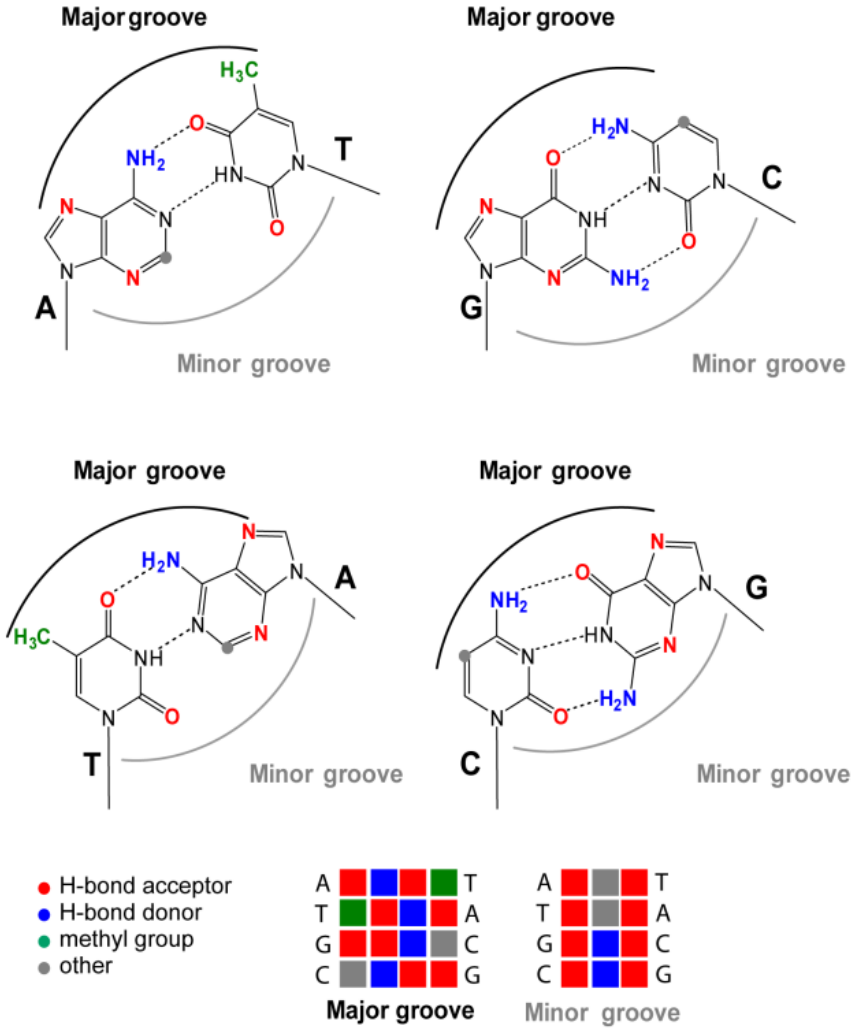

Figure 4. DNA Base Readout
Cartoon representation of the four different base pairs and their distinct patterns of functional groups in the minor and major grooves. Each pattern in the major groove is unique while only two unique patterns exist in the minor groove, one shared by AT and TA base pairs and the other by GC and CG base pairs. Adapted from Schneider et al, 2014.

3.3 The *Hind*III Restriction Endonuclease as an Example of Sequence-Based DNA Recognition

A great example of sequence-based recognition can be found in bacterial restriction endonucleases. One of the important roles of these endonucleases is to protect the bacteria against bacteriophage DNA integration into the genome[57-60]. To do so, restriction endonucleases recognize and cleave a specific sequence in foreign DNA (e.g., from bacteriophage). Restriction endonucleases are unable to recognize and cleave the same specific sequence in the host DNA, as these sites are protected via methylation markers[57-60]. A particularly well-known and studied restriction endonuclease is *Hind*III from *Haemophilus influenza*. *Hind*III recognizes and binds to the DNA sequence 5′-A/AGCTT-3′, cutting on the top and bottom strands after the first adenine (as indicated by the "/"). Binding between *Hind*III and AAGCTT is dominated by base-specific contacts[55] (**Fig 5**). According to the DNA-*Hind*III X-ray diffraction structure, all six bases in the major groove on both the top and bottom strands are read by the enzyme via mono- and bidentate hydrogen bonds[55] (**Fig 5**). Binding of *Hind*III to DNA is at first non-specific but upon recognition of the sequence at the cleavage site, a conformational change is triggered in both the protein and DNA, and subsequently cleavage occurs[55, 61].

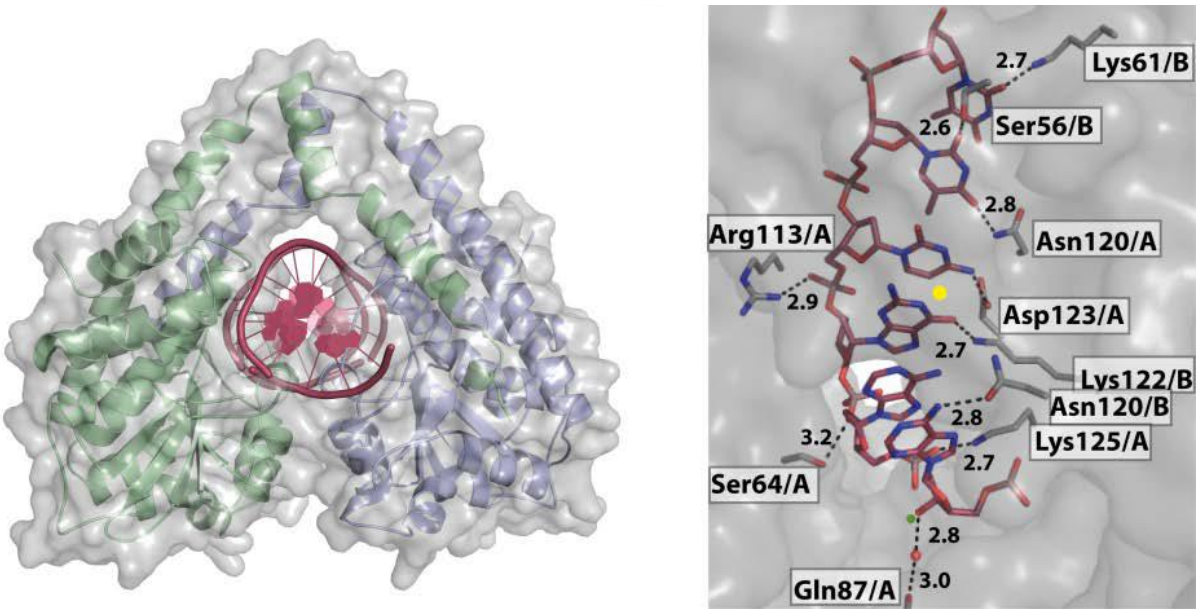

Figure 5. The Restriction Endonuclease *Hin*dIII
Structure of *Hin*dIII/DNA complex of the symmetrically bound dimer embedding the DNA. A enlarged view of the active site is shown with the palindromic target sequence AAGCTT. Direct and water mediated DNA contacts with specific residues are shown with dotted lines and their angstrom lengths listed. Water molecules and Mg2+ ions (required for catalysis), are shown as red and green spheres, respectively. The symmetry axis of the complex is shown as a yellow sphere and is located between bases GC/CG. Adapted from Schneider et al, 2014.

3.4 Overview of Structure-Based Recognition

The second mechanism proteins use to recognize DNA is structure-based recognition otherwise known as indirect readout[55, 56]. This mechanism is more difficult to define, and various definitions exist. In general, structure-based recognition can be defined as when a DNA-binding protein relies on the physical conformation and structural properties of DNA to recognize and bind to it, independent of a specific sequence. While it is true that sequence defines DNA structure, DNA structure is degenerate (in the same way codons are for amino acids). In other words, similar DNA structures and structural properties can be produced by more than one DNA sequence. It has also been defined as whenever sequence-based recognition is insufficient to account

for the specificity of an interaction. DNA-binding proteins recognize DNA structure primarily through interactions and contacts with phosphate and sugar groups of the backbone rather than the bases themselves. DNA structure can be measured in several key ways, such as minor groove width, roll, propeller twist, and helix twist (**Fig 6**). Minor groove width is measured as the inter-strand distance between phosphate atoms in the minor groove[62, 63] (**Fig 6**). Minor groove width has been found to be a specific structural feature recognized by DNA-binding proteins[64, 65]. For example, arginine (which has a positively-charged functional group) makes up 28% of all amino acid interactions in the negatively charged minor groove[64, 65]. Make the minor groove narrow as defined by <5Å compared to the 5.8Å in ideal B-DNA, and this percentage rises to 60%[64, 65]. This recognition is due to the enhanced electrostatic potential which is the most negative within the minor groove. This comes from an increased concentration of negative charge from the negatively charged phosphate groups of the backbone that are now closer together[64, 65]. Another way to measure the physical conformation of DNA, is by measuring roll. Roll is defined as the angle between two base pairs that opens toward the minor groove[62, 63] (**Fig 6**). Propeller twist is defined as the degree to which two bases in a base pair are out of plane with one another looking down the longitudinal axis[62, 63] (**Fig 6**). Helix twist is the degree to which a double-stranded DNA molecule is twisted counterclockwise from being perfectly straight[62, 63] (**Fig 6**).

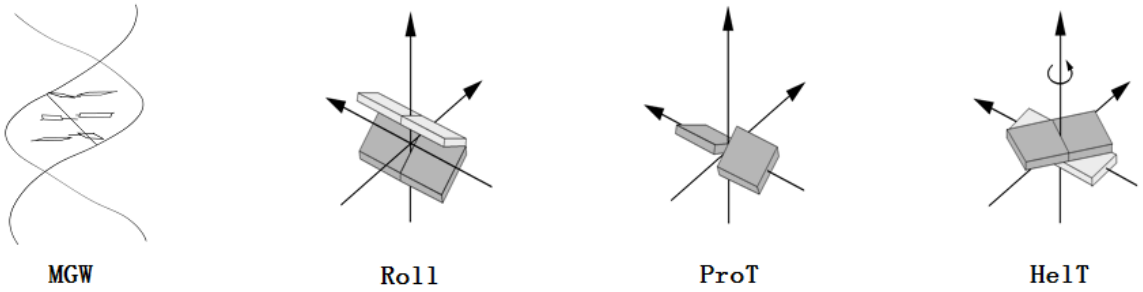

Figure 6. DNA Structural Parameters.

Cartoon representations of structural parameters of DNA base pairs predicted by DNAshape. Minor Groove Width (MGW), Roll, Propeller Twist (ProT), and Helical Twist (HelT) are shown. Adapted from Rohs et al, 2013.

Another important factor in determining DNA shape and structural properties are the base stacking interactions between one base pair and its neighboring base pair (also known as a base step). Base stacking plays a significant role in determining how bendable and or curved a DNA sequence is[55, 56, 66-68]. Bendability refers to the ability of a DNA sequence to be deformed. A DNA sequence is considered to be highly bendable if it is readily deformed or takes little energy to deviate from the ideal B form DNA. Base stacking interactions are primarily electrostatic and there is a direct correlation between the stacking area of a base step and stiffness[55] (**Fig 7**). The larger the stacking area (or overlap between base pairs) in a base step, the more interactions and stability gained and the stiffer the base step[55] (**Fig 7**). AT rich sequences are more bendable than GC rich sequences[66, 69-72]; however, the context of neighboring base pairs and base steps play a significant role in the bendability of any given sequence. For example, poly deoxyriboadenosine (dA) motifs are associated with very low bendability, but poly dAT:dAT motifs have higher bendability that poly dA motifs; CpG is considered one of the most bendable dinucleotides[66, 69-72]. As for curvature, it is defined as gradual and consistent bending over the course of a stretch of nucleotides. Context also

matters for curvature, but in general, A and AT tracks are more curved and rigid than other sequences[66, 69-72]. It is important to note that for these tracks and any other curved sequences, the base steps that are producing said curvature must be in phase with each other, otherwise the direction of their bends will cancel each other out and produce a relatively straight DNA molecule[70, 71, 73, 74]. Classifying the structure of sequences is further complicated by the fact that sequences can be bent and rigid, curved and rigid, curved and bendable, and bent and bendable. It should be noted that bendability is separate from a sequence that is bent. A sequence that is bent is not necessarily bendable and may be stiff while relatively locked into a single conformation.

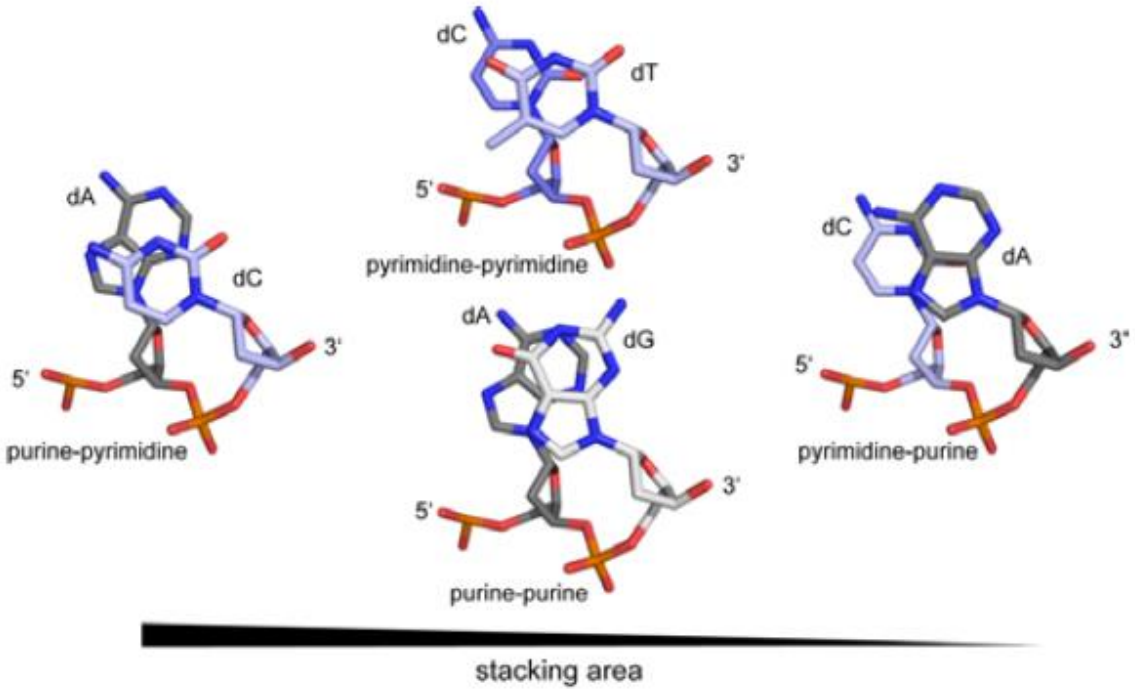

Figure 7. Purine and Pyrimidine Base Stacking Areas
The relative stacking areas of the four combinations of purine and pyrimidine base pairs ordered from highest to lowest. Adapted from Schneider et al 2014.

Then there is the consideration of when the bend is occurring in the context of protein binding. DNA can be recognized in its pre-bent form, already close to the final conformation it will occupy upon binding, with little bending induced by the protein itself. Alternatively, the initial conformation of the DNA can be different than the conformation adapted when bound by DNA-binding proteins. A DNA-binding protein can recognize DNA based on the intrinsic bendability of the DNA sequence and induce significant bending upon binding to the DNA. There are thermodynamic trade-offs to each mechanism with low entropic costs of binding corresponding to high enthalpic costs, and vice versa. This is because for a binding event to take place, the change in free energy must be negative. Upon protein binding, any loss of conformational variability in addition to vibration of trapped water molecules incurs an entropic cost and the replacement of water surrounding the DNA by protein or desolvation incurs an enthalpic cost[49, 55, 75, 76]. Loss of conformational variability results in an entropic cost because disorder has decreased. The DNA has been locked into a more limited range of conformations and now more organized. Desolvation incurs an enthalpic cost because displacement of water by the protein requires the disruption and breaking of bonds between the water and DNA, which requires energy[49, 55, 75, 76]. For example, if a DNA sequence's conformation isn't significantly altered upon binding to protein, this interaction would be advantageous for enthalpy but have some entropic cost. The loss of conformation variability and freedom leads to less disorder or entropy, which is disadvantageous as thermodynamic systems are biased towards increasing entropy (i.e., disorder). Conversely, since the DNA was already mostly in the desired conformation of the protein, conformational variability is relatively unrestrained, and the entropic cost

remains relatively low. Also, the enthalpic cost is favorable as the formation of new bonds between the DNA and protein release energy. The complex transitions from a higher energy state to a lower one and is not bending and destacking the DNA, which would require energy to be put into breaking and disrupting those bonds[49, 55, 75, 76]. If a protein is significantly bending the DNA from its free conformation, the entropic cost becomes very high and as a result, the change in enthalpy must compensate with more bonds formed between the DNA and protein than the previous example. Ultimately, enthalpy and entropy must be balanced in such a way that the free energy change isn't so favorable as to permanently fix the protein to the DNA but also not so weak as to have no specificity for the recognition site over non-specific DNA.

3.5 E2 Protein as an Example of Structure-Based Recognition of Pre-Bent DNA

An example of a protein-DNA interaction that requires pre-bent DNA is the binding of the E2 protein to the HPV-18 E2 binding site[56, 77] (**Fig 8**). The E2 protein is encoded by human papillomavirus, a DNA virus. E2 is a transcription factor that binds to several sites in the papillomavirus genome[56, 77]. The free target DNA has an A tract AATT in the central region of the helix that once bound, largely resembles its unbound conformation (**Fig 8**). While in the bound state, a series of direct contacts are made with the bases of the conserved regions outside the central A tract, but not all these contacts are physically possible if the DNA was not bent[56, 77]. It is ultimately the physical pre-bent structure of the binding site that makes it possible for the E2 protein to recognize it and further allow for base specific contacts to be formed.

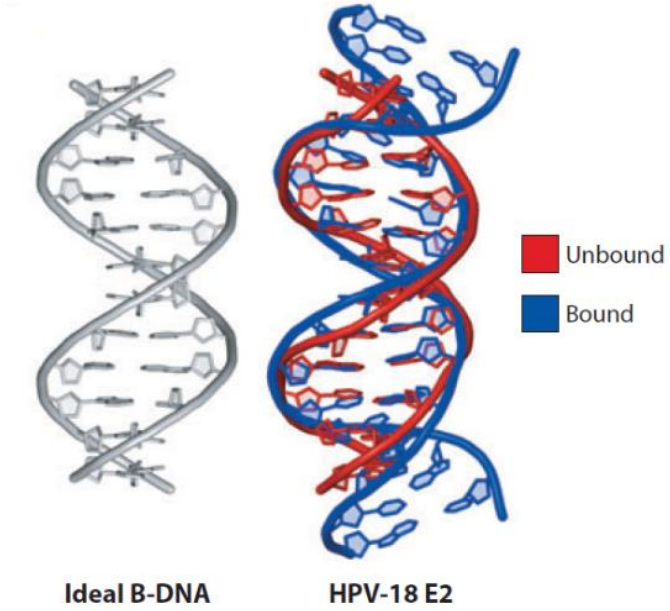

Figure 8. Pre-bent Physical Conformation of the E2 Binding Site of HPV-18

Unbound (red) and bound (blue) conformations of the HPV-18 E2 site compared to ideal B-DNA (grey) showing the DNA backbone and base pairs. Both bound and unbound structures display very similar conformations with both having a narrowed minor groove in the center relative to the ideal B-DNA. Adapted from Rohs et al, 2010.

3.6 HU and IHF Proteins as Examples of Structure-Based Recognition of DNA after DNA Bending

Two similar examples of DNA-binding proteins changing the structure of the DNA after being bound, are the binding of the bacterial chromosomal proteins Histone Like Protein from *E. coli* strain U93 (HU) and Integration Host Factor (IHF) to their target DNA sequences[55, 78] (**Fig 9**). HU and IHF bind to DNA to compact chromosomes and maintain supercoiling while also regulating DNA damage recognition, transcription, and DNA replication[55, 78]. HU and IHF bind more to A-tract, AT rich, pre-bent, and nicked or kinked sequences, without a preference for a specific DNA sequence[55, 78]. HU and IHF both recognize an increased helical twist at a critical and

bendable pyrimidine-purine base step in their respective binding sites[55, 78]. Upon binding, structures of these protein-DNA complexes reveal both proteins bend the DNA around 105-180°[55, 78] (**Fig 9**). Base pairs within a 9-nucleotide distance of binding also lose stacking interactions via two conserved proline residues intercalating between bases in the DNA and the DNA is kinked twice[55, 78] (**Fig 9**). IHF contacts the phosphate backbone and minor groove and makes no contacts in the major groove[55, 78] (**Fig 9**). Ultimately, while both HU and IHF recognize an existing structure of the DNA, they rely heavily on the bendability of the binding site to accommodate the severe 105-180° bending.

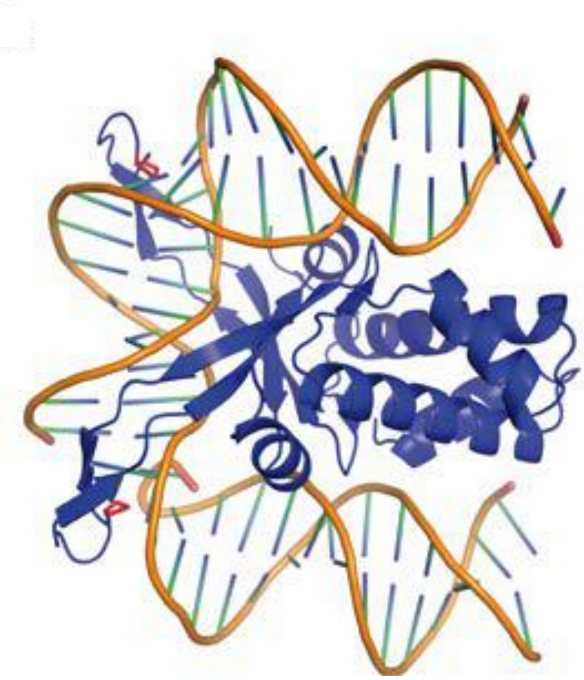

Figure 9. Significant Conformation Change of DNA upon IHF Binding
Structure of IHF-DNA complex with the DNA bent between 105°-180°. Two conserved proline residues on the B arms are highlighted in red and intercalate the DNA, destacking the base pairs and kinking the DNA twice. No contacts with the DNA major grove are made in this complex. Adapted from Schneider et al, 2014.

3.7 Hox Proteins and RNA Pol I as Examples of Structure-Based Recognition of DNA by Cofactors

Indirect and or direct base readout are not enough to explain the specificity of a protein-DNA interaction. Often, there are other cofactors and cooperative effects as well as chromatin state affecting specificity. However, why doesn't sequence and or structure recognition alone provide enough specificity to explain all protein-DNA interactions? One reason is that some DNA-binding proteins can tolerate some degree of variability in sequences that they can bind (e.g., transcription factors; TFs); a high degree of specificity is required to ensure that the TF binds to the correct DNA binding site. In addition, the typical binding site of DNA-binding proteins is approximately 6-10 bp. Consequently, there are potentially hundreds or thousands of identical sequences throughout the genome[79]. For a DNA-binding protein to bind to the correct site(s) and not the other potential binding sites, it requires specificity from other cofactors. It has also been proposed that there is an evolutionary pressure to allow for more complex systems of regulation than in turn allow for more complex organisms. There are also families of proteins that perform similar functions, and as a result, these proteins bind to very similar sequences. The Homeobox (Hox) family of proteins are an example[79, 80]. There are eight Hox orthologues in *Drosophila melanogaster* that control posterior-anterior patterning[79, 80]. The genes encoding these proteins are both located and expressed collinearly[79-81]. Each Hox protein has a distinct *in vivo* function and regulates the development of a specific region of the embryo[79-81]. The Hox protein distinguish one site from another by forming a heterodimer with the cofactor Extradenticle (Exd)[79, 80] (**Fig 10**). The binding of Exd to a specific Hox protein changes the preferred DNA sequence of the Hox protein by expanding the DNA-binding interface of the Hox protein

to include that of Exd. The complex then binds DNA specifically at the interface between them [79, 80] (**Fig 10**). In other words, heterodimerization with Exd changes the preferred sequence from what the Hox proteins would normally bind to on their own.

Figure 10. Latent Specificity of Motif Preferences by Hox-Exd Complexes. Aligned Hox-Exd binding motifs of different Hox proteins in complex with Exd displayed below an example Hox-Exd DNA bound structure. A hexameric "core" motif is underlined and both major and minor groove contacts are made here with Exd and Hox proteins. Flanking the core motif are the Exd or Hox regions where these bases contact just one protein or the other. Positions highlighted in yellow require Y across all complexes and underlined bases are contacted by Asn51 of the a3 recognition helices of the Exd and Hox homeodomains in the major groove, respectively. Positions highlighted in red can be N. The IUPAC symbols "W" denotes A or T, "R" denotes A or G, "Y" denotes C or T, "D" denotes not C, and N denotes any base. (Adapted from Slattery et al, 2011).

Another excellent example of cooperativity at work are the protein complexes that form around transcription start sites, such as RNA Polymerase I (Pol I). In order for Pol I transcription to start in vivo the Pol I pre-initiation complex must be formed[25]. In yeast, this complex consists of Upstream Activation Factor (UAF), TATA Box Protein (TBP), Core Factor (CF), Rrn3, and Pol I[48]. However, *in vitro*, only CF, Rrn3 and Pol I are required for a basal level of transcription[34]. When all proteins are present, they cooperatively enhance the binding, activation, and stabilization of the entire complex, ultimately leading to enhanced Pol I recruitment and transcription[34].

3.8 Chromatin as an Example of Structure-Based DNA Recognition

Another factor that can affect specificity and DNA recognition by proteins is chromatin. Chromatin exists as either euchromatin or heterochromatin. Euchromatin refers to chromatin that is unpacked or in an "open" state, which allows genes to be accessed by the transcription machinery and to be transcribed. Heterochromatin refers to chromatin that is densely packed or in a "closed" state, which prevents genes from being accessed by the transcription machinery and preventing transcription. Therefore, the state of chromatin dictates the accessibility of DNA-binding proteins to DNA. The state of chromatin is affected by two possible post-translational histone modifications. Euchromatin formation is induced by acetylation of histones by Histone Acetyltransferases[82], whereas heterochromatin formation is induced by methylation of histone by Histone Methyltransferases. Acetylation and methylation not only regulate transcription, but also play important roles in DNA repair[82]. In yeast, the histone H4K20 is methylated at sites of DNA damage[82]. H4K20 methylation recruits Crumbs

Cell Polarity Complex Component 2 (Crb2), a checkpoint protein that induces cell cycle arrest between the G2 and M phases of the cell cycle[82, 83]. Crb2 recognizes the H4K20 methylation through its double Tudor domains[82]. This methylation in concert with phosphorylation of H2A.X recognized by Crb2's Breast cancer associated 1 c-terminal (BRCT) domain creates the specificity for Crb2 to localize to DNA damage[82, 84]. In yeast, H3K56 acetylation is a marker of newly-synthesized histones during S phase[82]. In undamaged DNA, this marker is removed during G2; when DNA is damaged, deacetylases are downregulated, preventing deacetylation and allowing the H3K56 acetylation mark to persist. H3K56ac establishes euchromatin and keeps the DNA in an "open" state that is accessible by the DNA repair machinery[82]. The influence of chromatin on DNA recognition is not limited acetylation or methylation, and encompasses more processes than DNA repair, transcription, and compaction.

Disease Relevance

RNA Pol I is required for cell growth, functional activity and development. Pol I is directly responsible for synthesizing ribosomal RNA (rRNA), which makes up the structural and catalytic components of the ribosome[2-4]. Ribosome production and protein synthesis rates vary from cell to cell, tissue to tissue, and across stages of development, requiring intricate metabolic regulation. Mutation or dysregulation of Pol I often leads to disease. Up-regulation of RNA polymerase I (Pol I) transcription has been linked to a variety of human cancers, as increased protein production can facilitate the rapid growth of cancer cells[12]. c-MYC is a transcription factor that positively regulates Pol I transcription (as well as general metabolism), and is over-expressed in around 50% of all cancers [12, 85, 86]. Over-expression of c-MYC leads to increased Pol I activity,

which drives tumorigeneses and is often associated with poor prognosis[12, 87]. c-MYC can increase Pol I transcription by several mechanisms, one of which involves c-MYC binding to the Pol I transcription factor SL1 [12, 88]. This binding helps stabilize the SL1/UBF complex and increases Pol I transcription [12, 88].

The link between Pol I upregulation and disease has been known for some time and as a result, many efforts have been made at regulating and targeting Pol I transcription. One of the most promising Pol I anticancer compounds is CX-5461. CX-5461 has shown great specificity towards cancer cells where c-MYC is upregulated and p53 is not mutated [8]. The hypothesis is that c-MYC up-regulation leads to up-regulation of rRNA synthesis by Pol I, which in turn leads to a concomitant increase in ribosomal proteins synthesis [8]. Inhibiting rRNA synthesis in cancer cells with CX-5461 would create an excess of free ribosomal proteins [8]. Accumulation of free ribosomal proteins activates p53 and the nucleolar stress pathway, which collectively induce apoptosis [8]. Cancer cell death induced by the anti-Pol I inhibitor CX-5461 has made Pol I an attractive and emerging anti-cancer target. Despite its promise and universal sensitivity of hematologic malignancies to it, CX-5461 does not seem to be as nearly as effective in solid tumors [11, 12]. In addition, prolonged dosing in mice with Eu-MYC lymphomas eventually led to resistance [8, 12]. How CX-5461 inhibits Pol I transcription in cancer cells is currently unknown. Our lab has shown that CX-5461 binds to GC-rich DNA within the Pol I promoter *in vitro*, and Drygin *et al.* demonstrated that CX-5461 blocks SL1 binding and thereby, blocks Pol I transcription [11, 12, 48, 89]. Mars *et al.* proposed the conflicting hypothesis that CX-5461 specifically inhibits aberrant Pol I activity in

cancer cells by blocking promoter release of the Pol I-Rrn3 complex, trapping it in the pre-initiation stage [11, 12, 48, 89, 90].

However, dysregulation of Pol I is not limited to cancer, and is also implicated in other diseases, such as ribosomopathies (1). Ribosomopathies are caused by mutations in Pol I itself, factors that affect Pol I transcription, ribosomal proteins, and translation initiation and elongation factors [12]. Treacher Collins Syndrome (TCS) is a disease that occurs in 1/50,000 live births and is characterized by abnormal craniofacial development in early embryogenesis. This results in deformities such as cleft palate and down slanting palpebral fissures, as well as progressive hearing loss[12, 91, 92]. TCS can be caused by mutations in two subunits of Pol I, or in *TCOF1,* the gene encoding the protein Treacle which interacts with Pol I and other factors to promote transcription [9, 10, 12]. There are currently no known therapies for any ribosomopathies [12].

Critically, it is not yet known how CF is recognizing and binding to CE. To eventually develop more precise and effective Pol I inhibitors, understanding how CF recognizes CE is paramount. Determining the mechanism of CF's interaction with CE will allow for more future potential therapeutic opportunities to inhibit upregulated Pol I transcription in cancer and overcome shortcomings of drugs like CX-5461. Our lab has shown yeast CF binds to the human promoter CPE and that the CPE functions in yeast albeit in a positionally dependent manner [48]. In addition, rDNA promoter sequences share very little sequence conservation across a wide variety of species yet share structurally conserved features such as intrinsic curvature and kinks [5, 51, 93]. All this data together increases the possibility that any future findings regarding the CF and CE interaction will be applicable to SL1 and CPE.

Thesis Goal

Upregulation of RNA polymerase I (Pol I) transcription has been linked to a variety of cancers, making Pol I an attractive and emerging anticancer target. In addition, Pol I dysregulation is linked to ribosomopathies. One possible target to combat cancer is the Pol I transcription factor CF. However, it is not yet known how CF is recognizing and binding to the CE and whether it is more structurally or sequence based. Previous studies from our lab suggest that CF and its human orthologue, Selectivity Factor 1 (SL1), use an evolutionarily conserved mechanism to target DNA, governed by the structural features of their respective promoters (such as the GC minor groove, a unique surface rarely targeted by DNA-binding proteins). One of the most promising Pol I anticancer compounds is CX-5461, which binds GC rich DNA. However, CX-5461 is not universally effective and without flaws. By understanding the interaction between CF and CE, these findings could be used to develop better and more precise second-generation Pol I transcription inhibitors in the future. To this end, the specific focus of my project was to determine the structural features of CE that CF utilizes to recognize and bind to DNA using a structural mechanism.

1. Roeder, R.G. and W.J. Rutter, *Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms.* Nature, 1969. **224**(5216): p. 234-7.
2. Goodfellow, S.J. and J.C.B.M. Zomerdijk, *Basic Mechanisms in RNA Polymerase I Transcription of the Ribosomal RNA Genes*, in *Epigenetics: Development and Disease*. 2012, Springer Netherlands: Dordrecht. p. 211-236.
3. Reeder, R.H., *Regulation of RNA polymerase I transcription in yeast and vertebrates.* Progress in Nucleic Acid Research and Molecular Biology, 1999. **62**: p. 293-327.
4. Schneider, D.A., *RNA polymerase I activity is regulated at multiple steps in the transcription cycle: recent insights into factors that influence transcription elongation.* Gene, 2012. **493**(2): p. 176-184.
5. Moss, T., et al., *A housekeeper with power of attorney: the rRNA genes in ribosome biogenesis.* Cell Mol Life Sci, 2007. **64**(1): p. 29-49.
6. Woolford, J.L., Jr. and S.J. Baserga, *Ribosome biogenesis in the yeast Saccharomyces cerevisiae.* Genetics, 2013. **195**(3): p. 643-81.
7. Bywater, M.J., et al., *Dysregulation of the basal RNA polymerase transcription apparatus in cancer.* Nat Rev Cancer, 2013. **13**(5): p. 299-314.
8. Bywater, M.J., et al., *Inhibition of RNA polymerase I as a therapeutic strategy to promote cancer-specific activation of p53.* Cancer Cell, 2012. **22**(1): p. 51-65.
9. Dauwerse, J.G., et al., *Mutations in genes encoding subunits of RNA polymerases I and III cause Treacher Collins syndrome.* Nature Genetics, 2011. **43**(1): p. 20-22.
10. Dixon, M.J., et al., *The gene for Treacher Collins syndrome maps to the long arm of chromosome 5.* American Journal of Human Genetics, 1991. **49**(1): p. 17-22.
11. Drygin, D., et al., *Targeting RNA polymerase I with an oral small molecule CX-5461 inhibits ribosomal RNA synthesis and solid tumor growth.* Cancer Research, 2011. **71**(4): p. 1418-1430.
12. Hannan, K.M., et al., *Dysregulation of RNA polymerase I transcription during disease.* Biochimica et biophysica acta. Gene regulatory mechanisms, 2013. **1829**(3-4): p. 342-360.
13. Walker-Kopp, N., et al., *Treacher Collins syndrome mutations in Saccharomyces cerevisiae destabilize RNA polymerase I and III complex integrity.* Hum Mol Genet, 2017. **26**(21): p. 4290-4300.
14. Williamson, D., et al., *Nascent pre-rRNA overexpression correlates with an adverse prognosis in alveolar rhabdomyosarcoma.* Genes Chromosomes Cancer, 2006. **45**(9): p. 839-45.
15. Aprikian, P., B. Moorefield, and R.H. Reeder, *New Model for the Yeast RNA Polymerase I Transcription Cycle.* Molecular and Cellular Biology, 2001. **21**(15): p. 4847-4855.
16. Blattner, C., et al., *Molecular basis of Rrn3-regulated RNA polymerase I initiation and cell growth.* Genes Dev, 2011. **25**(19): p. 2093-105.
17. Engel, C., et al., *Structural Basis of RNA Polymerase I Transcription Initiation.* Cell, 2017. **169**(1): p. 120-131.e22.
18. Jackobel, A.J., et al., *Breaking the mold: structures of the RNA polymerase I transcription complex reveal a new path for initiation.* Transcription, 2018. **9**(4): p. 255-261.
19. Moorefield, B., E.A. Greene, and R.H. Reeder, *RNA polymerase I transcription factor Rrn3 is functionally conserved between yeast and human.* Proc Natl Acad Sci U S A, 2000. **97**(9): p. 4724-9.
20. Peyroche, G., et al., *The recruitment of RNA polymerase I on rDNA is mediated by the interaction of the A43 subunit with Rrn3.* EMBO J, 2000. **19**(20): p. 5473-82.

21. Sadian, Y., et al., *Structural insights into transcription initiation by yeast RNA polymerase I.* The EMBO journal, 2017. **36**(18): p. 2698-2709.
22. Siddiqi, I., et al., *Role of TATA binding protein (TBP) in yeast ribosomal dna transcription by RNA polymerase I: defects in the dual functions of transcription factor UAF cannot be suppressed by TBP.* Mol Cell Biol, 2001. **21**(7): p. 2292-7.
23. Steffan, J.S., et al., *The role of TBP in rDNA transcription by RNA polymerase I in Saccharomyces cerevisiae: TBP is required for upstream activation factor-dependent recruitment of core factor.* Genes Dev, 1996. **10**(20): p. 2551-63.
24. Hahn, S., *Structure and mechanism of the RNA polymerase II transcription machinery.* Nat Struct Mol Biol, 2004. **11**(5): p. 394-403.
25. Kostrewa, D., et al., *RNA polymerase II-TFIIB structure and mechanism of transcription initiation.* Nature, 2009. **462**(7271): p. 323-30.
26. Liu, X., et al., *Structure of an RNA polymerase II-TFIIB complex and the transcription initiation mechanism.* Science, 2010. **327**(5962): p. 206-9.
27. Choe, S.Y., M.C. Schultz, and R.H. Reeder, *In vitro definition of the yeast RNA polymerase I promoter.* Nucleic Acids Res, 1992. **20**(2): p. 279-85.
28. Kulkens, T., et al., *The yeast RNA polymerase I promoter: ribosomal DNA sequences involved in transcription initiation and complex formation in vitro.* Nucleic Acids Res, 1991. **19**(19): p. 5363-70.
29. Musters, W., et al., *Linker scanning of the yeast RNA polymerase I promoter.* Nucleic Acids Res, 1989. **17**(23): p. 9661-78.
30. Bischler, N., et al., *Localization of the yeast RNA polymerase I-specific subunits.* EMBO J, 2002. **21**(15): p. 4136-44.
31. Engel, C., et al., *RNA polymerase I structure and transcription regulation.* Nature, 2013. **502**(7473): p. 650-5.
32. Fernandez-Tornero, C., et al., *Crystal structure of the 14-subunit RNA polymerase I.* Nature, 2013. **502**(7473): p. 644-9.
33. Milkereit, P., P. Schultz, and H. Tschochner, *Resolution of RNA polymerase I into dimers and monomers and their function in transcription.* Biol Chem, 1997. **378**(12): p. 1433-43.
34. Bedwell, G.J., et al., *Efficient transcription by RNA polymerase I using recombinant core factor.* Gene, 2012. **492**(1): p. 94-99.
35. Keener, J., et al., *Reconstitution of yeast RNA polymerase I transcription in vitro from purified components. TATA-binding protein is not required for basal transcription.* J Biol Chem, 1998. **273**(50): p. 33795-802.
36. Pilsl, M., et al., *Structure of the initiation-competent RNA polymerase I and its implication for transcription.* Nat Commun, 2016. **7**: p. 12126.
37. Knutson, B.A., et al., *Architecture of the Saccharomyces cerevisiae RNA polymerase I Core Factor complex.* Nat Struct Mol Biol, 2014. **21**(9): p. 810-6.
38. Lalo, D., et al., *RRN11 encodes the third subunit of the complex containing Rrn6p and Rrn7p that is essential for the initiation of rDNA transcription by yeast RNA polymerase I.* J Biol Chem, 1996. **271**(35): p. 21062-7.
39. Lin, C.W., et al., *A novel 66-kilodalton protein complexes with Rrn6, Rrn7, and TATA-binding protein to promote polymerase I transcription initiation in Saccharomyces cerevisiae.* Mol Cell Biol, 1996. **16**(11): p. 6436-43.
40. Han, Y., et al., *Structural mechanism of ATP-independent transcription initiation by RNA polymerase I.* eLife, 2017. **6**: p. e27414.
41. Pilsl, M. and C. Engel, *Structural basis of RNA polymerase I pre-initiation complex formation and promoter melting.* Nat Commun, 2020. **11**(1): p. 1206.

42. Sadian, Y., et al., *Molecular insight into RNA polymerase I promoter recognition and promoter melting.* Nature Communications, 2019. **10**(1): p. 1-13.
43. Boukhgalter, B., et al., *Characterization of a fission yeast subunit of an RNA polymerase I essential transcription initiation factor, SpRrn7h/TAF(I)68, that bridges yeast and mammals: association with SpRrn11h and the core ribosomal RNA gene promoter.* Gene, 2002. **291**(1-2): p. 187-201.
44. Bell, S.P., H.M. Jantzen, and R. Tjian, *Assembly of alternative multiprotein complexes directs rRNA promoter selectivity.* Genes Dev, 1990. **4**(6): p. 943-54.
45. Denissov, S., et al., *Identification of novel functional TBP-binding sites and general factor repertoires.* EMBO J, 2007. **26**(4): p. 944-54.
46. Gorski, J.J., et al., *A novel TBP-associated factor of SL1 functions in RNA polymerase I transcription.* EMBO J, 2007. **26**(6): p. 1560-8.
47. Knutson, B.A. and S. Hahn, *Yeast Rrn7 and human TAF1B are TFIIB-related RNA polymerase I general transcription factors.* Science, 2011. **333**(6049): p. 1637-40.
48. Jackobel, A.J., et al., *DNA binding preferences of S. cerevisiae RNA polymerase I Core Factor reveal a preference for the GC-minor groove and a conserved binding mechanism.* Biochimica Et Biophysica Acta. Gene Regulatory Mechanisms, 2019. **1862**(9): p. 194408.
49. *A COMPARISON OF LAC REPRESSOR BINDING TO OPERATOR AND TO NONOPERATOR DNA.*
50. Kownin, P., E. Bateman, and M.R. Paule, *Eukaryotic RNA polymerase I promoter binding is directed by protein contacts with transcription initiation factor and is DNA sequence-independent.* Cell, 1987. **50**(5): p. 693-9.
51. Marilley, M. and P. Pasero, *Common DNA Structural Features Exhibited by Eukaryotic Ribosomal Gene Promoters.* Nucleic Acids Research, 1996. **24**(12): p. 2204-2211.
52. Marilley, M., et al., *DNA structural variation affects complex formation and promoter melting in ribosomal RNA transcription.* Mol Genet Genomics, 2002. **267**(6): p. 781-91.
53. Roux-Rouquie, M. and M. Marilley, *Modeling of DNA local parameters predicts encrypted architectural motifs in Xenopus laevis ribosomal gene promoter.* Nucleic Acids Res, 2000. **28**(18): p. 3433-41.
54. Smircich, P., M.A. Duhagon, and B. Garat, *Conserved Curvature of RNA Polymerase I Core Promoter Beyond rRNA Genes: The Case of the Tritryps.* Genomics Proteomics Bioinformatics, 2015. **13**(6): p. 355-63.
55. Harteis, S. and S. Schneider, *Making the Bend: DNA Tertiary Structure and Protein-DNA Interactions.* International journal of molecular sciences, 2014. **15**(7): p. 12335-12363.
56. Rohs, R., et al., *Origins of Specificity in Protein-DNA Recognition.* Annual Review of Biochemistry, 2010. **79**(1): p. 233-269.
57. Arber, W. and S. Linn, *DNA Modification and Restriction.* Annual review of biochemistry, 1969. **38**(1): p. 467-500.
58. Loenen, W.A.M., et al., *Highlights of the DNA cutters: a short history of the restriction enzymes.* NAR Breakthrough Article, 2014. **42**(1): p. 3-19.
59. Mirkin, S.M., *Discovery of alternative DNA structures: a heroic decade (1979–1989).* Frontiers in bioscience, 2008. **13**(13): p. 1064-1071.
60. Williams, R.J., *Restriction endonucleases: classification, properties, and applications.* Molecular biotechnology, 2003. **23**(3): p. 225-243.
61. Pingoud, A. and A. Jeltsch, *Structure and function of type II restriction endonucleases.* Nucleic acids research, 2001. **29**(18): p. 3705-3727.
62. Dickerson, R.E., et al., *Definitions and nomenclature of nucleic acid structure components.* Nucleic acids research, 1989. **17**(5): p. 1797-1803.

63. Zhou, T., et al., *DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale.* Nucleic acids research, 2013. **41**(Web Server issue): p. W56-W62.
64. Liu, P., et al., *The role of DNA shape in protein-DNA recognition.* Nature (London), 2009. **461**(7268): p. 1248-1253.
65. Rohs, R., et al., *The role of DNA shape in protein-DNA recognition.* Nature, 2009. **461**(7268): p. 1248-1253.
66. Neugebauerová, S. and J. Kypr, *Invariant and Variable Base Stacking Geometries in B-DNA and A-DNA.* Journal of biomolecular structure & dynamics, 2000. **18**(1): p. 73-81.
67. Protozanova, E., P. Yakovchuk, and M.D. Frank-Kamenetskii, *Stacked–Unstacked Equilibrium at the Nick Site of DNA.* Journal of molecular biology, 2004. **342**(3): p. 775-785.
68. Yakovchuk, P., E. Protozanova, and M.D. Frank-Kamenetskii, *Base-stacking and base-pairing contributions into thermal stability of the DNA double helix.* Nucleic acids research, 2006. **34**(2): p. 564-574.
69. Burkhoff, A.M. and T.D. Tullius, *Structural details of an adenine tract that does not cause DNA to bend.* Nature (London), 1988. **331**(6155): p. 455-457.
70. Gabrielian, A. and S. Pongor, *Correlation of intrinsic DNA curvature with DNA property periodicity.* FEBS letters, 1996. **393**(1): p. 65-68.
71. Gabrielian, A., A. Simoncsits, and S. Pongor, *Distribution of bending propensity in DNA sequences.* FEBS letters, 1996. **393**(1): p. 124-130.
72. Okonogi, T.M., et al., *Sequence-Dependent Dynamics of Duplex DNA: The Applicability of a Dinucleotide Model.* Biophysical journal, 2002. **83**(6): p. 3446-3459.
73. Brukner, I., et al., *Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides.* The EMBO journal, 1995. **14**(8): p. 1812-1818.
74. Goodsell, D.S. and R.E. Dickerson, *Bending and curvature calculations in B-DNA.* Nucleic acids research, 1994. **22**(24): p. 5497-5503.
75. Jen-Jacobson, L., *Protein-DNA recognition complexes: Conservation of structure and binding energy in the transition state.* Biopolymers, 1997. **44**(2): p. 153-180.
76. Von Hippel, P.H. and O.G. Berg, *On the Specificity of DNA--Protein Interactions.* Proceedings of the National Academy of Sciences - PNAS, 1986. **83**(6): p. 1608-1612.
77. Rozenberg, H., et al., *Structural Code for DNA Recognition Revealed in Crystal Structures of Papillomavirus E2-DNA Targets.* Proceedings of the National Academy of Sciences - PNAS, 1998. **95**(26): p. 15194-15199.
78. Lynch, T.W., et al., *Integration Host Factor: Putting a Twist on Protein–DNA Recognition.* Journal of molecular biology, 2003. **330**(3): p. 493-502.
79. Ansari, Aseem Z. and Kimberly J. Peterson-Kaufman, *A Partner Evokes Latent Differences between Hox Proteins.* Cell, 2011. **147**(6): p. 1220-1221.
80. Slattery, M., et al., *Cofactor Binding Evokes Latent Differences in DNA Binding Specificity between Hox Proteins.* Cell, 2011. **147**(6): p. 1270-1282.
81. Hueber, S.D., et al., *Improving Hox protein classification across the major model organisms.* PloS one, 2010. **5**(5): p. e10820.
82. Kouzarides, T., *Chromatin Modifications and Their Function.* Cell, 2007. **128**(4): p. 693-705.
83. *Gene Cards The Human Gene Database
*.
84. Leung, C.C.Y. and J.N.M. Glover, *BRCT domains.* Cell cycle (Georgetown, Tex.), 2011. **10**(15): p. 2461-2470.

85. van Riggelen, J., A. Yetil, and D.W. Felsher, *MYC as a regulator of ribosome biogenesis and protein synthesis.* Nature Reviews. Cancer, 2010. **10**(4): p. 301-309.
86. White, R.J., *RNA polymerases I and III, non-coding RNAs and cancer.* Trends in genetics: TIG, 2008. **24**(12): p. 622-629.
87. Dang, C.V., *MYC on the Path to Cancer.* Cell, 2012. **149**(1): p. 22-35.
88. Grummt, I., *Wisely chosen paths – regulation of rRNA synthesis.* The FEBS Journal, 2010. **277**(22): p. 4626-4639.
89. Haddach, M., et al., *Discovery of CX-5461, the First Direct and Selective Inhibitor of RNA Polymerase I, for Cancer Therapeutics.* ACS Medicinal Chemistry Letters, 2012. **3**(7): p. 602-606.
90. Mars, J.-C., et al., *The chemotherapeutic agent CX-5461 irreversibly blocks RNA polymerase I initiation and promoter release to cause nucleolar disruption, DNA damage and cell inviability.* NAR cancer, 2020. **2**(4): p. zcaa032.
91. Conte, C., et al., *Novel mutations of TCOF1 gene in European patients with treacher Collins syndrome.* BMC Medical Genetics, 2011. **12**: p. 125.
92. Edery, P., et al., *Apparent genetic homogeneity of the Treacher Collins-Franceschetti syndrome.* American Journal of Medical Genetics, 1994. **52**(2): p. 174-177.
93. Sommerville, J., *RNA polymerase I promoters and transcription factors.* Nature, 1984. **310**(5974): p. 189-190.

Chapter 2

Specific Structural features of the RNA polymerase I core promoter element targeted by Core Factor

Nathan J. Munoff[1], Wayne A. Decatur[1], Brian J. Zeberl[1], Matthew A. Palmer[1], Zsuzsa K. Szemere[1], Aula M. Fakhouri[1], Bruce A. Knutson[1]*

[1]SUNY Upstate Medical University

Department of Biochemistry and Molecular Biology

750 East Adams Street, Syracuse, NY 13210

*Corresponding author, Phone 315-464-8709, Knutsonb@upstate.edu

Authorship Contributions

1. **Effect of Point Mutations on CF DNA Binding**
   NJM, BJZ
2. **Correlation between Competition and DNA Structural Features**
   NJM, BJZ
3. **Identification of Novel CEs by In Vitro SELEX**
   NJM
4. **SELEX Enriched Structural Features**
   NJM, BAK
5. **Two Classes of SELEX Sequences**
   NJM, BAK
6. **Importance of Structural Features of Top Repeat**
   NJM
7. **SELEX Validation**
   NJM
8. **Identification of Novel CEs by In Vivo Selection**
   NJM
9. **Enriched Structural Preferences In Vivo**
   NJM, BAK
10. **Two classes of In Vivo Sequences**
    NJM, BAK

Participated in Research Design: NJM, BAK
Conducted Experiments: NJM, BJZ, BAK, ZKS, MP, AMF
Performed Data Analysis: NJM, BJZ, BAK, ZKS, MP
Wrote or Contributed to Writing Manuscript: NJM, BAK

Abstract

RNA Polymerase I (Pol I) synthesizes ribosomal RNA and is one of the three essential DNA-dependent RNA polymerases in eukaryotes. In yeast, Core Factor (CF) is a critical and essential Pol I transcription factor that plays fundamental roles in the transcription process by recruiting Pol I and opening Pol I promoter DNA before initiation. CF binds to a ~24 bp region in the rDNA promoter called the Core Element (CE). However, it was previously unclear how CF precisely recognizes the CE. Proteins use two main mechanisms when interacting with DNA: base-pair readout and shape/structural readout. Base-readout is the most common mechanism, and it is dominated by hydrogen bonding between the amino acid residues and base-pair hydrogen bond donors and acceptors. The second mechanism centers around DNA shape readout, where proteins target specific DNA features such as curvature, bendability, and groove width. Previous studies from our lab have shown that CF and its human orthologue, Selectivity Factor 1 (SL1), use an evolutionarily conserved mechanism to target DNA. This mechanism is governed by interactions with the GC minor groove, a unique surface rarely targeted by DNA-binding proteins.

To further understand the extent of structural recognition, we have employed a variety of EMSA and selection-based methods to resolve the structural rules governing CF's interaction with DNA. Our results show that CF is particularly sensitive to any structural changes in the rigid A patch of the CE via single bp mutants for binding. Additionally, we demonstrate an overall preference of CF for more bendable novel sequences via In Vitro SELEX and In Vivo selection. Furthermore, we were able to classify novel sequences into rigid and flexible categories based on their bendability

profile. We also characterized the preferences and effects of the structural properties of Roll, Propeller Twist (ProT), Helix Twist (HelT), and Minor Groove Width (MGW) in the CE on CF binding. We found that ProT and HelT were especially important to CF binding and that CF preferred sequences with decreased ProT and HelT. Our findings support a model that CF-CE interaction is primarily governed by DNA shape-based structural features rather than sequence.

Introduction

Three DNA dependent RNA polymerases (Pol I, II, and III) carry out the critical transcription process in eukaryotic organisms[1]. Each polymerase produces different classes of RNA, with different sets of transcription factors aiding in their recruitment to respective promoters. Pol II mainly produces mRNA, and Pol III mainly tRNA[1]. Pol I is solely responsible for producing ribosomal RNA (rRNA) and transcribing the rDNA gene into 18S, 5. 8S, and 25S rRNA, which make up the catalytic and structural parts of the ribosome[2-4]. The ribosome produces all proteins in the cell, making the synthesis of rRNA and ribosomes vital[5, 6]. Additionally, Pol I transcription accounts for over 60% of total cellular transcription activity, necessitating high transcription rates[3]. Any dysregulation of Pol I transcription often leads to severe consequences, as human cancers and developmental disorders such as Treacher Collins syndrome demonstrate[7-14].

Pol I transcription begins at the rDNA promoter and features a structurally conserved bipartite composition called the upstream Control Element (UCE) and the Core Promoter Element (CE) across eukaryotes[15-20]. In yeast, the UCE and CPE are referred to as the upstream activation sequence (UAS) and the core element (CE),

respectively[18-20]. The UAS is located between positions -150 and -60, and the CE is found between positions -38 to -15 relative to the start site of transcription[18-20]. To begin transcription, the pre-initiation complex (PIC) must be formed at the rDNA promoter, consisting of all the necessary machinery[15-17]. In the yeast Pol I transcription system, Pol I, Upstream activation factor (UAF), CF, Rrn3, and TATA-box binding protein (TBP) make up the PIC[3, 4, 21-29]. UAF binds to the UAS, CF binds to the CE, TBP stabilizes the PIC, and Rrn3 induces and stabilizes the monomeric form of Pol I[3, 4, 21-33]. Although the complete system experiences a 40-fold increase in transcription rates in vitro over basal initiation, only Pol I, Rrn3, and CF are required for initial transcription and promoter escape[34-36].

CF is a heterotrimer consisting of proteins Rrn6, Rrn7, and Rrn11[34, 35, 37-39]. Rrn7 and Rrn11 mediate contact with DNA from positions -27 to -20 and -24 to -16, respectively[40, 41]. Rrn7, a paralog of general transcription factors TFIIB and Brf1 from Pol II and III, contains an N-terminal zinc ribbon that interacts with Pol I and two cyclin domains that bind to DNA[22, 23, 27, 38, 40-46]. Rrn11 has an N and C-terminal domain separated by a tetratricopeptide domain (TPR) and mediates DNA contacts through three out of the four helices within the NTD and helix 5 of the TPR domain[23, 27, 40, 41, 46]. Rrn7 and Rrn11 together secure the DNA on opposite faces[23, 24, 27, 40, 41, 46]. CF bound to rDNA resembles a right hand that holds the rDNA between the palm and fingers, with the N-terminal regions of Rrn11 and Rrn6 making up the palm and Rrn7 constituting the fingers[23, 24, 27, 40, 41, 46]. The Rrn6 C-terminus forms the knuckles[23, 24, 27, 40, 41, 46].

Eukaryotic rDNA promoters exhibit conserved structural features, such as intrinsic curvature and kinks, with many containing a bend located ~25bps upstream of the TSS[47-52]. In yeast, CE contains two kinks of 45° and 35° at positions -16 and -21, respectively[40]. Although the Pol I transcription system possesses many structurally conserved features, there is a distinct lack of rDNA promoter sequence conservation among eukaryotic species[49]. This could be due to the tandem repetitive nature of rDNA gene loci which may contribute to the greater degree of species specificity seen in Pol I systems as opposed to Pol II or III systems[53-58]. This species specificity then leading to an increased rate of concerted evolution or divergence[53, 57]. Despite this lack of sequence conservation, yeast CF can bind the human core promoter element in vitro, and the hCPE functions in yeast, albeit in a positionally dependent manner[59].

Yeast CF has also been found to prefer the GC-rich minor groove in a recent DNA binding study[59]. It has been suggested that this preference may be a conserved structural feature, as *Acanthamoeba castellani's* orthologous protein TIF-IB has similar DNA binding preferences and mainly contacts the minor groove[50]. The narrowness of the minor groove is believed to be crucial for binding, as point mutations that widened the narrowest parts of the CPE resulted in the largest decreases in transcription[50]. TIF-IB was also shown to induce bends in the CPE upon binding, which weakened helical stability towards the TSS and unwound the double helix[50]. Both yeast CF and its human ortholog SL1 are likely influenced by multiple structural features of rDNA in their DNA binding mechanisms as opposed to a specific sequence.

In this study, we describe the structural binding preferences of yeast Pol I transcription factor CF. Our results revealed that CF is capable of tolerating mutations at

some positions of the CE while mutation in the rigid "A" patch being particularly sensitive to mutations changing structural properties. Along with conditional tolerance for sequence mutations, our results show that CF prefers a variety of structural features such as overall increased bendability and decreased curvature as well as specific profiles of bendability[60, 61]. We categorized sequences into rigid (WT like) and flexible groupings based on bendability profiles and found CF preferred flexible sequences more than rigid ones. Furthermore, we describe the preferences of CF for the parameters of helix twist, propeller twist, roll, and minor groove width[62]. CF's preferences in vitro closely matched what we observed in vivo with a few exceptions. All this data together, strongly suggests that CF recognizes and binds to the CE in more of a structural than sequence dependent manner. Here, we characterize the specific structural properties of this structural manner.

Results

3.01 Effect of Point Mutations on CF DNA Binding

We first investigated the effect and significance of sequence on CF-CE binding. We did so by testing the impact of point mutations to the CE sequence on CF binding via competition based Electrophoretic Mobility Shift Assays (EMSA). We created double stranded DNA competition probes spanning -31 to -14 of the promoter mutating positions -28 to -17. Each position was mutated to the other 3 possible bases in addition to Inosine. Inosine was chosen as it is a nucleotide analogue of Guanine but lacks the C2 amino group that would normally occupy the minor groove and can bind to any nucleotide although it prefers cytosine[63, 64]. Loss of the amino group allows for investigation

into the importance of the minor groove itself for targeting by proteins without base specific contacts and Inosine mimics an AT minor groove and a GC major groove. Competition oligos were incubated with IR labeled WT CE and CF and observed for changes in competition for CF relative to WT CE.

Some positions tolerated point mutants better than others. For instance, positions -28, -26, -23, and -19 tolerated any point mutations with very little loss of competition for CF relative to WT CE (**Fig. 1**). When position -27 was changed to a A, G, or I a reduction in competition was seen indicating that a pyrimidine is required at that position (**Fig. 1**). Position -25 may require having a purine due to its poor tolerance and reduced competition of mutations to T and C vs G and I (**Fig. 1**). The GC minor groove may be particularly important for positions -24 and -19 as they tolerate the changes to I more poorly than to A,T, or C (**Fig. 1**). This would indicate that are either base specific interactions occurring here in the minor groove or that a narrow minor groove is particularly important. Position -18 suffered significant losses in competition when mutated to an A or T vs C or I demonstrating a need for a GC base pair at this position. Mutations made to positions -22 to -20 had some of the greatest losses in competition for CF relative to WT particularly at position -20 being mutated to a C resulting in a 46% decrease in competition (**Fig. 1**). Overall, mutations were more tolerated outside positions -22 to -20 than within demonstrating this region to be particularly sensitive to change.

3.02 Correlation between Competition and DNA Structural Features

We then asked if there was a correlation between the level of competition and certain DNA structural features. We examined six predictable structural properties that include bendability, curvature, roll, propeller twist(ProT), minor groove width (MGW), and helix twist(HelT). We predicted these structural properties of each of the point mutation variants we tested in our competition assays to see if there was any correlation between loss of competition and a change in a DNA structure feature. For both bendability and curvature, there is a clear correlation between changes in bendability and curvature, and loss in competition against WT CE from positions -22 to -20 (**Fig. 1**). At these positions, the largest increases in bendability and decreases in curvature relative to WT strongly correlate with the largest losses in competition. This would suggest that at these positions, the structural features of low bendability and high curvature are important for CF binding. This correlation between competition and structural features also applies to increases in MGW and decreases in HelT, resulting in losses in competition at positions -21 and -20. This indicates that a narrow MGW and decreased HelT are also important at these positions. These findings match well with previous findings in *A. castellanii* of the importance of DNA conformation for TIF-IB and human SL1 binding. Roll and ProT do not follow this correlation in any clear pattern at any positions, however. The overall lack of toleration of most structural changes from positions -22 to -20 reveal the structural recognition occurring here and that this AT rich rigid patch of WT CE is crucial for CF binding (**Fig. 1**).

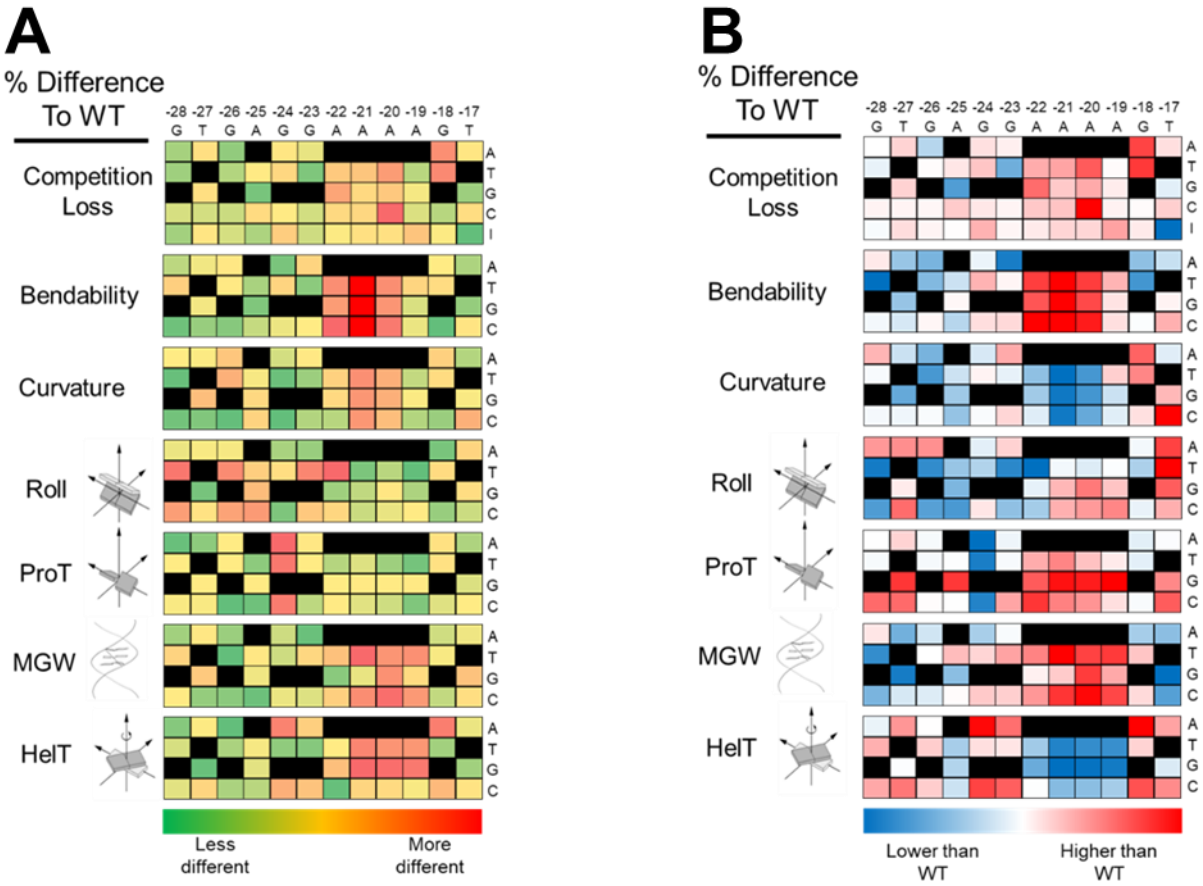

Figure 1. Effects of DNA Structural Properties on CF Binding.

Heatmaps showing changes in competition, bendability, curvature, roll, propeller twist(ProT), minor groove width(MGW), and helix twist(HelT) of single bp mutants of WT CE from positions -28 to -17. Mutations were made to all other possible nucleotides at each position as well as Inosine. The program Bend-it® was used to calculate GC% as well as the bendability and curvature of DNA sequences. Bendability and Curvature were calculated in Bend-it® using a sliding trimer window to determine these values at any given central base pair. Bendability refers to the ability of a DNA sequence to be deformed or change shape and occupy a range of conformations. Curvature refers to more static bends over a series of base pairs. The values calculated for bendability and curvature are relative arbitrary units (A.U.) based on DNaseI, nucleosome positioning, crystal structure, and other experimental data. The four structural properties of roll, propeller twist, minor groove width, and helical twist were calculated with the DNAshape program.

Heatmaps showing the **A)** % difference to WT and the **B)** % higher or lower than WT of competition, direct binding, bendability, curvature, roll, propeller twist, minor groove width, and helix twist of the single bp mutants.

3.03 Identification of Novel CEs by In Vitro SELEX

To determine the critical structural based features CF targets, we employed SELEX (Selective Evolution of Ligands with Exponential enrichment) to randomize the entire CE at once and look at a more diverse library of unique sequences. Each round of SELEX involved the incubation of CF with this library of randomized CE sequences flanked by WT Pol I promoter sequence. Sequences bound by CF were separated from those that didn't via an EMSA, and the CF bound sequences were extracted and enriched by PCR amplification for use in the next selection round. We completed seven SELEX rounds before encountering insurmountable byproduct formation due to repetitive PCR amplification[65].

Sequencing was carried out via insertion of the selected sequences into a plasmid via QuikChange that could be isolated and sequenced. Only the last three rounds 5, 6, and 7 were needed to be sequenced to show evidence of selection while avoiding having to sequence every round. Our goal of ~500 sequences sequenced for each round was chosen as sufficient to demonstrate enrichment across successive rounds of selection. Initial quantification of enrichment was measured by increases in repeated sequences found in the sequencing data of successive rounds against experimental and theoretical controls. We observed significant increases in repeated sequences from round to round with one sequence repeated more than the rest, dubbed the top repeat (**Fig. 2A**). The top repeat dominated the repeated sequences in rounds 6 and 7 demonstrating enrichment (**Fig. 2A**). However, it must be noted that the majority of sequences across all sequenced rounds of SELEX remained unique being only repeated once. Additionally, a lack of a conserved consensus sequence is seen in the weblogos of R5-7 besides a significant

increase in GC content (**Fig. 2B**). An analysis of these sequences using MEME suite confirms a lack of consensus sequence and or motif[66].

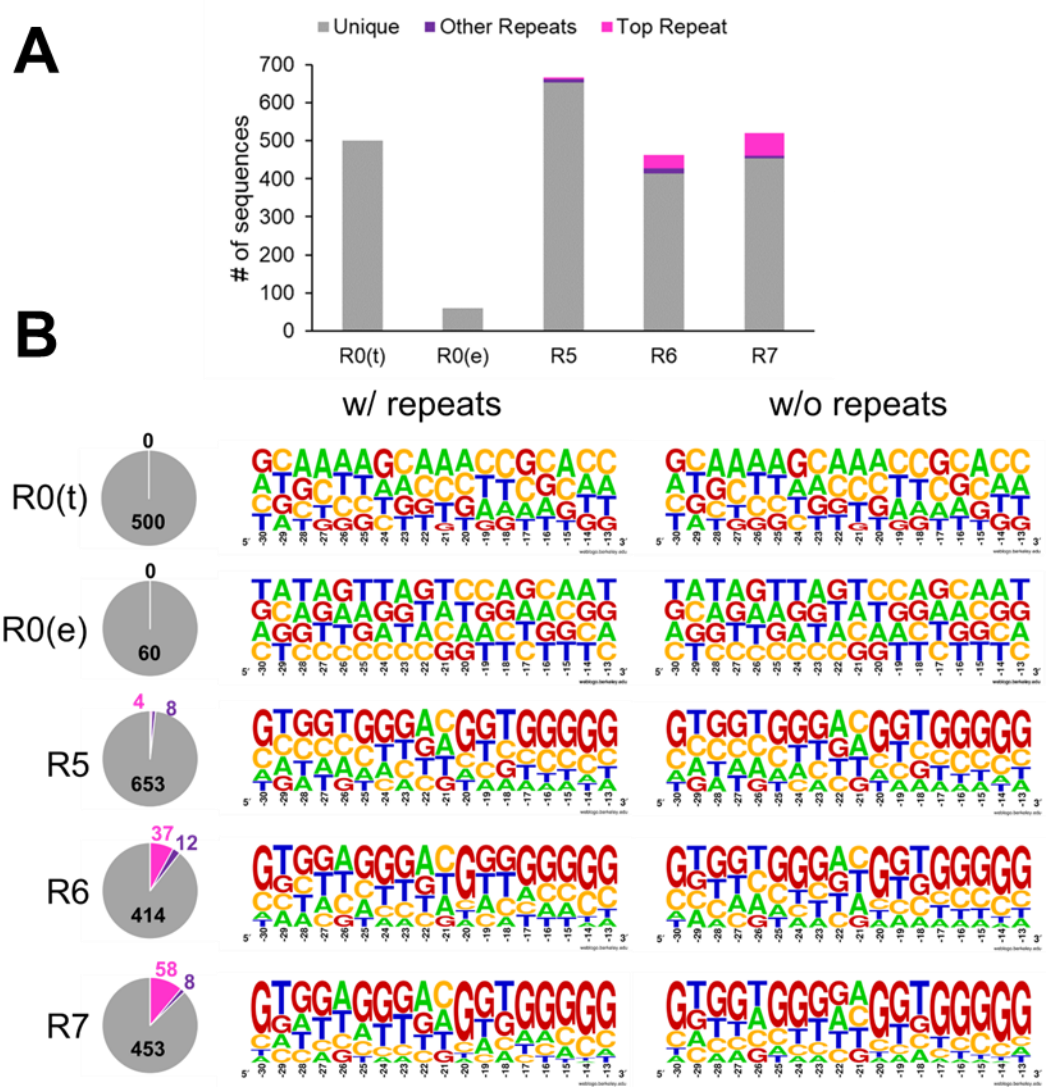

Figure 2. In Vitro SELEX Round Summary

A) Bar graph depicting the number and uniqueness of sequences found during sequencing of rounds 5-7 of in vitro SELEX and controls. Sequences only appearing once were classified as unique, twice or more as other repeats, and the sequence that appeared the most as the top repeat. Theoretical sequences were generated randomly in silico and experimental sequences were found by sequencing the original SELEX library before any rounds of in vitro selection with CF.

B) Weblogos depicting the nucleotide frequencies of the randomized regions of sequences found during sequencing of rounds 5-7 of in vitro SELEX and controls. Pie charts depicting the number and uniqueness of the sequences found

3.04 SELEX enriched Structural Features

To further examine the structural preferences of CF we analyzed all the selected sequences round by round for GC%, curvature, and bendability averaged over the randomized region of the CE. We observed a steady increase in GC% and bendability, and a decline in curvature from R5-7 with or without the inclusion of repeated sequences (**Fig 3. A,C,E**). These results indicate having overall increased GC content and bendability may be especially important for CF as well as these two properties have a strong direct correlation with one another.

Next, we examined these features and properties at the bp level. The clearest trends observed were that CF prefers curvature at the upstream end of the CE particularly at positions -28 and -27 (**Fig. 3C**). However, even at these positions with each successive round the level of curvature decreases. Bendability on the other hand, is more widely distributed with small patch of low bendability or rigidity found at positions -26 to -24, made more pronounced when including repeats (**Fig. 3F**). Increasing GC content is most pronounced from positions -17 to -13 at the downstream end (**Fig. 3B**). Overall, these results indicate high GC% and bendability with low curvature are important features of DNA for CF binding and recognition.

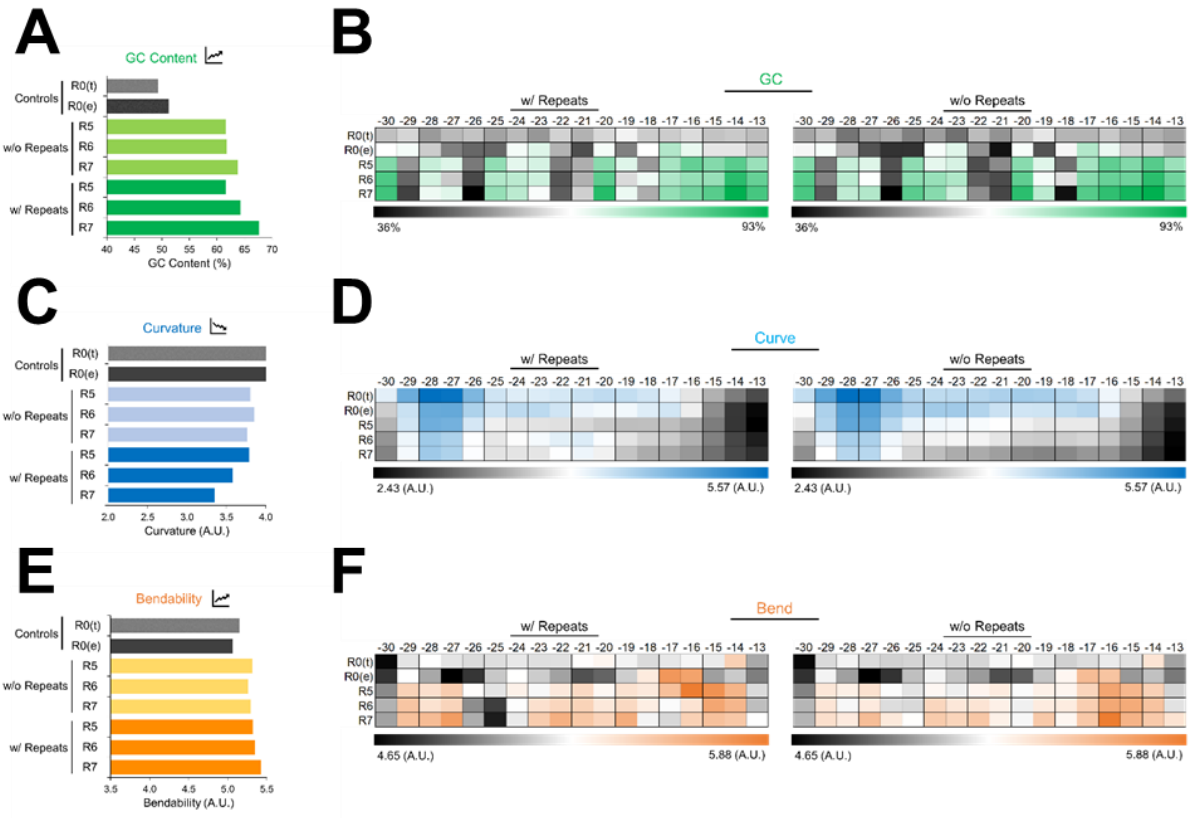

Figure 3. Bend-it Analysis of In Vitro SELEX Sequences Enriched by Yeast RNA Polymerase I CF

The program Bend-it® was used to calculate GC% as well as the bendability and curvature of DNA sequences with and without duplicates across rounds (R) 5-7 of *in vitro* SELEX compared to theoretical (R0(t)) and experimental (R0(E)) controls. The theoretical control consists of 500 random 18mer sequences generated *in silico* and the experimental control of 60 sequences directly cloned and sequenced from the starting SELEX library. Bendability and Curvature were calculated in Bend-it® using a sliding trimer window to determine these values at any given central base pair. Bendability refers to the ability of a DNA sequence to be deformed or change shape and occupy a range of conformations. Curvature refers to more static bends over a series of base pairs. The values calculated for bendability and curvature are relative arbitrary units (A.U.) based on DNaseI, nucleosome positioning, crystal structure, and other experimental data. The three parameters of GC%, bendability, and curvature were calculated both as an average over the entire randomized region of selected sequences as well as at each position individually.

A, C, E) Bar graphs of the GC%, curvature, and bendability of sequences averaged over the entire randomized region of the selected sequences from positions -30 to -13 upstream of the transcription start site with and without duplicates.

B, D, F) Heatmaps of the GC%, curvature, and bendability averaged at each base pair of the entire randomized region of the selected sequences from positions -30 to -13 upstream of the TSS, with and without duplicates.

We also analyzed all the selected sequences round by round for ProT, HelT, MGW, and Roll averaged over the randomized region of the CE. We observed a steady decline across rounds in helix twist which structurally widens the minor groove (**Fig. 4A**) Propeller twist also steadily declines although less so when including repeats (**Fig. 4C**). Propeller twist correlates with and likely determines another structural feature known as slide. Slide is the displacement of one base-pair to its neighbor along the long axis of a dinucleotide step[62]. High degrees of propeller twist are associated with low degrees of slide via a mechanism of stereochemical locks resisting slide[61, 67, 68]. This decrease in propeller twist fits with the increase in bendability seen across rounds as slide becomes less inhibited. Roll steadily increased across rounds when repeats were not included but remained steady when they were (**Fig. 4G**). Positive roll bends DNA towards the major groove thus opening base pairs towards the minor groove and widening it. It's clear that the increasing appearances of the top repeat are contributing to the plateauing of roll when repeats are included in the rounds. Despite this, there is clear enrichment of positive roll which agrees with previous studies that have demonstrated that the GC minor groove is important for CF binding[59]. Lastly, minor groove width showed fluctuations across rounds both including and not including repeats with no overall major trend, although R7 had the highest minor groove width both with and without repeats (**Fig. 4E).** However, it's not clear why there is such a disparity between the theoretical and experimental controls for Roll and MGW. These results confirm the importance of the GC minor groove and structural features that contribute to increased bendability.

Again, we examined these features at the bp level. We observed decreased helix twist mostly contained to positions -17 to -13 (**Fig. 4**). Any widening of the minor

groove was mostly contained to bps -28 to -26, -22 to -21, and -18 to -17 with more widening occurring in the upstream half of sequences than the downstream half (**Fig. 4F**). The regions of least propeller twist were found at the downstream end from -17 to -13 potentially indicating that this region is generally preferred to be flexible by CF (**Fig. 4D**). Lastly, areas with higher roll both with and without repeats were found at -27, -25, -22, -20, and -17 (**Fig. 4G**) These results indicate CF prefers upstream widening of the minor groove, making important base specific contacts on the upstream half of the promoter and prefers a stiffer downstream end of the promoter.

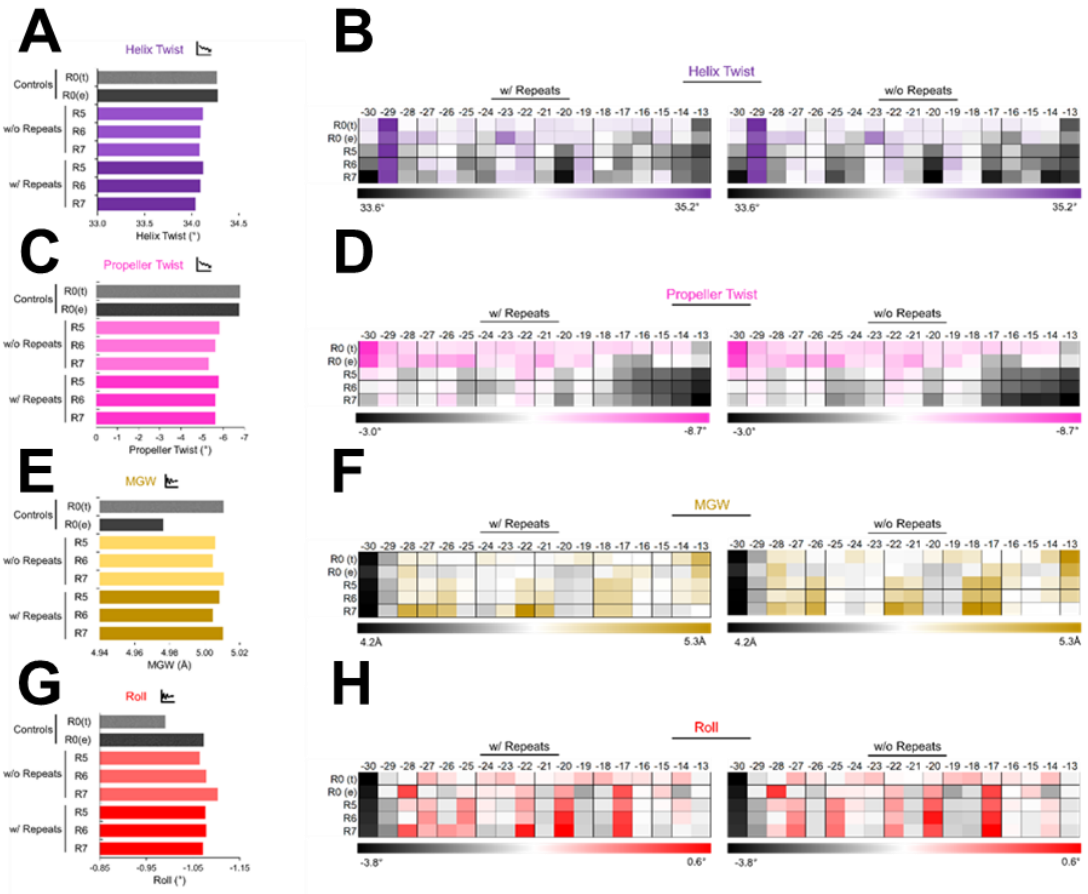

Figure 4. GB Shape Analysis of In Vitro SELEX Sequences Enriched by Yeast RNA Polymerase I CF.

The genome browser database for DNA shape annotations GB shape was used to calculate the Helical Twist, Propeller Twist, Minor Groove Width, and Roll of DNA sequences with and without duplicates across rounds (R) 5-7 of *in vitro* SELEX compared to theoretical (R0(t)) and experimental (R0(E)) controls. The theoretical control consists of 500 random 18mer sequences generated *in silico* and the experimental control of 60 sequences directly cloned and sequenced from the starting SELEX library. The DNA shape annotations were derived with a high-throughput method for DNA shape predictions and constitute the whole-genome complement to a motif database of transcription factor binding sites. All four DNA shape parameters were calculated both as an average over the entire randomized region of selected sequences as well as at each position individually.

A, C, E, G) Bar graphs of the Helical Twist, Propeller Twist, Minor Groove Width, and Roll of sequences averaged over the entire randomized region of the selected sequences from positions -30 to -13 upstream of the transcription start site with and without duplicates.

B, D, F, H) Heatmaps of the Helical Twist, Propeller Twist, Minor Groove Width, and Roll averaged at each base pair of the entire randomized region of the selected sequences from positions -30 to -13 upstream of the TSS, with and without duplicates.

3.05 Two classes of SELEX sequences

While analyzing sequences from our SELEX results we noticed the presence of two distinct categories of sequences emerge. In particular, we noticed sequences from round 7 could be grouped based on their bendability profiles into rigid and flexible groupings. We classified a rigid sequence as one that had a large negative spike in bendability or rigid patch as WT CE does. We then further classified both the rigid and flexible groups into different subgroups. For rigid groupings we categorized them based on where the center of the dip in bendability fell in terms of bp location. For flexible groupings we categorized them by their average bendability values across the randomized region of the sequence. We again first looked at the GC content, bendability, and curvature of these two groupings as a whole focusing on sequences from round 7. Predictably we saw flexible sequences had more bendability (**Fig. 5E**). We also observed that rigid sequences had more curvature than their flexible counterparts (**Fig. 5C**). Lastly, we saw that Flexible sequences had more GC content than rigid ones (**Fig. 5A**).

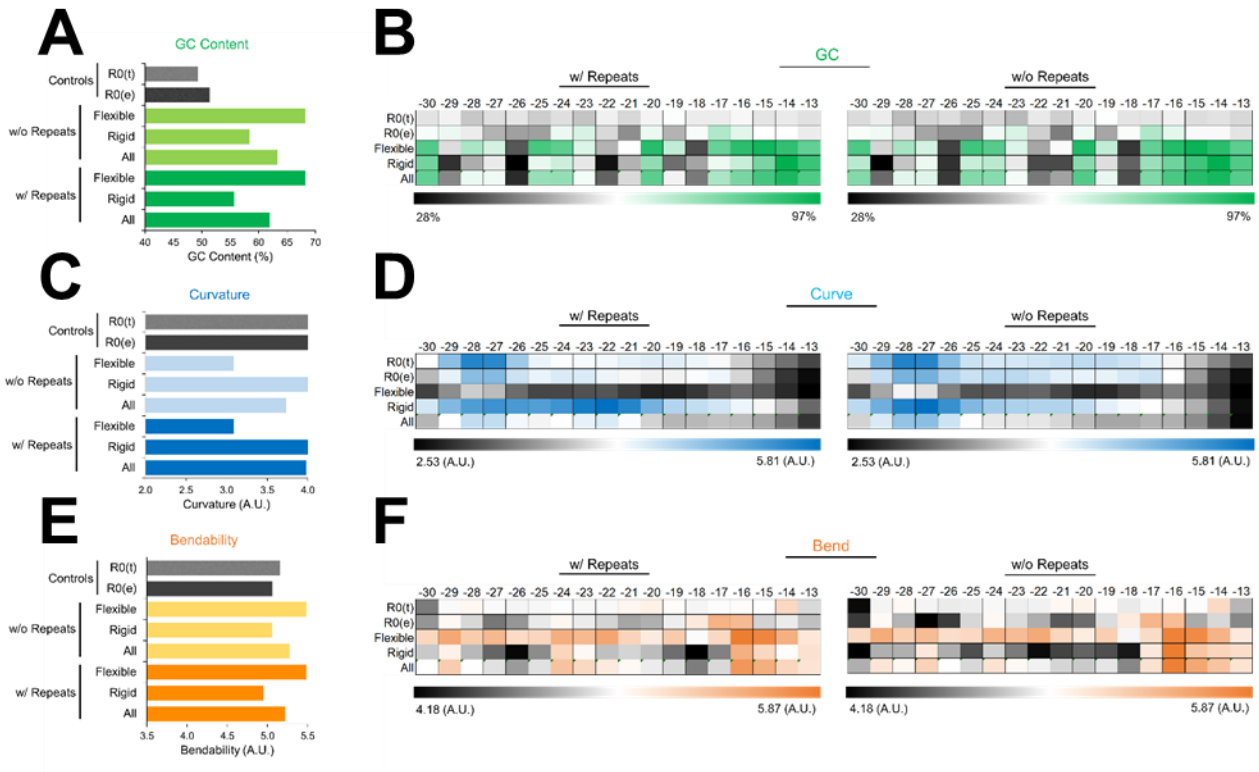

Figure 5. Bend-it Analysis of In Vitro SELEX Sequences Found in Round 7 Classified into Rigid or Flexible Based on Bendability.

The program Bend-it® was used to calculate GC% as well as the bendability and curvature of DNA sequences with and without duplicates of sequences classified as rigid or flexible from round 7 of *in vitro* SELEX compared to theoretical (R0(t)) and experimental (R0(E)) controls. Sequences were classified as rigid if they had a bendability value of <3 from bp 29-46 of the SELEX library which corresponds to -31 to -12 of the promoter. Sequences were classified as flexible if they did not have a bendability value <3 of the aforementioned library and promoter positions. The theoretical control consists of 500 random 18mer sequences generated *in silico* and the experimental control of 60 sequences directly cloned and sequenced from the starting SELEX library. Bendability and Curvature were calculated in Bend-it® using a sliding trimer window to determine these values at any given central base pair. Bendability refers to the ability of a DNA sequence to be deformed or change shape and occupy a range of conformations. Curvature refers to more static bends over a series of base pairs. The values calculated for bendability and curvature are relative arbitrary units (A.U.) based on DNaseI, nucleosome positioning, crystal structure, and other experimental data. The three parameters of GC%, bendability, and curvature were calculated both as an average over the entire randomized region of selected sequences as well as at each position individually.

A, C, E) Bar graphs of the GC%, curvature, and bendability of sequences averaged over the entire randomized region of the selected sequences from positions -30 to -13 upstream of the transcription start site with and without duplicates.

B, D, F) Heatmaps of the GC%, curvature, and bendability averaged at each base pair of the entire randomized region of the selected sequences from positions -30 to -13 upstream of the TSS, with and without duplicates.

We then analyzed the additional structural features of helix twist, propeller twist, roll, and minor groove width of the flexible and rigid groupings as a whole. For helix twist it was higher in the rigid grouping both with and without repeats than the flexible grouping (**Fig. 6A**). The opposite was true of MGW with it being the lowest in the rigid groupings both with and without repeats (**Fig. 6E**). Propeller twist was the highest in the rigid grouping with and without repeats, continuing to support the mechanism of stereochemical locking (**Fig. 6C**). Lastly, roll was the lowest in the flexible grouping both with and without repeats (**Fig. 6G**).

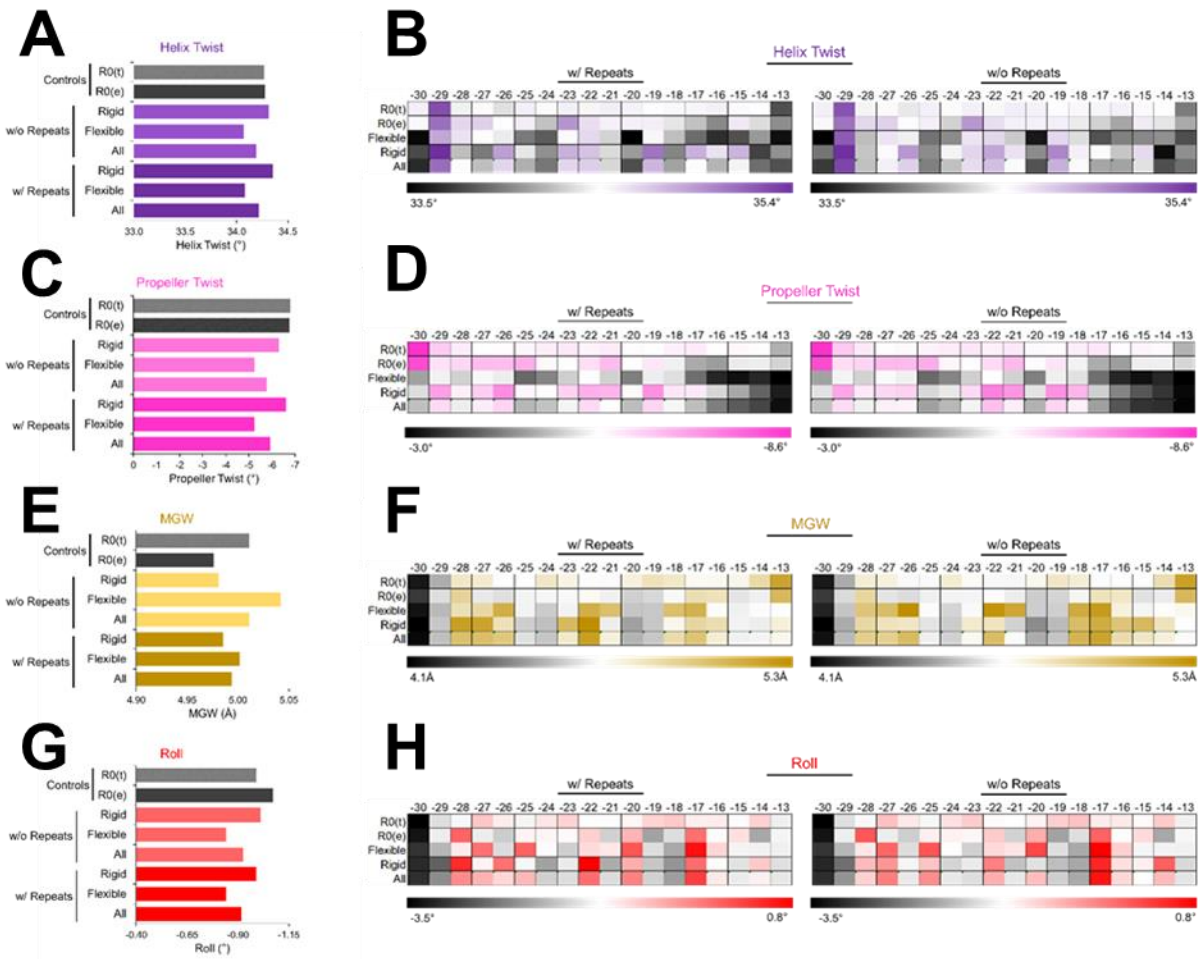

Figure 6. GB Shape Analysis of In Vitro SELEX Sequences Found in Round 7 Classified into Rigid or Flexible Based on Bendability.

The genome browser database for DNA shape annotations GB shape was used to calculate the Helical Twist, Propeller Twist, Minor Groove Width, and Roll of DNA sequences with and without duplicates of round 7 sequences of *in vitro* SELEX grouped into rigid and flexible categories based on bendability. These sequences were compared to theoretical (R0(t)) and experimental (R0(E)) controls. The theoretical control consists of 500 random 18mer sequences generated *in silico* and the experimental control of 60 sequences directly cloned and sequenced from the starting SELEX library. The DNA shape annotations were derived with a high-throughput method for DNA shape predictions and constitute the whole-genome complement to a motif database of transcription factor binding sites. All four DNA shape parameters were calculated both as an average over the entire randomized region of selected sequences as well as at each position individually.

A, C, E, G) Bar graphs of the Helical Twist, Propeller Twist, Minor Groove Width, and Roll of sequences averaged over the entire randomized region of the selected sequences from positions -30 to -13 upstream of the transcription start site with and without duplicates.

B, D, F, H) Heatmaps of the Helical Twist, Propeller Twist, Minor Groove Width, and Roll averaged at each base pair of the entire randomized region of the selected sequences from positions -30 to -13 upstream of the TSS, with and without duplicates.

3.06 Importance of Structural Features of Top Repeat

In terms of sequence, our top repeat and WT CE were very different, so we investigated further to ascertain what it was about the top repeat that made it so popular. Looking instead at the structural properties of intrinsic bendability and curvature, WT CE and our top repeat are very similar with a few key differences. For curvature, the profile of the top repeat is nearly identical to that of the WT except shifted ~4 bp downstream (**Fig. 7A**). For bendability, the profile is also close to that of WT and shifted ~5bp downstream but with a second major dip in bendability on the upstream end (**Fig. 7A**). The divergence in sequence but relative similarity of structural features such as bendability and curvature between the WT CE and our top repeat, indicate structure is playing a significant role in CF recognition of the CE. This theme of low sequence conservation and high structural feature conservation is also true for rDNA promoter sequences in general. Across many species, rDNA promoters share very little sequence conservation, yet share structurally conserved features such as intrinsic curvature and kinks[47-52]. Sometimes this means CF from one species can recognize and bind the CE from another such as with yeast CF and the human Pol I promoter in yeast[59].

To determine the relative specificity of the top repeat against WT for CF we performed EMSA based competition assays. Surprisingly, we saw that our top repeat initially performed quite poorly regardless of titration. Since we observed a similar but shifted bendability and curvature profile of the top repeat relative to WT CE, we hypothesized that this may be the reason that it competed for CF so poorly. The SELEX library of oligos were 75bp long but the competitors oligos used in the competition EMSAs were only 20bp long. It was possible that the loss of extra flanking sequence

was affecting the top repeat competitor binding. We then decided to investigate and refine where these new potential boundaries for CF binding were of the top repeat. We designed a series of new oligoes with the TR sequence flanked by varying lengths of WT sequence on either side to narrow down the new minimal boundaries needed for improved binding more comparable to WT CE (**Fig. 7B**). One oligo TRV10 competed for CF on par with WT CE (**Fig. 7C,D**). TR V10 has lost 5bp on the upstream end and gained 9 on the downstream end relative to the original top repeat but performs well. This fits with the idea that the downstream shift in bendability and curvature profiles is important for CF binding. Previous studies have demonstrated a similar effect showing that the human promoter which has very similar bendability and curvature profiles to the yeast promoter, functions in yeast in a positionally dependent manner[59]. These results indicate that specific bendability and curvature profiles of DNA are important for CF binding and recognition of the top repeat.

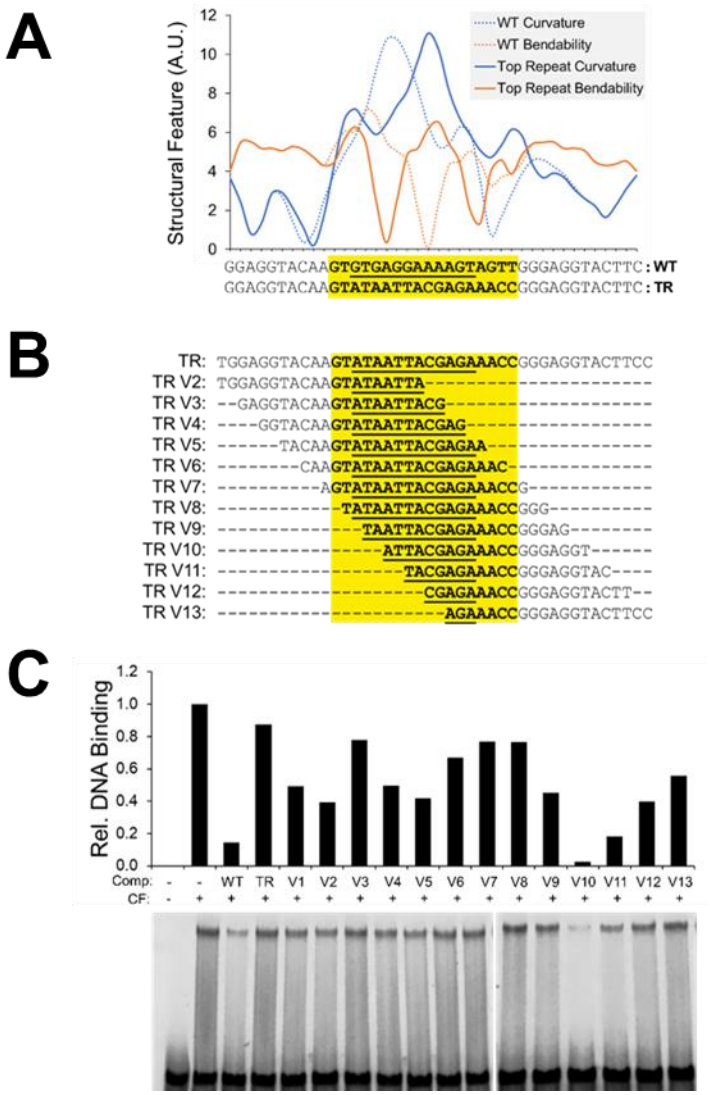

Figure 7. Top Repeat Binding Profile Analysis

The program Bend-it® was used to calculate the bendability and curvature profiles of the yeast WT Pol I promoter and the top or most repeated sequence found in round 7 of in vitro SELEX. Bendability and Curvature were calculated in Bend-it® using a sliding trimer window to determine these values at any given central base pair. Bendability refers to the ability of a DNA sequence to be deformed or change shape and occupy a range of conformations. Curvature refers to more static bends over a series of base pairs. The values calculated for bendability and curvature are relative arbitrary units (A.U.) based on DNaseI, nucleosome positioning, crystal structure, and other experimental data.

A) Line graph of the bendability and curvature of the yeast WT Pol I promoter and the TR sequence from round 7 of in vitro SELEX. Highlighted in yellow is the region of the Pol I promoter which was randomized during in vitro SELEX and the minimal CE is underlined.

B) Sequences of competitors to determine the binding boundaries of the TR. TR is the top repeat sequence from round 7 of in vitro SELEX in which competitors are based on.

Dashed lines indicate excluded base pairs. Highlighted in yellow is the region of the Pol I promoter which was randomized during in vitro SELEX and the minimal CE is underlined.

C) Bar graph depicting relative CF binding (R.B.) quantitation after unlabeled competition experiments. All values shown are relative to CF DNA binding in the absence of competitor oligos which is set at 1.0.

D) Effect of indicated TR competitors on CF DNA binding to the yeast ΔUAS IR-labeled promoter. Representative EMSA results are shown.

3.07 SELEX Validation

We also wanted to validate the binding ability of the other sequences found during in vitro SELEX. We took the consensus sequences of all R7 sequences both with and without repeats and we performed EMSA based competition assays (**Fig. 8A**). The consensus sequences both without repeats performed slightly better than the consensus sequence with repeats likely due to effects of the top repeat (**Fig. 8B,C**). Both performed nearly as well as WT, confirming the ability of these sequences to compete for CF and validating the selection of the sequences from in vitro SELEX.

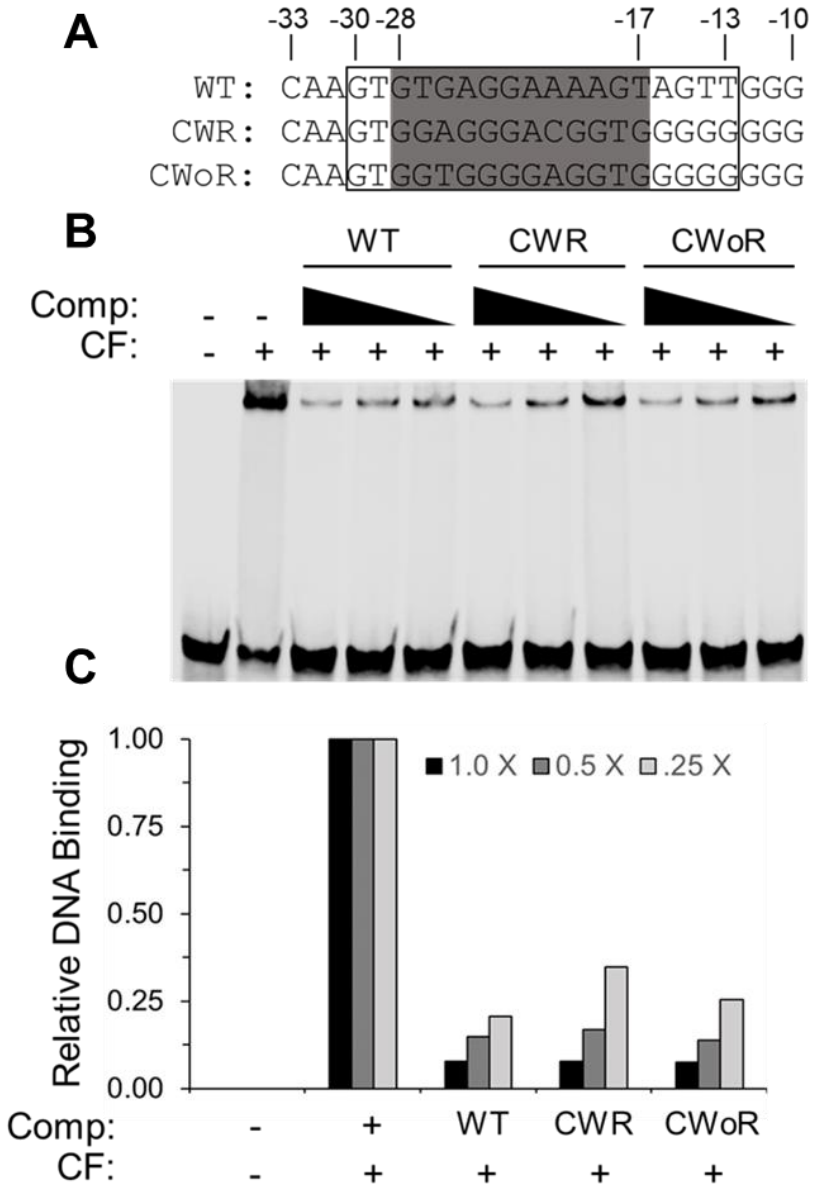

Figure 8. Validation of the In Vitro SELEX Round 7 Consensus Sequences Enriched by CF.

Consensus sequences were generated from all sequences found in R7 of SELEX both including and excluding any duplicated sequences. These sequences competed for CF against IR labeled yeast WT CE probe and were compared to WT CE competitor.

A) Sequences of competitors used. The minimal CE sequence needed for CF binding is shaded and the randomized region of the CE sequences found from SELEX boxed.
B) Effects of competitors on CF DNA binding to the yeast WT CE IR labeled promoter. Representative EMSA results are shown using a range of competitor concentrations.
C) Bar graph depicting the relative CF DNA binding quantitation after unlabeled competition assays. All values shown are relative to CF DNA binding in the absence of a competitor oligo which is set at 1.00.

3.08 Identification of Novel CEs by In Vivo Selection

We also employed an In Vivo selection process to complement and lend a more biologically relevant context to any critical structural based features of DNA that CF targets. We first randomized CE sequences which were then inserted into pMF150, a derivative of pNOY373 [2] with the -28 to -17 region of the promoter deleted, such that the sequences replaced the deleted region of the promoter. This vector was transformed into a yeast strain with the chromosomal rDNA deleted and moved to a plasmid with a URA marker. The URA marker aids in selection by producing a toxic product when grown on FOA. Any yeast growing on FOA will have successfully kicked out the plasmid with the URA marker and CF will have been able to bind to the mutant CE as failure to do so would result in death. The plasmids from yeast that successfully grew on FOA were then isolated and sequenced.

Initial quantification of enrichment was measured by increases in repeated sequences found in the sequencing data against experimental and theoretical controls. We observed a significant increase in repeated sequences from round to (**Fig. 9A**). Other repeats dominated any repeated sequences demonstrating enrichment (**Fig. 9A**). However, it must be noted that the majority of sequences remained unique. Additionally, a lack of a conserved consensus sequence is seen in the weblogo of RX besides a significant increase in GC content (**Fig. 9B**). An analysis of these sequences using MEME suite confirms a lack of consensus sequence and or motif.

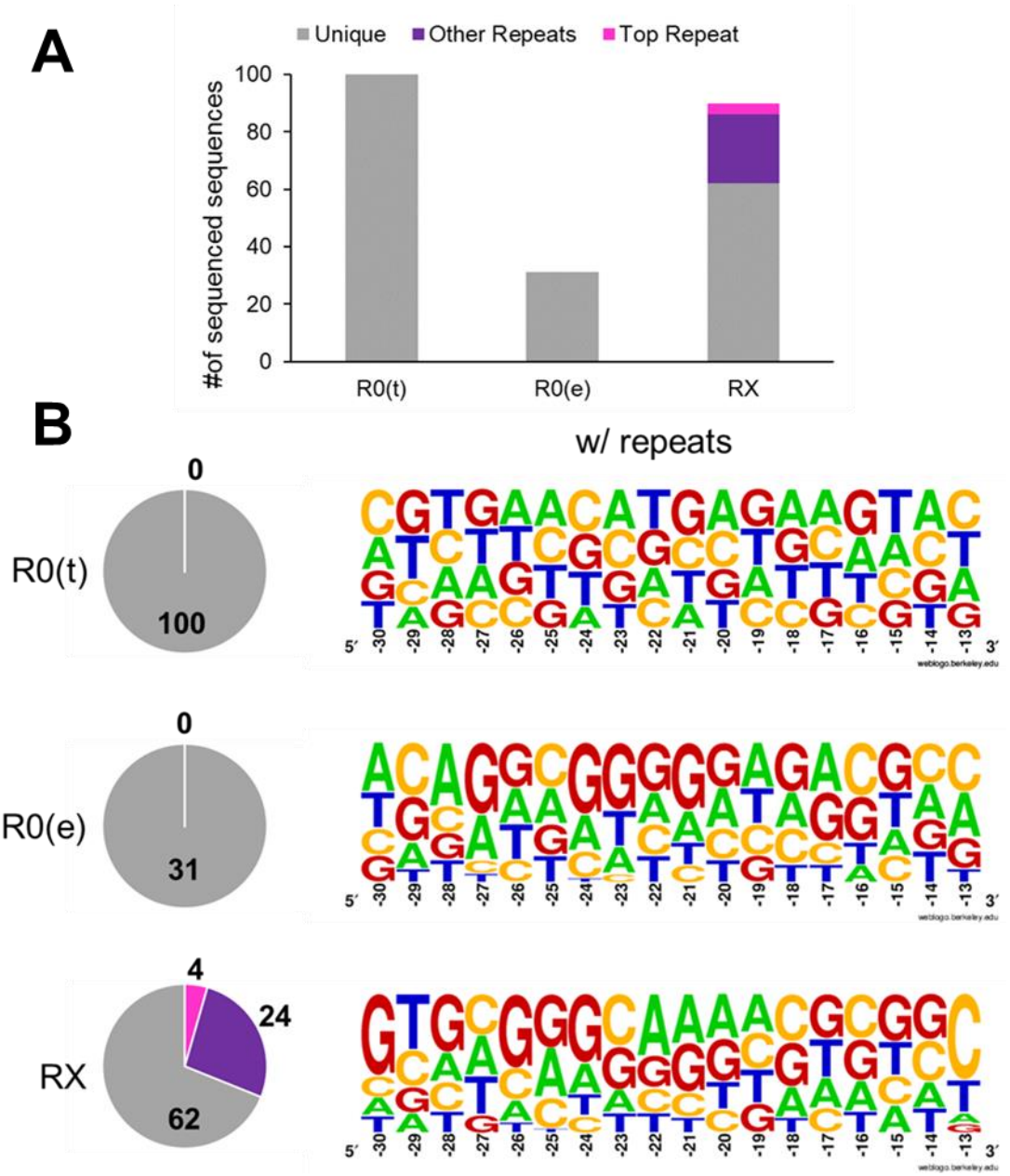

Figure 9. In Vivo Selection Summary

A) Bar graph depicting the number and uniqueness of sequences found during sequencing of in vivo selection controls. Sequences only appearing once were classified as unique, twice or more as other repeats, and the sequence that appeared the most as the top repeat. Theoretical sequences were generated randomly in silico and experimental sequences were found by sequencing the original SELEX library before any rounds of in vitro selection with CF.

B) Weblogos depicting the nucleotide frequencies of the randomized regions of sequences found during in vivo selection and controls. Pie charts depicting the number and uniqueness of the sequences found.

3.09 Enriched Structural Preferences In Vivo

Following the same format of analysis as we did for our in vitro selection, we began examining the selected sequences over the entire element. We observed an increase in GC% and bendability as well as a decrease in curvature when factoring in repeated sequences relative to the sequencing of the initial starting library (**Fig. 10A,C,E**). The trends of these results closely match those of our in vitro selection. Then looking at these features on a bp level, we see the highest bendability at positions -16 and -15 or -15 and -14 with and without repeats respectively (**Fig. 10F**). This follows closely with in vitro data especially without repeats. The same story is true for curvature with both the in vivo and in vitro data showing the highest curvature at the upstream end of the sequence (**Fig. 10D**). Interestingly, GC content diverges here and does not match well the in vitro data. In vivo, the GC content is somewhat evenly distributed across all positions, lacking a downstream GC rich patch seen in the in vitro data (**Fig. 10A**). This data indicates that even in an in vivo environment, upstream curvature coupled with downstream bendability as well as increased bendability overall is important for CF binding.

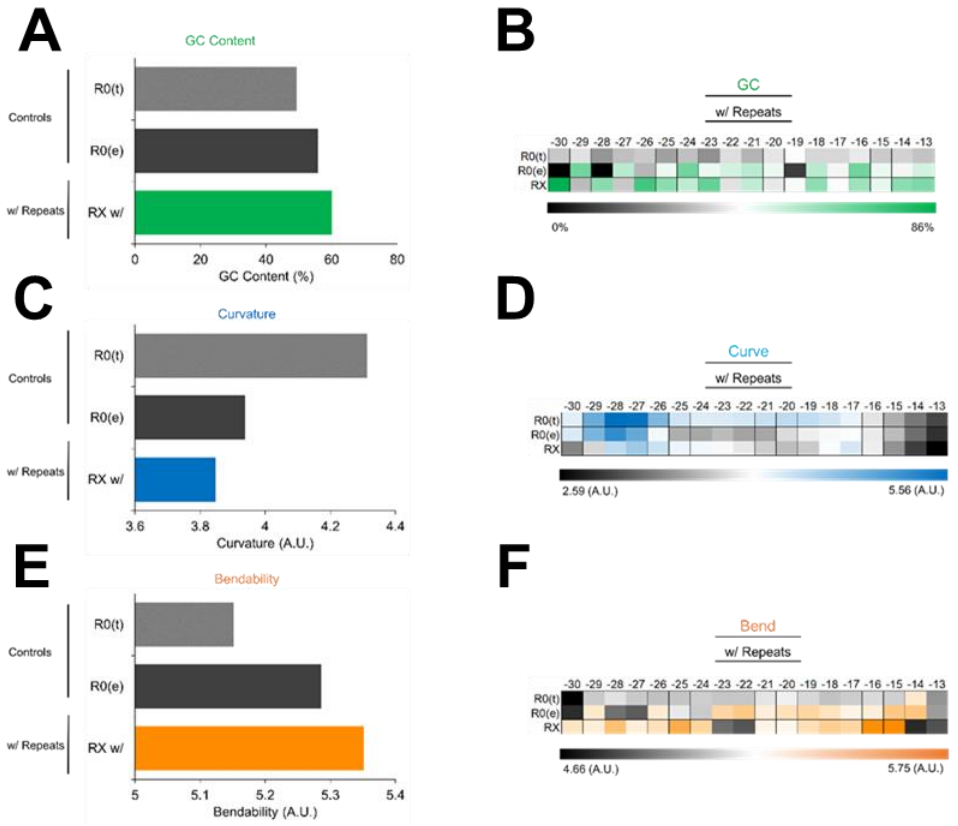

Figure 10. Bendit Analysis of In Vivo Selection Sequences Enriched by Yeast RNA polymerase I CF

The program Bend-it® was used to calculate GC%, bendability, and curvature of DNA sequences with duplicates from in vivo selection. These sequences were compared to theoretical (R0(t)) and experimental (R0(E)) controls. The theoretical control consists of 500 random 18mer sequences generated *in silico* and the experimental control of 30 sequences directly cloned and sequenced from the starting library. Bendability and Curvature were calculated in Bend-it® using a sliding trimer window to determine these values at any given central base pair. Bendability refers to the ability of a DNA sequence to be deformed or change shape and occupy a range of conformations. Curvature refers to more static bends over a series of base pairs. The values calculated for bendability and curvature are relative arbitrary units (A.U.) based on DNaseI, nucleosome positioning, crystal structure, and other experimental data. The three parameters of GC%, bendability, and curvature were calculated both as an average over the entire randomized region of selected sequences as well as at each position individually.

A, C, E) Bar graphs of the GC%, curvature, and bendability of sequences averaged over the entire randomized region of the selected sequences from positions -30 to -13 upstream of the transcription start site with duplicates.

B, D, F) Heatmaps of the GC%, curvature, and bendability averaged at each base pair of the entire randomized region of the selected sequences from positions -30 to -13 upstream of the TSS, with duplicates.

We also again looked at the additional structural features such as helix twist, propeller twist, roll, and minor groove width. Looking at the averages of these features over the entire CE, we see a decrease in helix twist relative to the control with the lowest twist found when including repeats, following the same trend as the in vitro data (**Fig. 11A**). We also saw an increase in minor groove width same as the in vitro data (**Fig. 11C**). Looking at propeller twist we also saw a clear decrease which matches the trend of the in vitro data (**Fig. 11E**). We also see a decrease in roll over the experimental control (**Fig. 11G**). On a base pair level, helix twist is the highest at position -29 the same as the in vitro data shows both with and without repeats (**Fig. 11B**). The data for MGW in vivo diverges from that of in vitro having the highest groove widening occur at the downstream end whereas for in vitro, some of the smallest minor groove values are found at the downstream end (**Fig. 11D**). Propeller twist in vivo also varies from in vitro with the highest propeller twist occurring at position -26 both with and without repeats whereas in vitro the highest values are found in a clear cluster at the downstream end of the sequence (**Fig. 11F**). Finally, looking at roll, we see a scattering of high values throughout the sequence similar to that of the in vitro data (**Fig. 11H**).

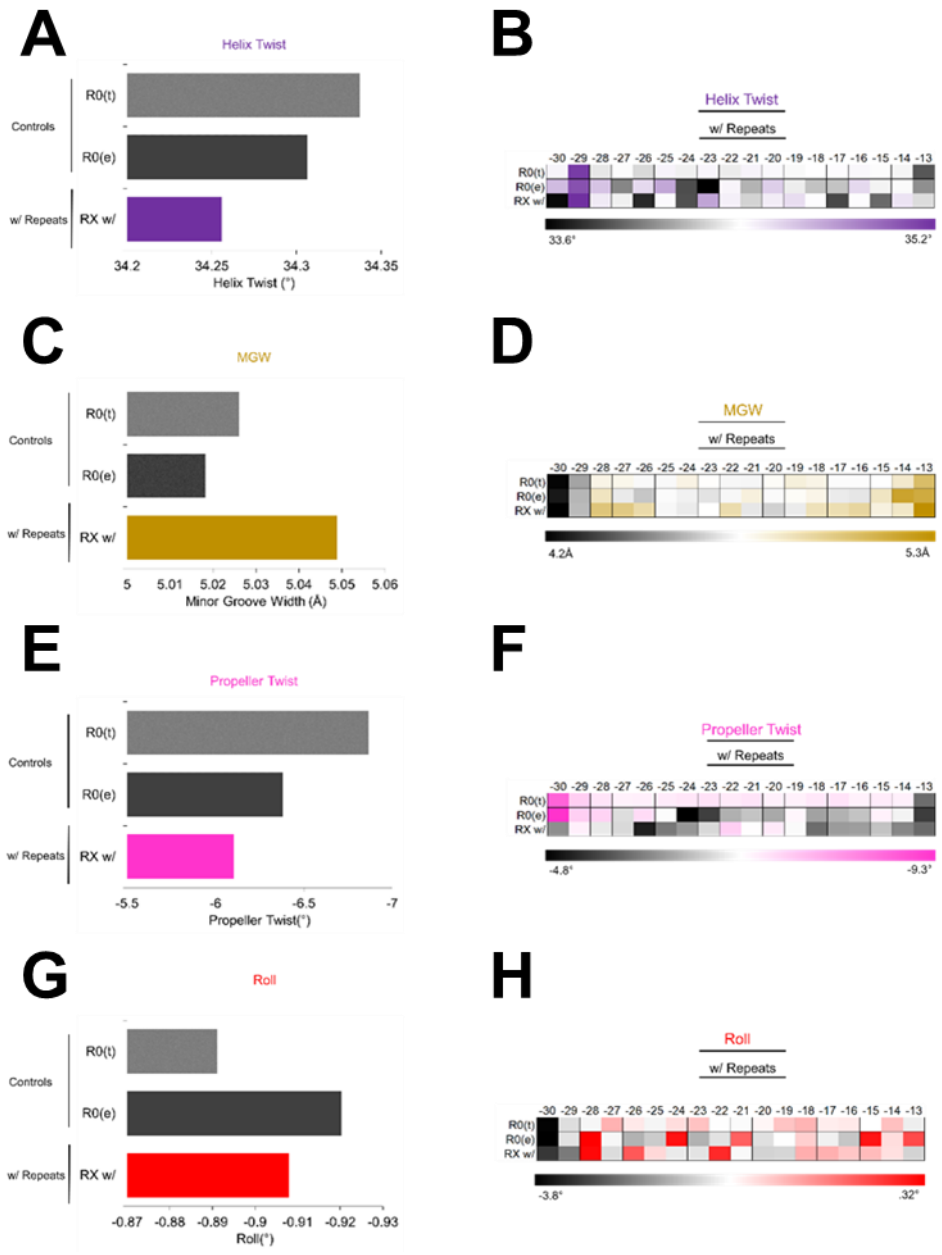

Figure 11. GB shape analysis of in vivo selection sequences enriched by yeast RNA polymerase I CF

The genome browser database for DNA shape annotations GB shape was used to calculate the Helical Twist, Propeller Twist, Minor Groove Width, and Roll of DNA sequences with duplicates of sequences of *in vivo* selection. These sequences were compared to theoretical (R0(t)) and experimental (R0(E)) controls. The theoretical control consists of 500 random 18mer sequences generated *in silico* and the experimental control of 30 sequences directly cloned and sequenced from the starting library. The DNA shape annotations were derived with a high-throughput method for DNA shape predictions and constitute the whole-genome complement to a motif database of

transcription factor binding sites. All four DNA shape parameters were calculated both as an average over the entire randomized region of selected sequences as well as at each position individually.

A, C, E, G) Bar graphs of the Helical Twist, Propeller Twist, Minor Groove Width, and Roll of sequences averaged over the entire randomized region of the selected sequences from positions -30 to -13 upstream of the transcription start site with duplicates.

B, D, F, H) Heatmaps of the Helical Twist, Propeller Twist, Minor Groove Width, and Roll averaged at each base pair of the entire randomized region of the selected sequences from positions -30 to -13 upstream of the TSS, with duplicates.

3.10 Two classes of In Vivo Sequences

We classified the sequences found from our in vivo selection the same way we did for our in vitro SELEX with the exception that we did not further break up rigid and flexible grouping into subgroups. We again first looked at the GC content, bendability, and curvature of these two groupings as a whole. Just as we saw before in vitro, we saw flexible sequences had more bendability (**Fig. 12E**). We also saw that rigid sequences had more curvature than their flexible counterparts (**Fig. 12C**). Lastly, we saw that Flexible sequences had more GC content than rigid ones (**Fig. 12A**). Looking at these properties at a base pair level, we see a relatively even distribution of GC content throughout the CE (**Fig. 12B**). For curvature, the rigid grouping had the most curvature which was localized to positions -29 to -26 and the flexible grouping has very little curvature at any positions (**Fig. 12D**). Bendability was highest in the flexible grouping and found at positions -18 to -14. The rigid grouping predictably had low bendability at mot positions but particularly low bendability at positions -22 and -21 (**Fig. 12F**).

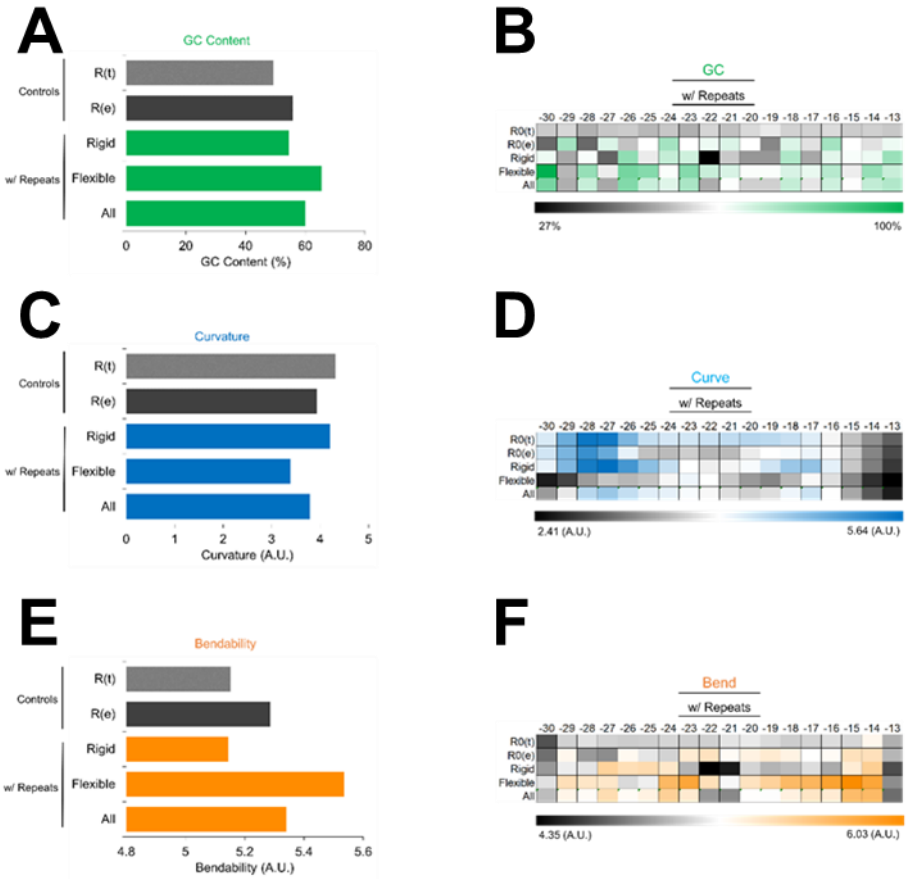

Figure 12. Bendit Analysis of In Vivo Selection Sequences Enriched by Yeast RNA Polymerase I CF Grouped into Rigid and Flexible Categories by Bendability

The program Bend-it® was used to calculate GC%, bendability, and curvature of DNA sequences with duplicates from in vivo selection grouped into flexible categories based on bendability. These sequences were compared to theoretical (R0(t)) and experimental (R0(E)) controls. The theoretical control consists of 500 random 18mer sequences generated *in silico* and the experimental control of 30 sequences directly cloned and sequenced from the starting library. Bendability and Curvature were calculated in Bend-it® using a sliding trimer window to determine these values at any given central base pair. Bendability refers to the ability of a DNA sequence to be deformed or change shape and occupy a range of conformations. Curvature refers to more static bends over a series of base pairs. The values calculated for bendability and curvature are relative arbitrary units (A.U.) based on DNaseI, nucleosome positioning, crystal structure, and other experimental data. The three parameters of GC%, bendability, and curvature were calculated both as an average over the entire randomized region of selected sequences as well as at each position individually.

A, C, E) Bar graphs of the GC%, curvature, and bendability of sequences averaged over the entire randomized region of the selected sequences from positions -30 to -13 upstream of the transcription start site with duplicates.

B, D, F) Heatmaps of the GC%, curvature, and bendability averaged at each base pair of the entire randomized region of the selected sequences from positions -30 to -13 upstream of the TSS, with duplicates.

Once more, we looked at the properties of helix twist, propeller twist, minor groove width, and roll of these groupings. Starting with helix twist, it is readily apparent that the flexible groupings with and without repeats have less twist than the rigid grouping (**Fig. 13A**). It is also apparent the rigid group is more twisted and the flexible one less twisted than the controls. Looking at propeller twist, there is a clear trend of the rigid group being more twisted than the flexible group (**Fig. 13E**). This matches up well with the in vitro data as well as the idea that stereochemical locking and reduced slide from increased propeller twist affects bendability. For minor groove width we saw the flexible group had a significantly larger minor groove width than the rigid grouping (**Fig. 13C**).

Examining roll, we see the flexible group has significantly lower roll than the rigid group (**Fig. 13G**). Since roll opens base pairs towards the minor groove by bending the DNA away from the minor groove, this may indicate that the rigid sequences are locked in a more bent conformation and indeed this fits well with the increased curvature seen of the rigid group over the bendable one. Looking at these properties at a base pair level, helix twist remained relatively low across all groups (**Fig. 13B**). A region of higher MGW could be found -28 to -26 of both rigid and flexible groups and the flexible grouping had the highest MGW at positions -18 and -17 (**Fig. 13D**). Propeller twist was highest in the rigid grouping especially at positions -22 to -20 and remained low across the flexible grouping (**Fig. 13F**). Lastly, higher roll was distributed in scattered patches across both groups with higher roll being shared -22 (**Fig. 13H**). The trends of the in vivo data closely match our in vitro data and further confirm their validity.

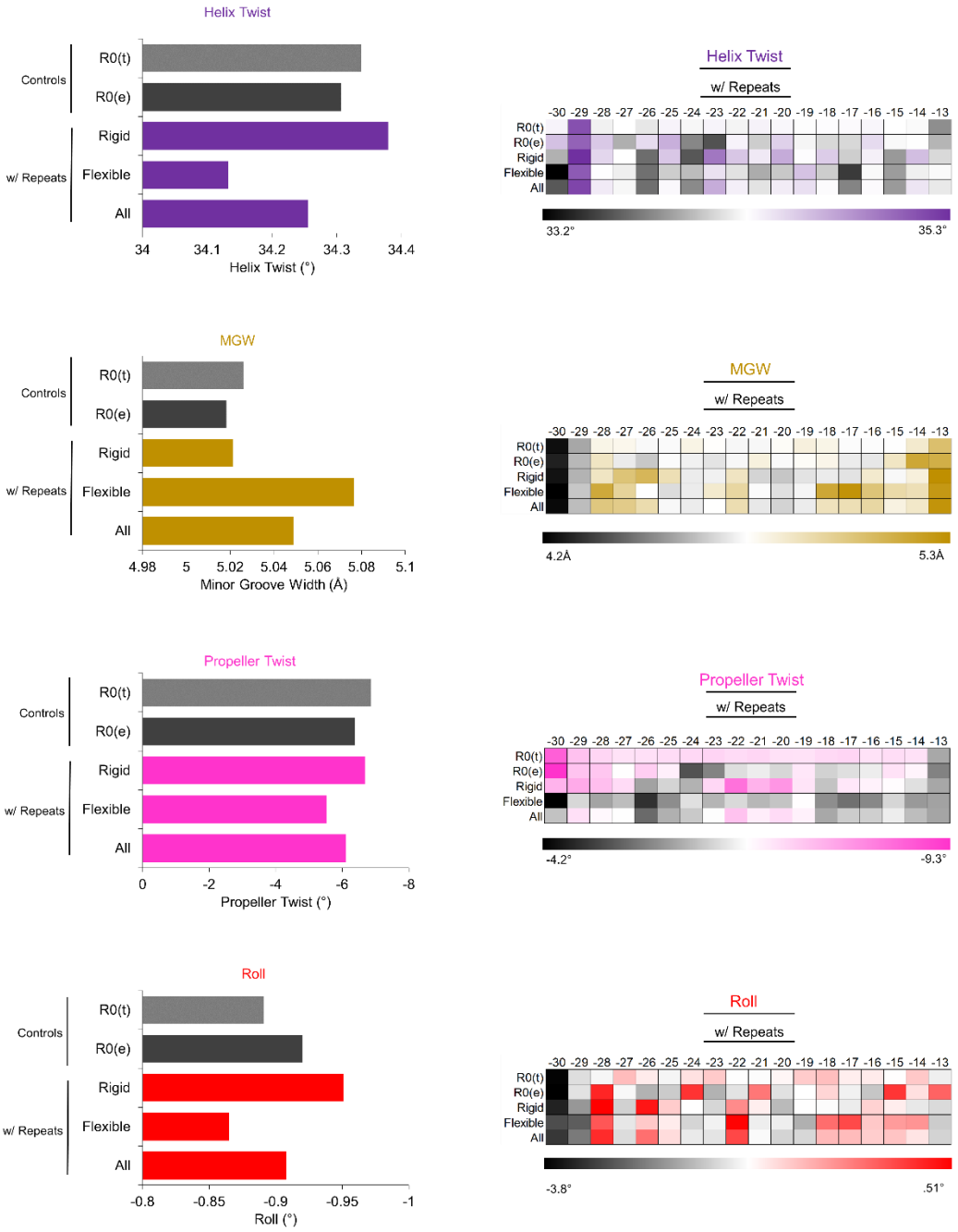

Figure 13. GB Shape Analysis of In Vivo Selection Sequences Enriched by Yeast RNA Polymerase I CF Grouped into Rigid and Flexible Categories by Bendability

The genome browser database for DNA shape annotations GB shape was used to calculate the Helical Twist, Propeller Twist, Minor Groove Width, and Roll of DNA sequences with and without duplicates of round 7 sequences of *in vitro* SELEX grouped into flexible categories based on bendability. These sequences were compared to theoretical (R0(t)) and experimental (R0(E)) controls. The theoretical control consists of 500 random 18mer sequences generated *in silico* and the experimental control of 30 sequences directly cloned and sequenced from the starting library. The DNA shape annotations were derived with a high-throughput method for DNA shape predictions and

constitute the whole-genome complement to a motif database of transcription factor binding sites. All four DNA shape parameters were calculated both as an average over the entire randomized region of selected sequences as well as at each position individually.

A, C, E, G) Bar graphs of the Helical Twist, Propeller Twist, Minor Groove Width, and Roll of sequences averaged over the entire randomized region of the selected sequences from positions -30 to -13 upstream of the transcription start site with duplicates.

B, D, F, H) Heatmaps of the Helical Twist, Propeller Twist, Minor Groove Width, and Roll averaged at each base pair of the entire randomized region of the selected sequences from positions -30 to -13 upstream of the TSS, with duplicates.

Discussion

In this chapter, we have further characterized and revealed the specific structural DNA binding preferences of yeast RNA Polymerase I CF. We found that CF is able to tolerate single bp mutations with little to no loss in competition across the CE except in the rigid AT patch. We also found that losses in competition in this region highly correlated with changes in DNA structure. Additionally, we characterized the structural properties of novel randomized CE sequences selected by CF in vitro and in vivo finding a lack of sequence conservation but conservation of specific structural features in both. Finally, we found these selected sequences fell into two distinct structural categories of rigid and flexible. Overall, we have shown that CF relies heavily on structure to recognize the CE.

Oftentimes, specific base pairs or properties of a DNA binding site for a protein can be identified as more important whether for a sequence or structure-based reason. Previously, much work has been done to help define the boundaries of CE required for CF binding but there exists less data on what bps or properties of the CE may be more

important for CF binding and why. Here, we utilized single bp mutations across the minimal region of the CE in yeast to examine experimentally the specific bps and regions of importance for CF binding and determine why these regions are important. In agreement with the importance of DNA structure for CF binding, we found that CF was able to tolerate most single bp mutations for much of the CE, indicating that a specific sequence at these positions is of less importance. However, we found that mutations to the AT rich patch in the CE sequence led to drastic reductions in CF binding. These losses in competition correlated particularly well with the three properties of bendability, curvature, MGW, and HelT. Increases in all four properties in this region correlated strongly with decreases in CF binding. For bendability and curvature, this would indicate that a rigid patch that perhaps is more locked into an ideal conformation is preferred by CF for binding.

For MGW and HelT, the same effect was seen with the orthologue TIF-IB in *Acanthamoeba castellani*. In an analysis by Marilley et al, widening of the minor groove via single bp mutations at the four most narrow sites of the core promoter resulted in significantly decreased transcriptional activity but narrowing at these sites increased transcriptional activity. This matched quite well with what we saw occur in the AT rich patch. Not only did we see that increases in MGW decreased CF binding but that MGW in the AT rich patch was the narrowest in the CE(not shown in figures). Favoring of a narrow minor groove is further illustrated by bends both intrinsic and induced in the core promoter. Across many rDNA promoters, there is a conserved bend around 25bps upstream of the TSS and both TIF-IB and CF further bends the rDNA upon binding. TIF-IB bends then DNA ~45° near -23 and CF by ~35° and ~45° near -21 and -16

respectively. According to modeling by Marilley et al, these inherent and additional bends by TIF-IB to the core promoter narrow the minor groove further indicating a narrow minor groove is desired at least for part of the CE. Given the similarities between TIF-IB and CF it is likely that the same occurs with CF. Additionally, these bends affect the property of helix twist. According to their model and some experimental data, these bends and narrowing of the minor groove result in unwinding and negative supercoiling of the helix downstream. Furthermore, this unwinding led to decreased thermal stability. This of course would aid in the ATP independent melting and opening of DNA at the TSS for Pol I. Our data corroborates this with our observation of decreased CF binding correlating with increased helical twist.

It has long been well documented that across eukaryotic species, there is very little sequence conservation on the rDNA promoter. Instead, there is a distinct conservation of structural features across rDNA promoters such as an intrinsic bend in the DNA around 25bp upstream of the TSS. This would explain how the hCPE is able to function in yeast albeit in a positionally dependent manner as they have very similar profiles of bendability and curvature. Similar shifts in position have also been shown to allow mouse Pol I factors to transcribe frog rDNA and further help to explain how species specificity and differentiation is possible despite conserved structural features. This emphasis on the importance of structure in Pol I factor recognition of the rDNA promoter matched well with our results from our in vitro and in vivo selection assays. Both of these theoretically gave CF all possible combinations of sequence to choose from, allowing us to examine conservation of sequence and or structure that wouldn't be possible with single bp mutations. Carrying out selection both in vitro and in vivo

enabled us to examine what role native conditions in yeast may play in CF recognition of the rDNA promoter in comparison to CF alone. While there were slight differences between our in vitro and in vivo data, overall, we saw the same conservation and trends of structural features selected by CF.

We first looked at the properties of GC%, curvature, and bendability both in vitro and in vivo over the entire element. In this case, bendability stands for the ability of the DNA sequence to be deformed. A more easily deformed sequence or a sequence that takes less energy to deviate from the ideal B form than another is considered more bendable. One factor that plays a critical role in the bendability of DNA is base stacking. Base stacking interactions are primarily electrostatic and there is a direct correlation between the stacking area of a base step and stiffness. The larger the stacking area or overlap between base pairs in a base step, the more interactions and stability gained and the stiffer it is. We observed increased GC% and bendability and decreased curvature of selected sequences over controls. In particular, the increased GC% correlating with increased bendability was surprising as GC base pairs are generally regarded as less bendable than AT base pairs. However, the context of neighboring base pairs and base steps play a significant role in the bendability of any given sequence. For example, Poly dA motifs are associated with very low bendability but (TpA)*(TpA) steps with high bendability and GpG is considered one of the more bendable dinucleotides. GpG happened to be the most enriched dinucleotide from our selection(in vitro) and would explain why we see the increased bendability of sequences selected by CF. For curvature, we saw a clear decrease. It seems that overall, CF is favoring more bendable and less intrinsically bent or curved sequences.

At a base pair level, the downstream region of highest bendability correlates with the highest region of GC% confirming what we saw over the element overall in vitro. However, in vivo, GC enrichment appears to be more scattered without a clear downstream patch. This is likely due to the very different environmental conditions that are found in vivo relative to in vitro and require further investigation. Interestingly, we saw that curvature was mostly constrained to the upstream end. This makes sense as the more bendable the DNA, the less likely it is to be intrinsically curved or bent and fixed in any given conformation. Overall, our results indicate CF prefers a more bendable and GC rich sequence.

Looking further at bendability, we were able to classify all sequences into two distinct groups of rigid and flexible. Rigid sequences possessed a dip in bendability or rigid patch similar to WT and flexible sequences did not. There may be two competing strategies at play here for why we saw these two distinct groupings. One strategy is that the DNA is recognized in its pre-bent form already close to the final conformation it will occupy upon binding, with less bending being done by the protein itself. Another scenario could be the opposite case where the initial conformation of the DNA is vastly different than the bound form with significant bending done by the protein with it relying more on the intrinsic bendability of the DNA for recognition and binding. There are thermodynamic trade-offs to each method with low entropic costs of binding corresponding with high enthalpic costs and vice versa. For example, if a DNA sequence doesn't require much alteration to its conformation upon binding to protein, this interaction would be advantageous for enthalpy but have some entropic cost. If a protein is significantly bending the DNA from its free conformation, the entropic cost becomes

very high and as a result the change in enthalpy must compensate with more bonds formed between the DNA and protein than in the previous example. Overall, we noticed that significantly more sequences fell into the flexible category indicating that CF would rather rely on the intrinsic flexibility of the DNA than a rigid WT like bendability profile. Having increased bendability would help to offset the higher entropic cost of molding the DNA to a desired conformation and it appears that bendability overall, is a driving force of CF-DNA binding.

Looking at the properties of MGW, Roll, HelT, and ProT, the trends were generally less clear. As we observed with the single bp mutations, the importance of some of these properties in the CE is positionally dependent. So, averaging these properties across the entire CE could obscure their importance. Only HelT showed a consistent decrease across the element, whether or not repeats were included. Examining HelT at a bp level, we observed the greatest decrease at the downstream end of the randomized region. This ties in well with the observations from Marilley et al that reduced helix twist at this region allows for easier melting and transcription initiation at the TSS. As for MGW, any widening is scattered across the element in patches some of which overlap where the WT rigid AT patch would normally be and is in contrast to our single bp mutant data. It's likely there is some sort of compensation occurring when randomizing the entire CE that is allowing for these differences and could benefit from further investigation. ProT decreased overall with much of it contained at the downstream end of the element although less so in vivo which allows for sliding of base pairs and contributes to the bendability of the DNA correlating well with the highest increases in GC% that are found here. Lastly, roll is scattered similarly to MGW. It is

clear that downstream decreases in HelT and ProT of the CE are preferred by CF, but the more investigation is needed to pick apart and identify the interplay of MGW, Roll, HelT, and ProT and how changes in one can be compensated by another to facilitate CF binding.

CF both in vitro and in vivo relies upon very similar structural features of the CE to recognize and bind to it. In addition, the structural binding characteristics we described of CF closely resemble those of TIF-1B and even SL1 as found in previous literature. This shared structural binding mechanism between these orthologous factors along with our consistent characterization of CF, strongly indicate our findings could have future applications in therapeutic attempts to regulate Pol I transcription in cancers and ribosomopathies.

Material and Methods

PCR Synthesis of infrared-labeled EMSA probes

Infrared-labeled DNA probes were synthesized according to the procedure found in Jackobel et al, 2019

Infrared labeling of EMSA probes

Infrared labeling of EMSA probes followed the procedure found in Jacobel et al, 2019 with the following modification. The strand containing the 5' C6 amino ester modification was incubated with dye for 16 hr protected from sunlight.

Electrophoretic Mobility Shift Assay(EMSA)

Samples were prepared with 100ng of infrared labeled DNA probe and .07ug of CF protein in gel shift buffer (20 mM Tris-HCl pH 8.0, 60 mM KCl, 5% Glycerol, 5mM MgCl2), and H2O to a total reaction volume of 30 uL. CF and DNA probe were incubated at 25°C for 45min shaking at 400rpm in an Eppendorf Thermomixer F1.5. 15uL of reaction was run on a precast 5%TBE Gel from Bio-Rad in 1X TBE (.089 M Tris, .089 M Boric acid, and .002 M EDTA) for 40 min at 150v. The gel was imaged using a LiCOR Odyssey FC scanner.

Competition Assay of Selected In vitro and In Vivo Sequences

Competitor oligos were generated by annealing the top and bottom strands according to this procedure using a Bio-Rad C1000 Touch™ thermocycler: 95ºC for 2min, [95ºC for 1min decreasing by ºC each cycle] for 83 cycles. Reagents were added as follows. Equimolar amounts of the top and bottom strand were combined in annealing buffer(10mM Tris, pH 7.5-8, 50mM NaCl, 1mM EDTA) and H2O to a final volume of 50uL. EMSA were performed as above with the following modifications. DNA probe was preincubated with competitor for 30min prior to addition of CF. In a 30uL reaction, 40mM of annealed competitor oligo(referred as 1X) was incubated with 30ng of probe in gel shift buffer were mixed and incubated for 30min, followed by a 45min incubation after the addition of .4ug of CF.

Competition Assay of Single bp mutant oligos

Single base pair mutant oligos of the WT CE sequence from positions -30 to -15 were generated the same way as described above. Base pair mutants consisted of A, T, C, G, and I. EMSA were performed as above with the following modifications. In a 15uL reaction volume, 20mM of annealed competitor was incubated with 15ng of probe. .025ng of CF was used.

In Vivo Selection Library

The In Vivo selection library was designed using a variation of QuikChange Megaprimer mutagenesis [1]. Random sequences were inserted into pMF150, a

derivative of pNOY373 [2] with the -28 to -17 region of the promoter deleted, such that the sequences replaced the deleted region of the promoter. PCR was conducted using KODX Hot Start polymerase(Millipore). Forward Primer- InvivoSelex-F1:CATGGAGTACAANNNNNNNNNNNNNNNNNNGGGAGGTACTTCATGCGAAACG, Reverse Primer 1- P1pro-invivo-R1: GTATAGAGACTAGGC AGATCTGAC, Reverse Primer 2- P1pro-invivo-R2:TAGCGACTCTCCACCGTTTG AC,Reverse Primer 3- P1pro-invivo-R3: TTCCCAAATTGT ATCTCTTCAATAC. The PCR was conducted as follows: 94°C for 2 minutes, [98°C for 10°C seconds, 55°C for 30 seconds, 68°C for 45 seconds] for 17 cycles, followed by [98°C for 10 seconds, 68°C for 15 minutes] for 17 cycles. Reagents were added as follows for a total volume of 50uL: 25uL of 2X Xtreme buffer from Sigma-Aldrich, (need conc of plasmid), 30nM of FP and RP, .4mM DNTP, .7 U KOD Xtreme™ Hot Start DNA Polymerase from Sigma-Aldrich, and H2O to final reaction volume. The resulting PCR product was digested using DPN1 for 4 hours. DH10b *Escherichia Coli*[3] were transformed with resultant plasmid and grown on LB plates containing ampicillin. Colonies were selected, grown overnight in LB with ampicillin, and subject to plasmid isolation using the E.N.Z.A Plasmid isolation kit (Omega Bio-Tek), following the manufacturers protocol.

In Vivo Selection/Yeast Growth Assay and Plasmid Isolation

The strain of *S. cerevisiae* used for this study *in vivo*, was rdnDD. This strain has the chromosomal rDNA deleted and moved to a plasmid, for genetic manipulation. Additionally this vector has a URA3 marker that will poison the cell when in contact with 5-Fluoroorotic Acid (5 FOA)[2]. All yeast growth occurred in a 30°C incubator and took

approximately 3 days to grow. Frequently used plasmids in this study were pMF150(derivative of pNOY373), pCC150 (Wild-Type), and pRS425 (Parent Empty Vector). The media used to grow up yeast cells, was Glucose Complete without Leucine (GC-L). Plates of media used for selectively isolating our plasmid were, GC-L, and GC-L+FOA. The isolated plasmids from the mutant CE library were plated to GC-L. The colonies that grew on these plates were replica plated to GC-L+FOA plates and incubated. Colonies that grew on FOA were able to kick out the original vector from the *rdn*DD strain and take in the vector from the QuikChange PCR. These colonies were then plated back to GC-L, and grown in GC-L liquid culture overnight. These cultures were then subject to plasmid isolation. Plasmid isolation of yeast cells was conducted by first breaking open the cells with zymolase. After adding zymolase to the culture, 50 mM EDTA was added to control the activity of nucleases present in the cell. Next the E.N.Z.A. Plasmid isolation kit (Omega Bio-Tek), was used following the manufacturers protocol.

The theoretical control consisted of random sequences generated in silico and the experimental control of sequences directly cloned and sequenced from the starting SELEX library.

In Vitro SELEX Library Design

The In Vitro SELEX library was designed using WT CE from positions -58 to +15 relative to the TSS with an 18bp randomized region from -30 to -13(GTAAAACGACGGCCAGCATGGAGTACAANNNNNNNNNNNNNNNNNGGGAGGTACTTCCATGGTCATAGCTGTT). Previous studies have indicated that CF

requires a minimal 12bp region from positions -28 to -17 in order to bind to the CE. Based on previous competition assays that showed a sharper drop off in competition on the 5' side as bases of the WT CE were removed compared to the 3' end, we chose to extend the minimal 12bp region by 2 bp on the 5' side and by 4bp on the 3' side. The DNA oligomers were ordered from Sigma Aldrich.

CF Purification

Core Factor (CF) was expressed as previously described [46,48] with the following modifications. Briefly, CF was expressed from the pET-Duet CF vector containing His6-Rrn7-Rrn11-His6-Rrn6 in LOBSTRBL21 (DE3)-RIL Escherichia coli cells. Recombinant CF protein was expressed in Autoinducing Terrific Broth (0.024% w/v tryptone, 0.048% yeast extract w/v, 0.4% v/v glycerol, 17 mM KH2PO4, and 72mM K2HPO4) supplemented with 20 mL per liter 50X5052 (25% v/v glycerol, 2.5% w/v glucose, and 10% w/v alpha lactose monohydrate) and 2mM MgSO4. Inoculated media was grown to an OD600 of 0.6 then shifted to 20 °C overnight. Cells were harvested by centrifugation, pellets were washed in Tris-buffered saline (50mM Tris-HCl pH 7.6, 150mM NaCl) supplemented with 1X PMSF and 1mM DTT and stored at−80 °C. Cells were thawed and resuspended in 5 mL per gram of Extraction Buffer (50mM HEPES pH 8.0, 500mM KCl, 10mM Imidazole, 5mM MgCl2, 0.1mM EDTA, 20% glycerol) and supplemented with 1X PMSF and 1mM DTT. Lysozyme was added to resuspended cells at 1 mg/mL and incubated on ice for 30 min. The cells were lysed by sonication using a Branson Sonifier 450 (VWR Scientific). The extract was clarified by centrifugation at 4 °C for 15 min at 5,000rpm. The clarified extract was added to Ni-NTA Sepharose beads (Biotool) washed in Extraction Buffer and incubated at 4 °C overnight in batch. Protein

bound beads were washed three times with high salt Wash Buffer (Extraction Buffer but with 1M KCl) and three times with low salt Wash Buffer (Extraction Buffer but with 200mM KCl). Bound proteins were eluted with 2-3 bead volumes of Elution Buffer (50mM HEPES pH 8.0, 200mM KCl, 200mM Imidazole, 5mM MgCl2, 0.1mM EDTA, 20% glycerol). Eluted CF was then further purified over a HiTrap Heparin HP column (GE Healthcare) using a linear gradient of Buffer A (50mM HEPES pH 8.0, 200mM KCl, 5mM MgCl2, 0.1mM EDTA, 5% glycerol) to Buffer B (Buffer A with 1M KCl) over 10 column volumes. CF was eluted between 800 and 1000mM KCl. Peak fractions were pooled, desalted in Buffer C (50mM HEPES pH 8.0, 0.1mM EDTA, 5% glycerol). CF was then further purified over a HiTrap Q HP column (GE Healthcare) using a linear gradient of Buffer A to Buffer B (as described above). Protein eluted between 400 and 600mM KCl. Peak fractions were pooled and diluted in Buffer D (50mM HEPES pH 8.0, 200mM KCl, 5% glycerol, 0.1mM EDTA, 0.1% Tween-20) to 0.1 mg/mL, aliquoted and stored at −80 °C.

In Vitro SELEX

The novel CE sequences were selected using in vitro SELEX. First, the ds DNA library was generated by annealing the top and bottom strands according to this procedure using a Bio-Rad C1000 Touch™ thermocycler: 95ºC for 2min, [95ºC for 1min decreasing by ºC each cycle] for 83 cycles. Reagents were added as follows: The annealed DNA was then diluted to 400ng/uL. 400ng of DNA was incubated with a titration of core factor protein between .25ng and 3ng each round, 15uL of gel shift buffer (20mM Tris-HCl pH 8.0, 60mM KCl, 5% Glycerol, 5mM MgCl2) and H2O for a total

volume of 30uL in a 1.5mL tube at room temp at 400rpm for 45min. An EMSA was conducted with 15uL of each sample loaded into a well of a 3% TBE gel made with certified™ low range ultra agarose from Bio-Rad and run at 90v for 90 minutes. The gel was then rinsed with DI H2O and stained in 50mL of TBE with 5uL of 10,000X in H_2O GelRed® stain from Biotium on a shaker at 100rpm for 30 minutes. The gel was then de-stained in DI H2O for 10 minutes at 100rpm on a shaker. The gel was visualized using a LiCOR Odyssey FC scanner and the shifted CF-CE complex band was excised using a UV illuminator. The excised gel fragment was nebulized using the ULTRAFREE® DNA Extraction from Agarose kit from Millipore. The nebulized DNA was then amplified via asymmetric PCR according to this procedure: 94ºC for 2 min, [98ºC for 10 s, 67ºC for 30 s, 68ºC for 15 s] for 17 cycles, and then a final 68ºC extension for 2min. Reagents were added as follows for a total volume of 30uL: 15uL 2X Xtreme buffer from Sigma-Aldrich, .6uL template DNA, 30 nM of FP and RP each, 0.4 mM DNTP, 60 mM TMAC, 3% DMSO, 0.5 mg BSA, and 0.42 U KOD Xtreme™ Hot Start DNA Polymerase from Sigma-Aldrich, and H2O to final reaction volume. This PCR procedure was run twice, first with only the forward primer (GTAAAACGACGGCCAGCATGGAGTA) and again with the reverse primer(AACAGCTATGACCATGGAAGTA) added to the reaction. The DNA was then ethanol precipitated. The dried pellet was resuspended in 40uL of EB. This amplified DNA was titrated up against a 400ng/uL control sample titrated down to determine concentration. The concentration determined the titration of CF to add to each lane of the EMSA. The titration of CF was shifted down each successive round of SELEX.

The theoretical control consisted of random sequences generated in silico and the experimental control of sequences directly cloned and sequenced from the starting SELEX library.

Sequencing of In Vitro Sequences

The PRS 425 plasmid was selected as the vector for inserting the novel CE sequences into. The insertion site was shortened from 300bp to 80bp to improve insertional mutagenesis cloning efficiency by the following overlapping mutagenic primers (FP-5'sequence-3'; RP-5'seq-3;). Quick Change PCR was done according to this procedure: 94º for 2min, [98º for 10sec, 50ºC for 30sec, 68ºC for 7 minutes] for 17 cycles, then held at 12ºC. Reagents were added as follows for a total volume of 50 uL: 25uL buffer, 100 ng plasmid template DNA, 0.7 mM DNTP, 0.7 U KOD Xtreme™ Hot Start DNA Polymerase from Sigma-Aldrich. Water and template DNA were adjusted to increase transformation efficiency on a round by round basis. The PRS 425 vector was transformed into dh10b cells and grown on LB plates containing ampicillin. Colonies were selected, grown overnight in LB with ampicillin and Plasmid DNA was extracted using the E.Z.N.A.® Plasmid Mini Kit from Omega Bio-tek and protocol with the following changes. A vacuum manifold was used for all steps requiring a centrifuge except for the final drying and elution steps and HBC buffer was not used. McLab sequencing services were used with the Selex Seq F2 primer GATGTGCTGCAAGGCGATTAAGTT.

Sequence Analysis

Sequences from McLab were extracted and aligned via the flanking regions in Excel and screened for repeated sequences. The sequences were then analyzed for curvature, bendability, and GC content using the bend.it® program by Jedlik laboratories[69-74]. We developed an environment and pipeline stitched together with Jupyter and Python, and served via MyBinder.org, featuring the standalone program from Vlahovicek et al. (2003) at the core to calculate values for bendability and curvature of the sequences in a high-throughput manner, see https://github.com/fomightez/bendit-binder. Windows of length of three with consensus scale option for the curvature parameter were the specific settings for the standalone program with these worked out using the online form at the bend.it Server. Curvature, bendability, and GC content averages were plotted at each base pair as well as over the entire 18bp randomized region. Weblogos were made to visualize the consensuses of each round with and without repeats using the weblogo generator from Berkeley https://weblogo.berkeley.edu/logo.cgi. Analysis of the four structural properties of propeller twist, helical twist, minor groove width, and roll were carried out using DNAshape by the Rohs lab[75-78].

1. Roeder, R.G. and W.J. Rutter, *Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms.* Nature, 1969. **224**(5216): p. 234-7.
2. Goodfellow, S.J. and J.C.B.M. Zomerdijk, *Basic Mechanisms in RNA Polymerase I Transcription of the Ribosomal RNA Genes*, in *Epigenetics: Development and Disease*. 2012, Springer Netherlands: Dordrecht. p. 211-236.
3. Schneider, D.A., *RNA polymerase I activity is regulated at multiple steps in the transcription cycle: recent insights into factors that influence transcription elongation.* Gene, 2012. **493**(2): p. 176-184.
4. Reeder, R.H., *Regulation of RNA polymerase I transcription in yeast and vertebrates.* Progress in Nucleic Acid Research and Molecular Biology, 1999. **62**: p. 293-327.
5. Moss, T., et al., *A housekeeper with power of attorney: the rRNA genes in ribosome biogenesis.* Cell Mol Life Sci, 2007. **64**(1): p. 29-49.
6. Woolford, J.L., Jr. and S.J. Baserga, *Ribosome biogenesis in the yeast Saccharomyces cerevisiae.* Genetics, 2013. **195**(3): p. 643-81.
7. Dixon, M.J., et al., *The gene for Treacher Collins syndrome maps to the long arm of chromosome 5.* American Journal of Human Genetics, 1991. **49**(1): p. 17-22.
8. Williamson, D., et al., *Nascent pre-rRNA overexpression correlates with an adverse prognosis in alveolar rhabdomyosarcoma.* Genes Chromosomes Cancer, 2006. **45**(9): p. 839-45.
9. Dauwerse, J.G., et al., *Mutations in genes encoding subunits of RNA polymerases I and III cause Treacher Collins syndrome.* Nature Genetics, 2011. **43**(1): p. 20-22.
10. Walker-Kopp, N., et al., *Treacher Collins syndrome mutations in Saccharomyces cerevisiae destabilize RNA polymerase I and III complex integrity.* Hum Mol Genet, 2017. **26**(21): p. 4290-4300.
11. Bywater, M.J., et al., *Dysregulation of the basal RNA polymerase transcription apparatus in cancer.* Nat Rev Cancer, 2013. **13**(5): p. 299-314.
12. Bywater, M.J., et al., *Inhibition of RNA polymerase I as a therapeutic strategy to promote cancer-specific activation of p53.* Cancer Cell, 2012. **22**(1): p. 51-65.
13. Drygin, D., et al., *Targeting RNA polymerase I with an oral small molecule CX-5461 inhibits ribosomal RNA synthesis and solid tumor growth.* Cancer Research, 2011. **71**(4): p. 1418-1430.
14. Hannan, K.M., et al., *Dysregulation of RNA polymerase I transcription during disease.* Biochimica et biophysica acta. Gene regulatory mechanisms, 2013. **1829**(3-4): p. 342-360.
15. Hahn, S., *Structure and mechanism of the RNA polymerase II transcription machinery.* Nat Struct Mol Biol, 2004. **11**(5): p. 394-403.
16. Liu, X., et al., *Structure of an RNA polymerase II-TFIIB complex and the transcription initiation mechanism.* Science, 2010. **327**(5962): p. 206-9.
17. Kostrewa, D., et al., *RNA polymerase II-TFIIB structure and mechanism of transcription initiation.* Nature (London), 2009. **462**(7271): p. 323-330.
18. Choe, S.Y., M.C. Schultz, and R.H. Reeder, *In vitro definition of the yeast RNA polymerase I promoter.* Nucleic Acids Res, 1992. **20**(2): p. 279-85.
19. Musters, W., et al., *Linker scanning of the yeast RNA polymerase I promoter.* Nucleic Acids Res, 1989. **17**(23): p. 9661-78.
20. Kulkens, T., et al., *The yeast RNA polymerase I promoter: ribosomal DNA sequences involved in transcription initiation and complex formation in vitro.* Nucleic Acids Res, 1991. **19**(19): p. 5363-70.

21. Aprikian, P., B. Moorefield, and R.H. Reeder, *New Model for the Yeast RNA Polymerase I Transcription Cycle.* Molecular and Cellular Biology, 2001. **21**(15): p. 4847-4855.
22. Blattner, C., et al., *Molecular basis of Rrn3-regulated RNA polymerase I initiation and cell growth.* Genes Dev, 2011. **25**(19): p. 2093-105.
23. Engel, C., et al., *Structural Basis of RNA Polymerase I Transcription Initiation.* Cell, 2017. **169**(1): p. 120-131.e22.
24. Jackobel, A.J., et al., *Breaking the mold: structures of the RNA polymerase I transcription complex reveal a new path for initiation.* Transcription, 2018. **9**(4): p. 255-261.
25. Moorefield, B., E.A. Greene, and R.H. Reeder, *RNA polymerase I transcription factor Rrn3 is functionally conserved between yeast and human.* Proc Natl Acad Sci U S A, 2000. **97**(9): p. 4724-9.
26. Peyroche, G., et al., *The recruitment of RNA polymerase I on rDNA is mediated by the interaction of the A43 subunit with Rrn3.* EMBO J, 2000. **19**(20): p. 5473-82.
27. Sadian, Y., et al., *Structural insights into transcription initiation by yeast RNA polymerase I.* The EMBO journal, 2017. **36**(18): p. 2698-2709.
28. Siddiqi, I., et al., *Role of TATA binding protein (TBP) in yeast ribosomal dna transcription by RNA polymerase I: defects in the dual functions of transcription factor UAF cannot be suppressed by TBP.* Mol Cell Biol, 2001. **21**(7): p. 2292-7.
29. Steffan, J.S., et al., *The role of TBP in rDNA transcription by RNA polymerase I in Saccharomyces cerevisiae: TBP is required for upstream activation factor-dependent recruitment of core factor.* Genes Dev, 1996. **10**(20): p. 2551-63.
30. Bischler, N., et al., *Localization of the yeast RNA polymerase I-specific subunits.* EMBO J, 2002. **21**(15): p. 4136-44.
31. Engel, C., et al., *RNA polymerase I structure and transcription regulation.* Nature, 2013. **502**(7473): p. 650-5.
32. Fernandez-Tornero, C., et al., *Crystal structure of the 14-subunit RNA polymerase I.* Nature, 2013. **502**(7473): p. 644-9.
33. Milkereit, P., P. Schultz, and H. Tschochner, *Resolution of RNA polymerase I into dimers and monomers and their function in transcription.* Biol Chem, 1997. **378**(12): p. 1433-43.
34. Bedwell, G.J., et al., *Efficient transcription by RNA polymerase I using recombinant core factor.* Gene, 2012. **492**(1): p. 94-99.
35. Keener, J., et al., *Reconstitution of yeast RNA polymerase I transcription in vitro from purified components. TATA-binding protein is not required for basal transcription.* J Biol Chem, 1998. **273**(50): p. 33795-802.
36. Pilsl, M., et al., *Structure of the initiation-competent RNA polymerase I and its implication for transcription.* Nat Commun, 2016. **7**: p. 12126.
37. Lin, C.W., et al., *A novel 66-kilodalton protein complexes with Rrn6, Rrn7, and TATA-binding protein to promote polymerase I transcription initiation in Saccharomyces cerevisiae.* Mol Cell Biol, 1996. **16**(11): p. 6436-43.
38. Knutson, B.A., et al., *Architecture of the Saccharomyces cerevisiae RNA polymerase I Core Factor complex.* Nat Struct Mol Biol, 2014. **21**(9): p. 810-6.
39. Lalo, D., et al., *RRN11 encodes the third subunit of the complex containing Rrn6p and Rrn7p that is essential for the initiation of rDNA transcription by yeast RNA polymerase I.* J Biol Chem, 1996. **271**(35): p. 21062-7.
40. Han, Y., et al., *Structural mechanism of ATP-independent transcription initiation by RNA polymerase I.* eLife, 2017. **6**: p. e27414.
41. Sadian, Y., et al., *Molecular insight into RNA polymerase I promoter recognition and promoter melting.* Nature Communications, 2019. **10**(1): p. 1-13.

42. Hahn, S. and S. Roberts, *The zinc ribbon domains of the general transcription factors TFIIB and Brf: conserved functional surfaces but different roles in transcription initiation.* Genes Dev, 2000. **14**(6): p. 719-30.
43. Knutson, B.A. and S. Hahn, *Yeast Rrn7 and human TAF1B are TFIIB-related RNA polymerase I general transcription factors.* Science, 2011. **333**(6049): p. 1637-40.
44. Knutson, B.A. and S. Hahn, *TFIIB-related factors in RNA polymerase I transcription.* Biochim Biophys Acta, 2013. **1829**(3-4): p. 265-73.
45. Naidu, S., et al., *TAF1B is a TFIIB-like component of the basal transcription machinery for RNA polymerase I.* Science, 2011. **333**(6049): p. 1640-2.
46. Pilsl, M. and C. Engel, *Structural basis of RNA polymerase I pre-initiation complex formation and promoter melting.* Nat Commun, 2020. **11**(1): p. 1206.
47. *A COMPARISON OF LAC REPRESSOR BINDING TO OPERATOR AND TO NONOPERATOR DNA.*
48. Kownin, P., E. Bateman, and M.R. Paule, *Eukaryotic RNA polymerase I promoter binding is directed by protein contacts with transcription initiation factor and is DNA sequence-independent.* Cell, 1987. **50**(5): p. 693-9.
49. Marilley, M. and P. Pasero, *Common DNA Structural Features Exhibited by Eukaryotic Ribosomal Gene Promoters.* Nucleic Acids Research, 1996. **24**(12): p. 2204-2211.
50. Marilley, M., et al., *DNA structural variation affects complex formation and promoter melting in ribosomal RNA transcription.* Mol Genet Genomics, 2002. **267**(6): p. 781-91.
51. Roux-Rouquie, M. and M. Marilley, *Modeling of DNA local parameters predicts encrypted architectural motifs in Xenopus laevis ribosomal gene promoter.* Nucleic Acids Res, 2000. **28**(18): p. 3433-41.
52. Smircich, P., M.A. Duhagon, and B. Garat, *Conserved Curvature of RNA Polymerase I Core Promoter Beyond rRNA Genes: The Case of the Tritryps.* Genomics Proteomics Bioinformatics, 2015. **13**(6): p. 355-63.
53. Carter, R. and G. Drouin, *The evolutionary rates of eukaryotic RNA polymerases and of their transcription factors are affected by the level of concerted evolution of the genes they transcribe.* Mol Biol Evol, 2009. **26**(11): p. 2515-20.
54. Cavallini, B., et al., *A yeast activity can substitute for the HeLa cell TATA box factor.* Nature, 1988. **334**(6177): p. 77-80.
55. Dover, G.A. and R.B. Flavell, *Molecular coevolution: DNA divergence and the maintenance of function.* Cell, 1984. **38**(3): p. 622-3.
56. Ganley, A.R. and T. Kobayashi, *Highly efficient concerted evolution in the ribosomal DNA repeats: total rDNA repeat variation revealed by whole-genome shotgun sequence data.* Genome Res, 2007. **17**(2): p. 184-91.
57. Heix, J. and I. Grummt, *Species specificity of transcription by RNA polymerase I.* Curr Opin Genet Dev, 1995. **5**(5): p. 652-6.
58. Miesfeld, R. and N. Arnheim, *Species-specific rDNA transcription is due to promoter-specific binding factors.* Mol Cell Biol, 1984. **4**(2): p. 221-7.
59. Jackobel, A.J., et al., *DNA binding preferences of S. cerevisiae RNA polymerase I Core Factor reveal a preference for the GC-minor groove and a conserved binding mechanism.* Biochimica Et Biophysica Acta. Gene Regulatory Mechanisms, 2019. **1862**(9): p. 194408.
60. Harteis, S. and S. Schneider, *Making the Bend: DNA Tertiary Structure and Protein-DNA Interactions.* International journal of molecular sciences, 2014. **15**(7): p. 12335-12363.
61. Rohs, R., et al., *Origins of Specificity in Protein-DNA Recognition.* Annual Review of Biochemistry, 2010. **79**(1): p. 233-269.

62. Dickerson, R.E., et al., *Definitions and nomenclature of nucleic acid structure components.* Nucleic acids research, 1989. **17**(5): p. 1797-1803.
63. Xuan, J.C. and I.T. Weber, *Crystal structure of a B-DNA dodecamer containing inosine, d(CGCIAATTCGCG), at 2.4 A resolution and its comparison with other B-DNA dodecamers.* Nucleic Acids Res, 1992. **20**(20): p. 5457-64.
64. Krepl, M., et al., *Effect of guanine to inosine substitution on stability of canonical DNA and RNA duplexes: molecular dynamics thermodynamics integration study.* J Phys Chem B, 2013. **117**(6): p. 1872-9.
65. Tolle, F., et al., *By-Product Formation in Repetitive PCR Amplification of DNA Libraries during SELEX.* PLOS ONE, 2014. **9**(12): p. e114693.
66. Bailey, T.L., et al., *The MEME Suite.* Nucleic Acids Res, 2015. **43**(W1): p. W39-49.
67. Lavery, R., et al., *A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA.* Nucleic acids research, 2010. **38**(1): p. 299-313.
68. Olson, W.K., et al., *DNA Sequence-Dependent Deformability Deduced from Protein-DNA Crystal Complexes.* Proceedings of the National Academy of Sciences - PNAS, 1998. **95**(19): p. 11163-11168.
69. Brukner, I., et al., *Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides.* The EMBO journal, 1995. **14**(8): p. 1812-1818.
70. Gabrielian, A. and S. Pongor, *Correlation of intrinsic DNA curvature with DNA property periodicity.* FEBS letters, 1996. **393**(1): p. 65-68.
71. Gabrielian, A., A. Simoncsits, and S. Pongor, *Distribution of bending propensity in DNA sequences.* FEBS letters, 1996. **393**(1): p. 124-130.
72. Gabrielian, A., K. Vlahovicek, and S. Pongor, *Distribution of sequence-dependent curvature in genomic DNA sequences.* FEBS letters, 1997. **406**(1): p. 69-74.
73. Goodsell, D.S. and R.E. Dickerson, *Bending and curvature calculations in B-DNA.* Nucleic acids research, 1994. **22**(24): p. 5497-5503.
74. Vlahoviček, K., L.s. Kaján, and S.n. Pongor, *DNA analysis servers: plot.it, bend.it, model.it and IS.* Nucleic acids research, 2003. **31**(13): p. 3686-3687.
75. Chiu, T.-P., et al., *GBshape: a genome browser database for DNA shape annotations.* Nucleic Acids Research, 2015. **43**(Database issue): p. 103.
76. Liu, P., et al., *The role of DNA shape in protein-DNA recognition.* Nature (London), 2009. **461**(7268): p. 1248-1253.
77. Rohs, R., et al., *The role of DNA shape in protein-DNA recognition.* Nature, 2009. **461**(7268): p. 1248-1253.
78. Yang, L., et al., *TFBSshape: a motif database for DNA shape features of transcription factor binding sites.* NAR Breakthrough Article, 2014. **42**(Database issue): p. D148-D155.

Chapter 3

Conclusions and Future Directions

Nathan J. Munoff

Department of Biochemistry and Molecular Biology, SUNY Upstate Medical University,

Syracuse, NY 13210

1. Conclusions

Our data presented in chapter 2 has helped to elucidate the specific structural properties and characteristics of CE that CF uses to recognize and bind. Our research has revealed the sensitivity of CF binding to structural changes within the rigid AT patch of the CE. We also observed novel sequences selected by CF both in vitro and in vivo that lacked sequence conservation but had shared conserved structural features. Lastly, we were able to classify these sequences based on bendability into rigid and flexible categories with a preference by CF for more flexible sequences. By establishing the specific structural properties of CE preferred by CF in a level of detail not done before, we've set a strong foundation to further explore the CF-CE interaction.

To build on this strong foundation, we first need to ensure that the sequences we found with our in vitro and in vivo selection methods are not simply artefacts of the selection processes and can actually bind to CF. This could be carried out with EMSA and transcriptional assays. Following this, we need to experimentally validate that these selected sequences possess the properties we have predicted they do. It must be noted that the predictive programs mentioned previously are only as accurate as the data and algorithms on which they are based. Experimental data is ultimately still needed to confirm any prediction and to bring a level of accuracy not possible solely based on predictions. Here, I suggest several experiments that can more directly or indirectly verify the predicted structural properties of the novel sequences selected by CF. Some experiments such as NMR, CD, AFM, and single molecule manipulation techniques all more directly look at structure whereas methods such as FRET can look at structure more indirectly. We can also study the CF-CE interaction from the protein side to better

understand what residues of CF are critical for interacting with the CE and how. So far, we have mostly focused on characterizing the CE side of the interaction and describing what CF is looking for in DNA recognition but haven't yet delved into the properties of CF that are driving its specific mode of recognition. Additionally, given that CF doesn't have a specific consensus sequence it prefers, this greatly opens up the possibility that it could potentially be binding at other locations in the genome besides the rDNA promoter. This could be investigated through a technique such as CHEC-seq. Finally, alternative methods of SELEX could be employed in the future to gain more novel CE sequences as well optimize the selection process. In the end, there remains work to be done in understanding and characterizing protein-DNA interactions. It is here that the suggested future experiments can fill in this gap.

2. In vitro Analysis of Single Base Pair CE Mutants

2.1 Direct Binding EMSA

From our in vitro competitive EMSA binding studies, we were able to pinpoint which bases of the CE were more critical for CF recognition, finding that CF was particularly sensitive to mutations in the central AT patch of the CE. From these results we were also able to predict what structural changes in the CE these mutations were producing, and which changes were preferred over others in this region. It must be noted however, that EMSA results do not capture or tell the entire binding ability of the competitor to CF. Binding results may not be entirely attributed to the interaction between the competitor and CF and instead may also be affected in part by interactions between the competitor and the labeled probe itself. In addition, poor competitor binding

may be indistinguishable from sufficiently low competitor concentrations in a binding reaction. To complement competitor EMSAs, direct binding EMSAs of the single bp CE mutants can also be conducted. Results from the direct binding EMSAs would give us binding data free of any influence of competitor and probe interactions. Then we would be able to assess the effect of a mutation more directly on CF's binding ability.

2.2 In Vitro Transcription Assay

Additionally, CF and single bp mutant interactions could be examined from a transcriptional perspective via an in vitro transcription assay. This assay would provide a more biological context to how these mutations affect CF binding by observing the ability of transcription to be carried out as opposed to more narrowly focusing on just the CF-CE interaction. In this assay transcriptionally active extracts from yeast would be used to transcribe reporter promoter-controlled rDNA with introduced single bp mutations from a reporter plasmid[1]. The amount of RNA product produced would be analyzed. From this assay, we would expect to see a poorly performing competitor oligo also show poor levels of transcription and consequently RNA production. Additionally, we may see oligos that previously appeared to bind well as a competitor and directly but not necessarily well enough to enable significant levels of transcription. Thus, this assay would be a valuable tool in further contextualizing the single bp mutants.

3. Analysis of Physical Characteristics of DNA

3.1 NMR

As mentioned previously, the predictive programs used in our research to analyze the structural properties of novel CE sequences selected by CF while helpful and based on experimental data, cannot replace experiments directly examining structure. A variety of different methods could be employed to accomplish this aim. For example, Nuclear Magnetic Resonance(NMR) spectroscopy could be used to look at the structure of DNA[2, 3]. The benefit of NMR is that the DNA structure at atomic resolution can be observed in its natural solution state as opposed to a crystal lattice structure if one were to use a method such as X-ray crystallography. Additionally, NMR allows for observation of active dynamics of DNA such as bending, twisting, and flexibility by looking at measurements such as residual dipolar couplings, spin relaxation, and relaxation dispersion[2, 3]. When examining nucleic acids, the method of two-dimensional NMR is typically used[4].

In our application, we would take the novel CE sequences we found, single bp mutants, and in silico generated oligos and use water as our sample solvent. We could then employ a variety of techniques to look at different nuclei to determine nucleic acid structure and dynamics such as 1H NMR, 13C NMR, 15N NMR, and 31P NMR and 19F NMR. What NMR would allow us to do is to verify that sequences we have deemed as rigid or flexible for example, do in fact possess those properties. For example, we would expect an overall wider range of spectra and NMR measurements for a flexible sequence than a rigid one. We would also be able to verify if our other structural predictions of

curvature, propeller twist, helical twist, minor groove width, and roll match up with what we see in the NMR data.

3.2 Circular Dichroism

Circular dichroism is another method that can look at the structure and dynamics of DNA and is comparatively easier and simpler technique to run than NMR[5-7]. CD works by circularly polarizing light both in a left and right-handed manner and observing the differences in absorption by the molecule of interest. It is important to note, this technique is only viable for chiral molecules. Additionally, it does not provide as high of a resolution picture as NMR and single bp changes may not be observable. Again, this method would be able to tell us the structure of our CF selected sequences, single bp mutants, and in silico generated oligos and their dynamic properties like bendability but may not be sensitive enough to determine properties such as propeller twist. With CD we would expect to see more flexible sequences have wider ranges of differences in absorption than rigid ones.

3.3 Atomic Force Microscopy

Another technique that could be employed to study the structure and dynamics of DNA is atomic force microscopy(AFM)[8, 9]. AFM is a type of microscopy that uses positional and force measurements of a sharp tipped probe to build a structural picture of single biomolecules. For example, we could use this technique by combining the data from many images of a flexibly categorized DNA oligo to determine its range of conformation and see if it matches our previous predictions and vice versa for a rigid sequence. We would also be able to look at the more static property of curvature with this method. The entire system consists of a very fine tip attached to a cantilever and a

piezoelectrically controlled planar sample substrate along with a laser to measure the position of the cantilever[8](**Fig. 1**). The sample substrate is controlled very precisely to move the sample back and forth and up and down relative to the tip until it encounters the biomolecule, and a predetermined force limit is reached[8]. This process is repeated many times moving the planar substrate each time and scanning along the biomolecule until a 3-dimensional topographic map is built[8].

AFM can be done one of two ways. The first is called in-air AFM and consists of imaging sample on a planar substrate that was once in liquid solution but was allowed to dry before imaging[8]. This method allows for observation of a static snapshot of the biomolecule while it was previously in solution[8]. AFM in-liquid involves fixing or binding of the biomolecule to the planar substrate via a variety of methods before it can be imaged while still in solution[8]. This second method offers a higher resolution than the first[8]. Overall, AFM is capable of producing images in the nanometer range of resolution and when it comes to imaging DNA, can elucidate the overall shape, the helical pitch, and resolve both strands[8].

By employing both in-air and in-liquid methods, we could combine the information gained from the conformation of our novel CE sequences we found, single bp mutants, and in silico generated sequences in solution using the in-air method with the higher resolution images of the in-fluid method.

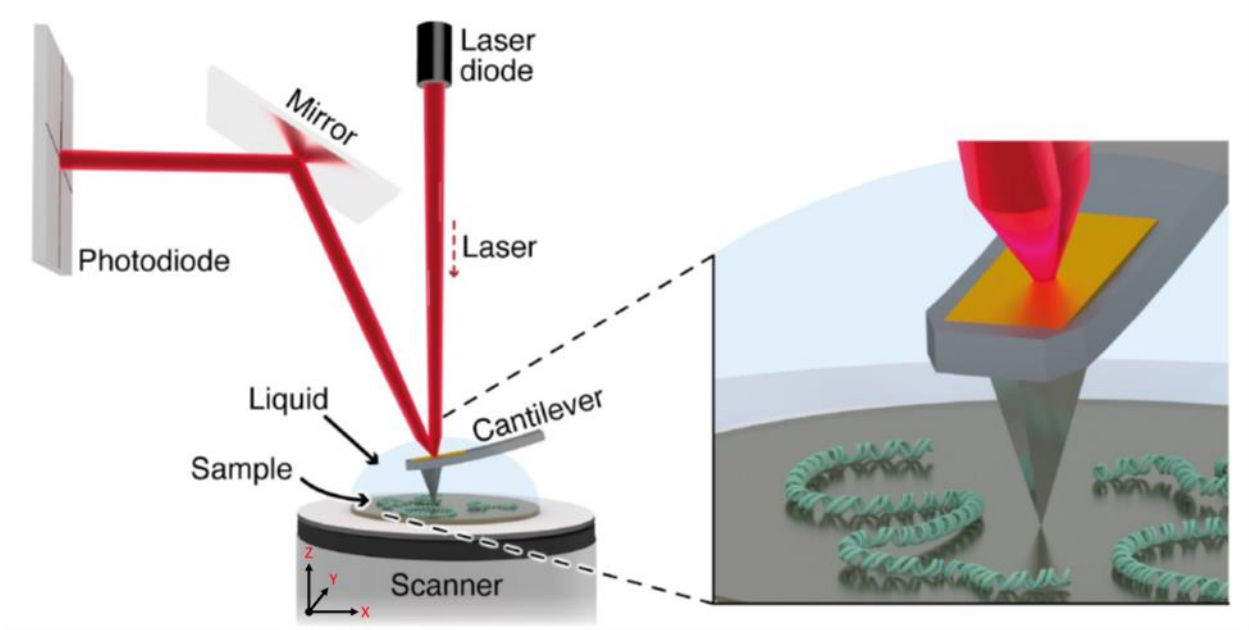

Figure 1. Atomic Force Microscopy Setup in Solution

This is the setup of AFM scanning in solution. The sharp tip attached to the cantilever scans line-by-line across the sample surface by moving a piezoelectrically controlled scanner or planar substrate surface in the x, y, and z axis relative to the tip. By measuring the position or bending of the cantilever, an image is built up of the surface topography. The bending of the cantilever is detected via deflection of a laser beam onto a 4-quadrant photodiode. An enlarged view of the tip and the sample adsorbed onto the scanner surface is shown to the right. Adapted from Haynes et al.

3.4 Fluorescence resonance energy transfer(FRET)

Single molecule FRET can also be used to study the flexibility of DNA oligos[10]. We would label each end of the sequences we want to investigate like our novel CE sequences, single bp mutants, and in silico generated oligos with the appropriate fluorophores and observe the FRET values. Sequences with higher FRET values would presumably be considered more flexible as the fluorophores would be

brought closer together than if the oligo was in perfect B conformation. However, this method is limited to studying flexibility and other specific structural features of the DNA are not able to be determined or resolved. In addition, this method may have trouble differentiating between an oligo that is highly curved but struggles to occupy other conformations against a more dynamic oligo that is able to change its conformation more freely.

3.5 Single Molecule Manipulation Techniques

There exist a multitude of single molecule manipulation techniques for studying DNA. These methods offer a direct approach to manipulate and observe DNA flexibility in a way that more passive observatory methods such as NMR cannot. By exertion of force on DNA in a controlled manner and observation of how the DNA molecule responds to these forces, conclusions can be drawn about its physical properties like its flexibility. One commonality across many of these methods, is the manipulation of DNA via beads[11]. The general technique is that one end of a single DNA molecule is affixed to a substrate and a bead is affixed to the other end[11]. The DNA molecule can then be stretched in a linear fashion by controlling the location of the bead such as through molecular or optical tweezers[11]. Fluid or electrokinetic flows can also be used for this purpose[11]. These methods could be used on the novel sequences we found selected by CF as well as our single bp mutants and in silico generated oligos to help verify our predictions of bendability for these sequences.

However, the methods I've just described are often limited in their manipulation of DNA to just in the linear sense. In other words, the entire DNA molecule can only be stretched in one direction and does not tell the complete story of a DNA molecules range

of flexibility or motion. Luckily, this is overcome with the use of a DNA pulley system[11]. The DNA pulley system is comprised of three parts similar to as described before[11](**Fig. 2**). The first is the glass capillary that serves as the substrate the DNA is fixed to and moves the DNA[11]. The second is the magnetic bead at the other end used to stretch and add tension to the DNA[11]. The third element is the silicon nitride knife which serves as the pulley of the system and has a magnet attached to it that attracts the magnetic bead[11]. How the system works is that the glass capillary can move up or down and drags the DNA molecule over the silicon nitride knife which remains fixed in place. The magnet helps keep the DNA molecule in tension and as the DNA is dragged over the knife a sharp bend is induced. By observing the position of the magnetic bead as the knife scans across the DNA, how much the DNA can be bent at any point can be determined. Further conclusions can then be made about flexibility along the molecule. We would expect to see more flexible sequences be more easily bent over the silicon nitride knife.

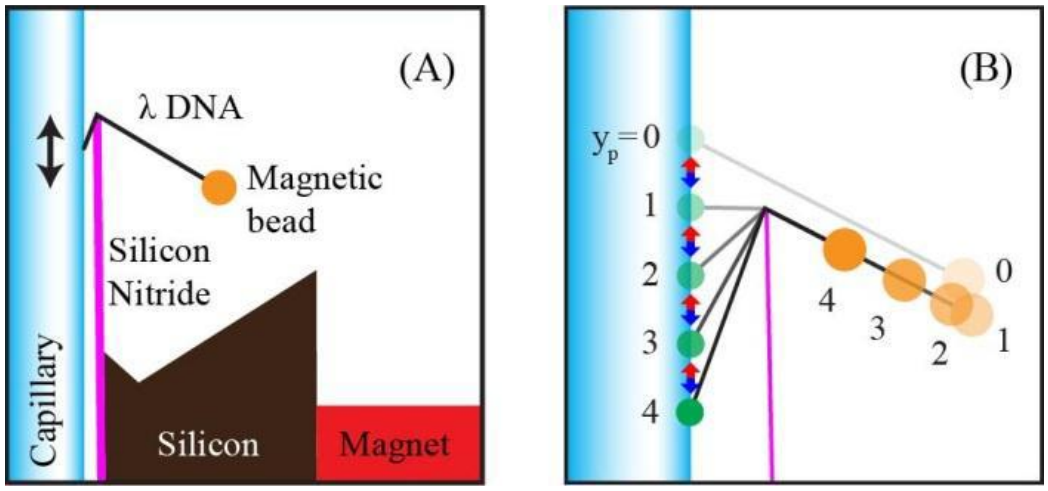

Figure 2. DNA Pulley Setup

(4) Cartoon representation of the DNA pulley system. Pictured is the λ DNA attached to the glass capillary and a magnetic bead being bent over the silicon nitride knife connected with a magnet. (B) The glass capillary is able to move up and down and five different positions of the capillary are shown marked by where the λ DNA is attached. The corresponding positions of the magnetic bead at the opposite end of the λ DNA are also marked.

4. Altered Specificity Assay

The idea of an altered specificity assay is that you mutate one partner of an interaction in order to disrupt it and then you mutate the second partner to see what mutations restore the interaction and compensate for the mutations in the first partner[12](**Fig 3**). The way in which the interaction is restored or rescued can tell you a lot about what features, and aspects of the DNA and protein are critical for interaction. For example, if a specific bond with a base is lost upon mutation of the DNA and isn't

restored upon mutation of the protein then this would indicate this particular bond wasn't critical to the interaction. This altered specificity assay can be performed both in vitro and in vivo.

In vitro, one would first mutate the DNA binding site of the protein and perform EMSAs on these mutant binding sites to find a mutant that bound poorly or not at all. We have already done this via our single bp mutant CE EMSAs. In vivo, we would insert the same single bp mutant CEs into yeast in a plasmid shuffle assay similarly as described in Chapter 2 and look for mutants with no or reduced growth phenotypes. Next, CF would be mutated and tested with the mutant DNA to see if binding or growth is recued. Ideally, choosing what residues of CF to mutate would be informed by structural data from Cryo-EM or NMR structures so that residues that are in close proximity to the mutation(s) in the DNA and potentially interact with them are selected. Determining which residues of CF are potentially interacting with which bases of the CE can be aided by a predictive program called DNAproDB[13]. DNAproDB aggregates data from a variety of databases and programs to predict both the location and type of contacts formed between any given protein and DNA structure[13-20]. We simply upload a structure of WT CF bound to WT CE and the program would predict where there are minor, major, base, sugar, or phosphate contacts between CF and CE[13]. Using this information, we could generate mutants of these critical residues. Any mutant CF complexes that lead to a recued growth phenotype in vivo or restored binding in an in vitro EMSA could have their predicted mutant CF-mutant CE structures compared to the WT CF-mutant CE and WT CF-WT CE structures to determine what contacts were lost and how they were restored, if they were restored at all[13]. By studying what residue

changes help to restore binding, we can further lay out the specific structural rules of binding between the residues of CF and the base pairs of the CE.

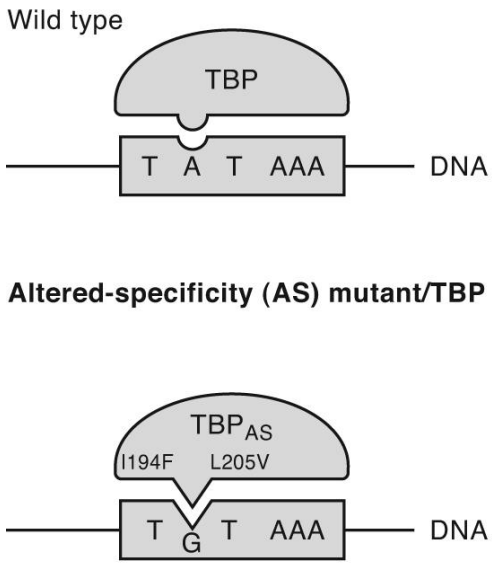

Figure 3. Example of Altered Specificity Assay Using TBP.
Pictured here is a cartoon representation of an altered specificity assay using TBP. At the top is a representation of wild type TBP bound its target TATA box DNA sequence. At the bottom, residues involved in DNA binding have been mutated and have thus altered the target DNA sequence. Adapted from Carey et al, 2009.

5. Coevolution Analysis

Finally, one can perform a coevolution analysis(**Fig 4**). Coevolution is the process by which two or more things reciprocally affect each other's evolution through natural selection. For example, coevolution can occur within proteins such as between two residues, between the residues of different proteins, and between proteins and

DNA(**Fig 4**). The way a coevolution analysis would work is by first taking orthologous sequences of both the DNA binding site and the protein of interest across a wide variety of species and aligning each set respectively(**Fig 4**). For our research, we would identify and align rDNA promoter sequences and orthologous CF protein sequences across a wide variety of eukaryotic species. Then, we would look at the frequency of changes that occur at each position of the DNA and amino acid sequence and observe any potential correlations(**Fig 4**). For example, one could say that a residue is coevolving and interacting with a particular base if every time the residue changes there is a corresponding and compensatory base change or vice versa. This is obviously an extreme example and more likely if coevolution were taking place, you would simply need to see a statistically and significantly higher correlation between the residue and base in question than would be expected by pure chance. If there is evidence of coevolution, then that tells you that particular residue and base in this example interact closely with each other and are important for the overall protein-DNA binding.

Coevolution analysis can also tell you is if the amino acids participating in contacts with DNA are following sequence or structure-based interactions. It is well documented that amino acids have certain preferences when it comes to interacting with DNA [21-24]. Some amino acids are known to mostly participate in hydrogen bonds with base pairs and sometimes prefer specific base pairs and could be classified as sequence specific amino acids [21-24]. Then there are amino acids that very rarely interface with base pairs and instead often hydrogen bond with the phosphate backbone of DNA and could be classified as structure specific amino acids[21-24]. For example, if almost every time a specific base pair changes in the CE and the corresponding amino

acid interacting with it in CF either doesn't change or changes to another amino acid that typically doesn't ever make base specific contacts, then this would suggest there is structure-based recognition taking place. Using these known characteristics, one can add a sequence or a structure-based label to these protein-DNA contacts. However, it must be noted that coevolution trends are not hard and fast "rules". While there are general preferences and trends for amino acids bound to DNA, there is a great deal of context dependence of each of these interactions that does not always fit the "rules" of a sequence or structure-based interaction.

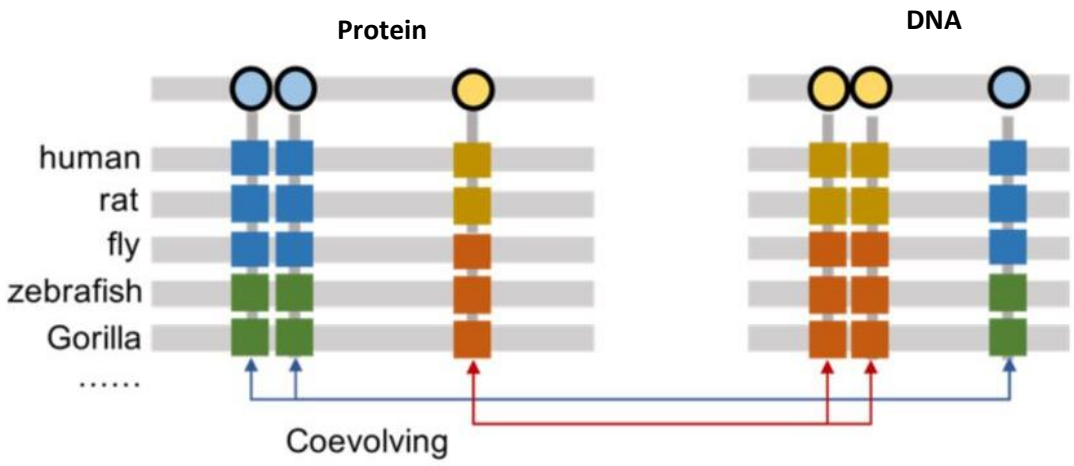

Figure 4. Example of Coevolution between a Protein and DNA.

Cartoon representation of protein and DNA sequence alignments of a variety of eukaryotic species. The key residues and bases are that are coevolving are highlighted in various colored squares and arrows are drawn between the respective coevolving pairs. For example, the transition from orange to yellow in the DNA sequences indicates a change in base and corresponds with a residue change at the same time in the coevolving protein of interest. Adapted from Fongang et al. 2019.

6. Mapping Genome Binding of CF with CHEC-seq

Chromatin endogenous cleavage is an enzymatic method of genome-wide mapping of protein binding sites or protein-DNA interactions[25]. It involves the fusion of the chromatin binding protein of interest and micrococcal nuclease(Mnase), an enzyme which cleaves unprotected DNA and depends on calcium for catalytic activity[25]. While levels of free calcium exist in yeast, multiple orders of magnitude more is required for reliable cleavage activity making this a viable method[25]. After cleavage, the DNA is purified and ready to be sequenced. What this technique would allow us to do is map CF binding across the genome, something which has yet to be done. It is obviously known that CF binds to the rDNA promoter, but it could potentially be binding elsewhere. If it does, we would then characterize these other binding sites based on their structural features as well as their affinity of binding.

There are a number of advantages to using CHEC-seq as opposed to other more traditional genome-wide protein-DNA interaction mapping methods such as ChIP-seq. For example, CHEC-seq results can be at near bp resolution whereas ChIP seq is limited to lower resolution as a result of the sonication step after protein binding and crosslinking[25, 26]. However, newer varieties of ChIP-seq are better able to combat this by replacing sonication with endonuclease digestion[26]. Additionally, antibody quality, crosslinking, and protein solubility remain potential issues and steps to be optimized in ChIP-seq but are avoided in CHEC-seq[25, 26]. Also, CHEC-seq cleavage activity can be tuned by calcium concentrations and cleavage activity can be halted at different time points[25]. This would allow for ranking of binding sites into high and low affinity based on if cleavage happened earlier or later respectively.

For my research, we would generate a vector with the Mnase sequence fused to one of the DNA binding subunits of CF either Rrn7 or 11 with a linker sequence in between. We would transform this into a yeast strain with either the Rrn7 or 11 subunits subsequently deleted from the genome. If CF is binding elsewhere besides the rDNA promoter in the genome, we would expect to see higher affinity sites enriched in the structural features we've deemed important for binding and see a greater preference for more bendable sequences. We'd also expect to see a lack of sequence conservation across all binding sites.

7. Alternative methods of SELEX

There exist multiple alternative methods to the traditional form of SELEX that we carried out in chapter 2. These methods go by many names such as one step SELEX, non-SELEX, CE-SELEX(capillary electrophoresis), FluMag-SELEX, and more[27-30]. What all these methods have in common is that either do away with or greatly reduce the repetitive amplification that is required of traditional SELEX. This helps to avoid increase of byproducts and bias in the library through successive rounds of SELEX. These methods also share that they greatly reduce the number of rounds needed to enrich the library for the target of interest. This helps to speed up the process of selection, reduce the potential loss of binding sequences that can occur from round to round, and ultimately increase specificity of the library for the target. These methods can also be coupled with high-throughput sequencing to greatly increase the number of sequences specific for the target. More sequences increase the significance and accuracy of any finding as well as potentially reveal additional preferences of CF for the CE.

8. Summary

High Priority

While a combination of all these methods would give the most complete picture of the CF-CE interaction, it is not feasible to execute each one. Therefore, priority must be given to some methods over others. One of more simple and necessary methods we could first carry out to build upon our existing knowledge are the direct binding EMSA and in vitro transcription assay previously mentioned. These methods are necessary for further validating that the sequences found in our in vitro and in vivo selection methods can in fact bind CF. The direct binding assays would again allow for direct proof of CF binding free of any influence of a competitor sequence. The in vitro transcription assay would do the same thing while adding a more biologically relevant context of transcription taking place. Both methods would be very easy to execute and are very simple to set up and similar to methods our lab already has experience with. The direct binding assay simply removes the presence of a competitor oligo and directly labels the DNA being bound but the rest of the setup is the same. The in vitro transcription assay also uses methodology that we are already familiar such as use of reporter plasmids and yeast strains with the rDNA locus genomically deleted etc. These methods would be the most likely to be done first.

Two other methods of higher priority are the altered specificity assay and the coevolution analysis. Both are relatively inexpensive and straightforward methods we could execute to further our research, and both offer similar information. They both are able to examine the relationship between CF and CE from the protein side, telling what residues are interacting with what bases of CE and potentially why. The altered

specificity assay setup is relatively straightforward and inexpensive with the only potential downside being the time it takes to generate and purify the requisite mutant proteins. Otherwise, we already have the sequences we know bind poorly or not at all to CF and simply need to generate and find mutant CF proteins that rescue binding. Testing of binding would be the same as the EMSAs we've performed before. As for the coevolution analysis, this would be even easier to execute as it would all be in silico and would have no cost associated with it. The only difficulty presented by this method is the determination of parameters and criterion of how the analysis is performed. Given the relative ease with which these methods can be executed and how they fill in a gap of knowledge of the CF-CE interaction from the protein side of things, these two methods should have a high priority of execution.

The next methods discussed of NMR, CD, AFM, and FRET can generally be grouped into the same category in terms of the type of information they generate and how they would be applied to my research. They would all be able to give some type of structural conformation information or assessment of properties like bendability of our novel CE sequences we've found. However, not all these methods are as easy to carry out. One of the easier methods to carry out is circular dichroism. CD requires less setup than the other methods. No fluorophore labeling required like in FRET, nor the more complicated experimental set up and access to more expensive equipment like in NMR or AFM. CD also gives the same amount and type of information as the other methods excepting NMR which can give more detailed structural information. The idea would be that CD could be a preliminary method to NMR where if we got valuable data from CD, we could then move onto NMR for more detailed structural resolution of the novel CE

sequences. For these reasons CD makes this most sense to carry out from this group and have a higher priority.

Low Priority

Up until this point all the proposed future methods have been standard and well documented. The single molecule manipulation studies in particular the DNA pulley system while documented, is not a method that sees as frequent use. It is a somewhat involved method also requiring access to specialized and expensive equipment and can only give you data on the bendability of DNA. However, it offers the most direct way out of any methods discussed here to study said bendability. No other method discussed allows for direct controlled manipulation of the DNA at any chosen point. For these reasons, it would be a method still worth doing but not be of the highest priority.

Another low priority future method is CHEC-seq. This method could potentially provide novel information of CF binding elsewhere in the genome besides the rDNA promoter, or it could not. If it were found that CF does indeed bind elsewhere on the yeast genome, we could also characterize these in vivo binding sites to further describe the preferences CF used for DNA binding and recognition in a more biological context. If it doesn't, then it simply tells us what we already know, and we don't gain any new information. While this method isn't expensive, it does require some setup and optimization and is something our lab has not done before. Thus, I view CHEC-seq as more of a novelty method of low priority.

Lastly, a method that would not receive as high of a priority would be the use of alternative methods of SELEX. While some of these methods are superior to the way we

chose to carry out in vitro SELEX, we chose the method we did for its simplicity and ease of use to someone like myself who at the time had little to no lab experience. Redoing our in vitro selection process would not serve much of a purpose other than to validate our findings with an alternative method and discover more sequences. However, one aspect from these alternative methods that we could borrow and apply to our existing method is the final step of high throughput sequencing. We could easily adapt our existing pools of selected sequences from our different rounds of selection to high throughput sequencing. This would allow us to greatly increase our number of selected sequences by CF which in turn would lend further significance to our findings and help to eliminate any potential biases of our previous sequencing methodology.

1. Knutson, B.A., et al., *Architecture of the Saccharomyces cerevisiae RNA polymerase I Core Factor complex.* Nat Struct Mol Biol, 2014. **21**(9): p. 810-6.
2. Bloomfield, V.A., D.M. Crothers, and I. Tinoco, *Nucleic Acids: Structure, Properties, and Functions*. 2000: University Science Books.
3. Shu, C. and H. Chien, *NMR in Biology and Medicine*. 1986: Raven Press. 248.
4. Feigon, J., et al., *Use of two-dimensional NMR in the study of a double-stranded DNA decamer.* Journal of the American Chemical Society, 1982. **104**(20): p. 5540-5541.
5. Bishop, G.R. and J.B. Chaires, *Characterization of DNA Structures by Circular Dichroism.* Current Protocols in Nucleic Acid Chemistry, 2002. **11**(1): p. 7.11.1-7.11.8.
6. Gray, D.M., R.L. Ratliff, and M.R. Vaughan, *Circular dichroism spectroscopy of DNA.* Methods Enzymol, 1992. **211**: p. 389-406.
7. Vorlickova, M., et al., *Circular dichroism spectroscopy of DNA: from duplexes to quadruplexes.* Chirality, 2012. **24**(9): p. 691-8.
8. Haynes, P.J., et al., *Atomic Force Microscopy of DNA and DNA-Protein Interactions.* Methods Mol Biol, 2022. **2476**: p. 43-62.
9. Pyne, A.L.B., et al., *Base-pair resolution analysis of the effect of supercoiling on DNA flexibility and major groove recognition by triplex-forming oligonucleotides.* Nat Commun, 2021. **12**(1): p. 1053.
10. Kang, J., J. Jung, and S.K. Kim, *Flexibility of single-stranded DNA measured by single-molecule FRET.* Biophys Chem, 2014. **195**: p. 49-52.
11. Shon, M.J., *Trapping and Manipulating Single Molecules of DNA*, in *Chemistry and Molecular Biology*. 2014, Harvard University. p. 102.
12. Carey, M., C.L. Peterson, and S.T. Smale, *Transcriptional regulation in eukaryotes*. 2. ed. ed. 2009, Cold Spring Harbor, N.Y: Cold Spring Harbor Laboratory Press.
13. Sagendorf, J.M., H.M. Berman, and R. Rohs, *DNAproDB: an interactive tool for structural analysis of DNA-protein complexes.* Nucleic acids research, 2017. **45**(W1): p. W89-W97.
14. Cock, P.J.A., et al., *Biopython: freely available Python tools for computational molecular biology and bioinformatics.* Bioinformatics, 2009. **25**(11): p. 1422-1423.
15. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers, 1983. **22**(12): p. 2577-2637.
16. Lavery, R. and H. Sklenar, *Defining the Structure of Irregular Nucleic Acids: Conventions and Principles.* Journal of biomolecular structure & dynamics, 1989. **6**(4): p. 655-667.
17. Lee, B. and F.M. Richards, *The interpretation of protein structures: Estimation of static accessibility.* Journal of molecular biology, 1971. **55**(3): p. 379,IN3-400,IN4.
18. Lu, X.J. and W.K. Olson, *3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures.* Nucleic acids research, 2003. **31**(17): p. 5108-5121.
19. McDonald, I.K. and J.M. Thornton, *Satisfying Hydrogen Bonding Potential in Proteins.* Journal of molecular biology, 1994. **238**(5): p. 777-793.
20. Mitternacht, S., *FreeSASA: An open source C library for solvent accessible surface area calculations.* 2016.
21. Luscombe, N.M., R.A. Laskowski, and J.M. Thornton, *Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level.* Nucleic acids research, 2001. **29**(13): p. 2860-2874.
22. Mandel-Gutfreund, Y. and H. Margalit, *Quantitative parameters for amino acid-base interaction: Implications for prediction of protein-DNA binding sites.* Nucleic acids research, 1998. **26**(10): p. 2306-2312.

23. Rohs, R., et al., *The role of DNA shape in protein-DNA recognition.* Nature, 2009. **461**(7268): p. 1248-1253.
24. Sousa, F., C. Cruz, and J.A. Queiroz, *Amino acids-nucleotides biomolecular recognition: from biological occurrence to affinity chromatography.* Journal of molecular recognition, 2010. **23**(6): p. 505-518.
25. Zentner, G.E., et al., *Corrigendum: ChEC-seq kinetics discriminates transcription factor binding sites by DNA sequence and shape in vivo.* Nat Commun, 2015. **6**: p. 10264.
26. Furey, T.S., *ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions.* Nat Rev Genet, 2012. **13**(12): p. 840-52.
27. Berezovski, M.V., et al., *Non-SELEX: selection of aptamers without intermediate amplification of candidate oligonucleotides.* Nature Protocols, 2006. **1**(3): p. 1359-1369.
28. Stoltenburg, R., T. Schubert, and B. Strehlitz, *In vitro Selection and Interaction Studies of a DNA Aptamer Targeting Protein A.* PLoS One, 2015. **10**(7): p. e0134403.
29. Wilson, R., et al., *Single-step selection of bivalent aptamers validated by comparison with SELEX using high-throughput sequencing.* PLoS One, 2014. **9**(6): p. e100572.
30. Yang, J. and M.T. Bowser, *Capillary electrophoresis-SELEX selection of catalytic DNA aptamers for a small-molecule porphyrin target.* Anal Chem, 2013. **85**(3): p. 1525-30.