

Characterizing Interactions of the Sox2 and LEF1 Transcription Factors with Non-B-Form Nucleic Acids

Abigail E. Hein
Department of Biochemistry
University of Colorado at Boulder

Defended March 29th, 2022

Thesis Advisor:
Dr. Robert T. Batey, Department of Biochemistry

Committee Members:
Dr. Jeffrey C. Cameron, Department of Biochemistry (Honors Council Representative),
Dr. Edward B. Chuong, Department of Molecular, Cellular, and Developmental Biology,
Dr. Deborah S. Wuttke, Department of Biochemistry

Table of Contents

Acknowledgments	3
Abstract	4
1. Introduction	5
1.1 Transcription Factors and mechanisms of DNA Recognition	5
1.2 Non-Canonical Transcription Factor-Nucleic Acid Interactions	7
1.2.1 Transcription Factor Binding to Nucleosomal DNA	8
1.2.2 Transcription Factor-RNA interactions	9
1.2.3 Transcription Factor interactions with G-quadruplexes	10
1.3 High-Mobility Group Box Proteins	14
1.3.1 Sox2 and Sox Family Proteins	15
1.3.2 LEF1 and TCF Family Proteins	17
1.4 Summary	17
2. Attempting to Determine the Structure of Sox2 in Complex with RNA through X-Ray Crystallography	19
2.1 Introduction	19
2.2 Results	23
2.2.1 Design of RNA Library	23
2.2.2 Surveying Sox2-RNA Crystallographic Conditions	25
2.3 Discussion and Future Directions	26
3. Developing an <i>In Silico</i> and <i>In Vitro</i> Pipeline for the Identification of G4-Binding HMGB Proteins	27
3.1 Introduction	27
3.2 Results	30
3.2.1 LEF1 Associates with G4s <i>In Vivo</i>	30
3.2.2 LEF1 Associations with G4s are Enriched in Functional Regions of the Genome	33
3.2.3 Verification of LEF1-G4 association by G4-seq	35
3.2.4 Sox2 Associates with G4s in Embryonic Stem Cells	36
3.2.5 Sox2 binds Genomic G4 Sites <i>In Vitro</i>	39
3.3 Discussion	45

4. Conclusion and Future Directions	48
5. Materials and Methods	50
5.1 Protein Purification	50
5.2 RNA Purification	51
5.3 Nucleic Acid preparation	54
5.4 Fluorescence Anisotropy Binding Assays	54
5.5 Electrophoretic Mobility-Shift Assay	55
5.6 Crystal screening	55
5.7 Bioinformatic Analysis Pipeline	56
5.8 Thioflavin-T Fluorescence Assay	57
Appendix 1: Sox2-RNA Hairpin Stoichiometric Binding Gels	58
Appendix 2: LEF1-G4 genomic associations in HEK293 cells	59
Appendix 3: LEF1-G4 associates with G4s predicted by G4 ChIP-seq and G4-seq	60
Appendix 3: Sox2-G4 associations in H9 cells	61
Appendix 4: ChIP Dataset Distances from TSS	62
Appendix 5: <i>In Vitro</i> Survey Genomic G4s	63
Appendix 6: Sox2-G4 Stoichiometric Binding Gel	63
Appendix 7: In Gel Verification of G4 Duplex Formation	64
Appendix 8: Nucleic Acid Constructs	65
Appendix 9: Protein Constructs	66
Data and Code Availability	66
References	66

Acknowledgments

I am deeply honored to have worked in the Batey lab for the past two years. I have been fortunate enough to work amongst passionate, dedicated, researchers without whom I would never have been able to complete this project.

I would like to extend my deepest thanks to my advisor, Dr. Robert T. Batey. He has gone above and beyond to support my research and has encouraged me to pursue science in my career. I most certainly would not have been able to progress as much into research had it not been for his patience, advice, and encouragement. It is because of his mentorship that I have been able to grow as a scientist and I cannot thank him enough.

Many thanks to the members of the Batey lab, all of whom have been nothing but welcoming and helpful to me. To Dr. Otto Kletzien, Lisa Hansen, Shea Siwik, Shelby Lennon, and Savannah Spradlin, your advice, feedback, and friendship has been invaluable, and I am eternally glad to have had the opportunity to work with you all, and I cannot overstate my appreciation. To Desmond Hamilton, my deepest thanks for all the opportunities you have shared with me and all the time you have taken to help me. Your insight and honesty are always appreciated. Thank you very much for all the support you have given

I would also like to thank Dr. Annette Erbes for her patience and support in training me, Dr. Mary Ann Allen for her invaluable help to get me started with my bioinformatic analysis, and my committee members Dr. Deborah Wuttke, Dr. Edward Chuong, and Dr. Jeffrey Cameron for their advice and direction.

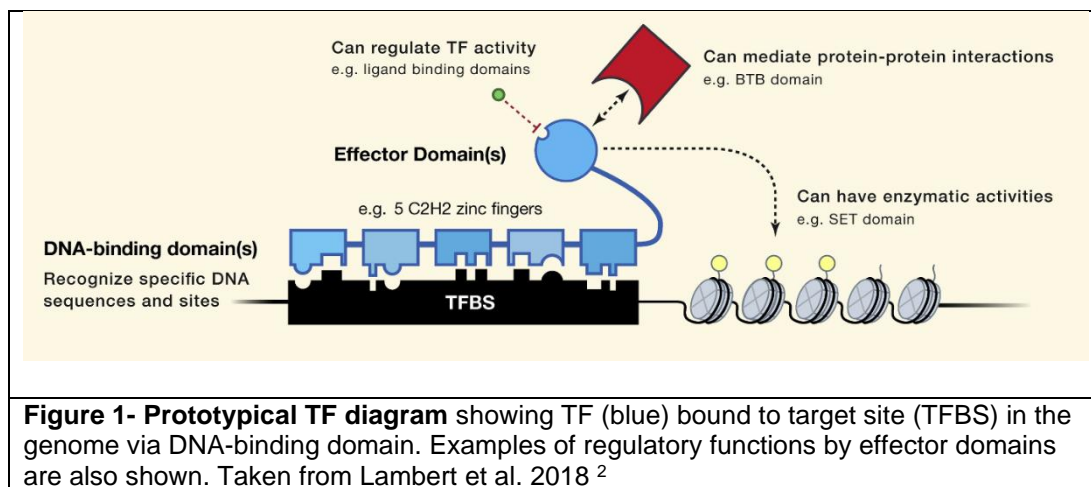
Abstract

Transcription factors have been increasingly found to directly interact with a broad range of nucleic acids distinct from their canonical targets of B-form DNA. These interactions play roles in the regulation of gene expression that remain poorly understood to date. HMGB proteins are one such family of transcription factors that have been repeatedly implicated in alternative regulatory roles via their ability to associate with a diverse set of nucleic acid structures. We hypothesize that these transcription factors may have the capability to bind nucleic acid targets with some form of structural selectivity. Specifically, we examine the capabilities of Sox2 to bind internally bulged RNA hairpins and investigate the possibility of Sox2 and LEF1 directly targeting G-quadruplex structures in the genome. We find that, both, Sox2 and LEF1 associate with G-quadruplexes in the genome and posit that these interactions are likely relevant for the biology of these proteins. Furthermore, we find that Sox2 binds genomic G-quadruplexes with high affinity *in vitro* and develop the framework for an *in vitro* workflow to thoroughly assess these interactions.

1. Introduction

1.1 Transcription Factors and mechanisms of DNA Recognition

Transcription factors (TFs) are a diverse class of proteins that alter gene expression in the cell by facilitating or repressing the transfer of genetic information from DNA to RNA.¹ As such, they are crucial for the regulation of gene expression at the level of transcription. The TF class of proteins has historically included any protein capable of altering gene expression². Currently however, TFs are generally considered to be proteins that attenuate transcription and are composed of at least one DNA binding domain responsible for recognizing target sites in the genome² (Fig. 1). The regulation of transcription can occur through the activity of one or more effector domains or through sterics.²



Canonically, these proteins act on regulatory elements of the genome, either binding the promoter region directly upstream of the transcription start site of a gene or binding a distal enhancer region.^{2,3} These binding events modulate the activity of RNA polymerase to transcribe the gene into mRNA, which is subsequently translated into

protein. Most TFs work as part of a regulatory system involving multiple related genes associated with a specific set of cellular functions. This allows for incredible precision in the control of gene expression.¹ TF expression is one of the driving factors of cellular differentiation, as there are distinct sets of TFs associated with each cell type within a given organism.

TFs often recognize a conserved “consensus sequence” of DNA that determines the protein’s target sites.¹ The process of DNA recognition through specific interactions between residues of the TF and a sequence of DNA bases is referred to as “base readout” recognition.^{3,4} However, given the prolific size of the genome, it is statistically impossible that TFs would be able to consistently localize to their functional target sites based on sequence recognition alone. Furthermore, it has been observed that only about 1-3% of a TF’s consensus sites in the genome are actually bound *in vivo*.^{5,6} Therefore, it has been determined that there must be unforeseen factors that contribute to TF localization. These include, but are not limited to, chromatin accessibility, the presence of cofactors, 3D structure of the DNA, and the presence of epigenetic signals (such as methylation).³

In addition to base readout recognition, TF proteins can also interact with DNA through what is commonly referred to as shape readout recognition. While base readout is primarily driven by hydrogen bonding between amino acid residues and nucleic acid bases, shape readout is driven by electrostatic and Van der Waals interactions (such as pi stacking).⁴ This leads the protein to recognize the local structure of the DNA rather than its sequence. It has been repeatedly shown that TFs often utilize both DNA recognition mechanisms, in concert, to bind target sites. This is exemplified by the p53

and GATA3 TFs, both of which recognize well defined consensus sequences. These TFs bind their putative targets through base specific hydrogen bonding in the major groove, but and recognition of minor groove geometry.^{7,8}

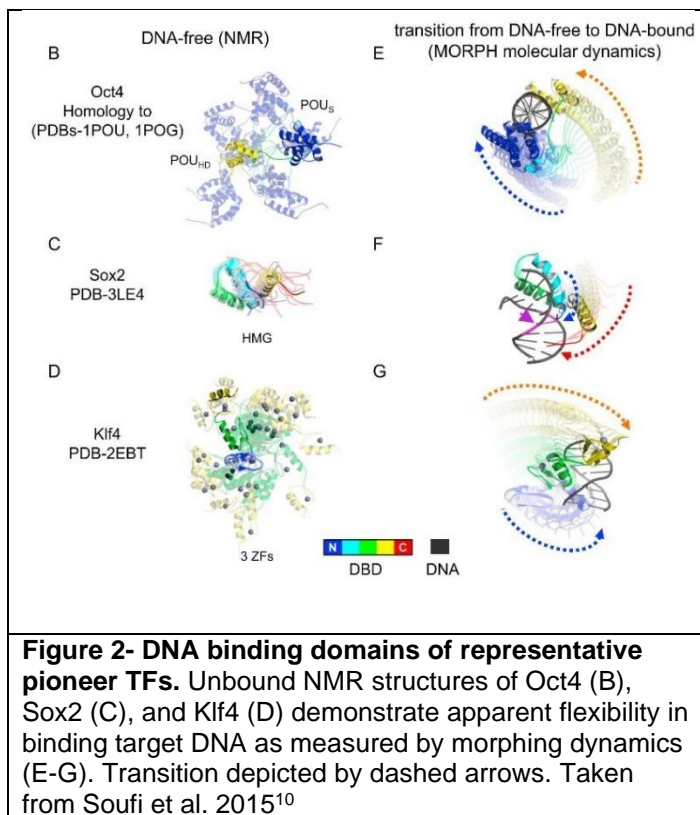
Furthermore, interactions with non-consensus DNA appear to play an important role in the localization of TFs. Kinetic studies of the model Sox2 TF have shown that this protein localizes to its target sites by sampling nonspecific sites in the genome followed by a period of one-dimensional sliding. If a consensus site is found, the TF then lingers at that site for a longer period of time.⁹ These data suggest a model of TF activity in which gene expression is not regulated exclusively through interactions with a consensus sequence, but rather with a diverse range of nucleic acid interactions.

1.2 Non-Canonical Transcription Factor-Nucleic Acid Interactions

Although TFs are most commonly studied in the context of their DNA-binding activities, a growing body of evidence indicates that gene regulation by TFs occurs through interactions with a wide variety of nucleic acids. Several TF families appear to play regulatory roles through non-canonical TF-nucleic acid interactions distinct from their consensus DNA-binding activities. Although interactions with B-form DNA are a well categorized function of these proteins, recent evidence indicates that interactions with non-B-form nucleic acids modulate gene regulation by TFs. These non-B-form nucleic acids include nucleosomal DNA, G-quadruplexes, and RNA.

1.2.1 Transcription Factor Binding to Nucleosomal DNA

When transcriptionally inactive, the DNA of eukaryotic organisms is tightly compacted around nucleosomes, which are composed of histone protein octamers. This allows the genome to be compacted into chromatin in which nucleosomal DNA is conformationally constrained by the histone octamer and is inaccessible to transcriptional machinery. As a result, gene expression is heavily influenced by local

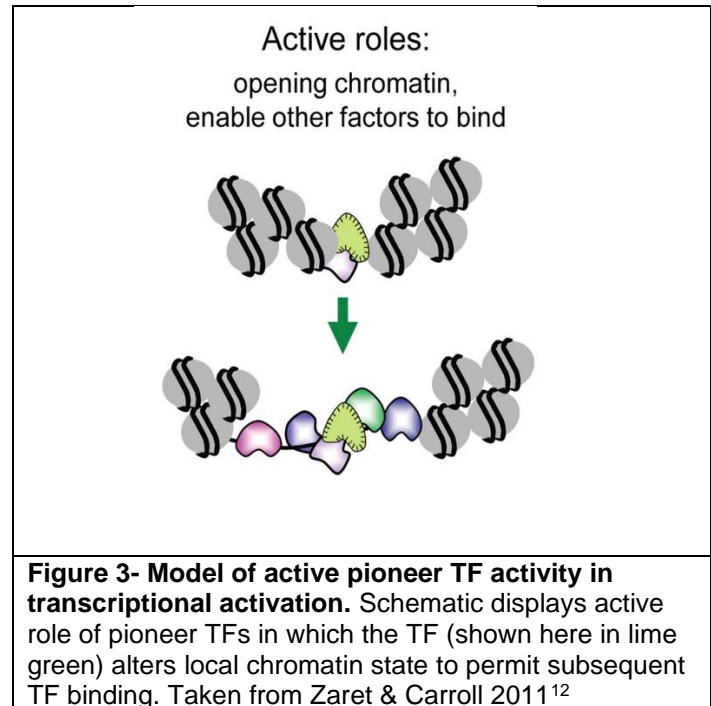


chromatin structure and DNA accessibility. There are a number of mechanisms for chromatin remodeling that allow for active transcription, either through histone modifications, nucleosome sliding, or through nucleosome binding proteins.¹¹ One such class of nucleosome binding proteins are “pioneer transcription factors,” including such proteins as FOXA,

Sox2, Oct4, myc, GATA and Klf4.^{10,12} This term refers to a group of TFs that are capable of binding condensed nucleosomal DNA. This capability is driven by the pioneer TFs flexibility and its ability to bind a partial consensus motif that is accessible from the nucleosome surface (Fig. 2).

The binding of pioneer TFs generally either causes a DNA distortion that loosens the DNA from the histone octamer or detaches terminal DNA from the octamer.^{13,14} This

initial binding event allows other TFs to bind cooperatively to their target site, leading to active transcription of the target gene(s) (Fig. 3). Additionally, the activity of pioneer TFs is critical for stem cell differentiation, as the presence of Sox2, Oct4, myc, and Klf4 is sufficient to induce pluripotency in differentiated cells.¹⁵



1.2.2 Transcription Factor-RNA interactions

Although TFs are commonly considered DNA-binding proteins, a growing body of evidence indicates that they also interact productively with RNA. These interactions have been implicated in RNA processing, often through direct binding to RNA. Several families of TFs have been observed to bind RNA, including Zinc finger proteins and homeodomain proteins.¹⁶ This indicates that many TF families may form an underappreciated subclass of DNA and RNA-binding proteins (DRBPs).¹⁷

Within the class of known DRBP TFs, there are several mechanisms by which RNA-binding could occur. The first and most intuitive of these mechanisms is through the use of a distinct domain, separate from the consensus DNA-binding domain, as in the case of P53¹⁸ or TFIIIA.¹⁹ However, it also appears common for DRBP TFs to bind RNA through their canonical DNA-binding domains, as has been observed with,^{20,21} NF- κ B,²² and RUNX1.²³ The binding of RNA through the DNA-binding domain is often

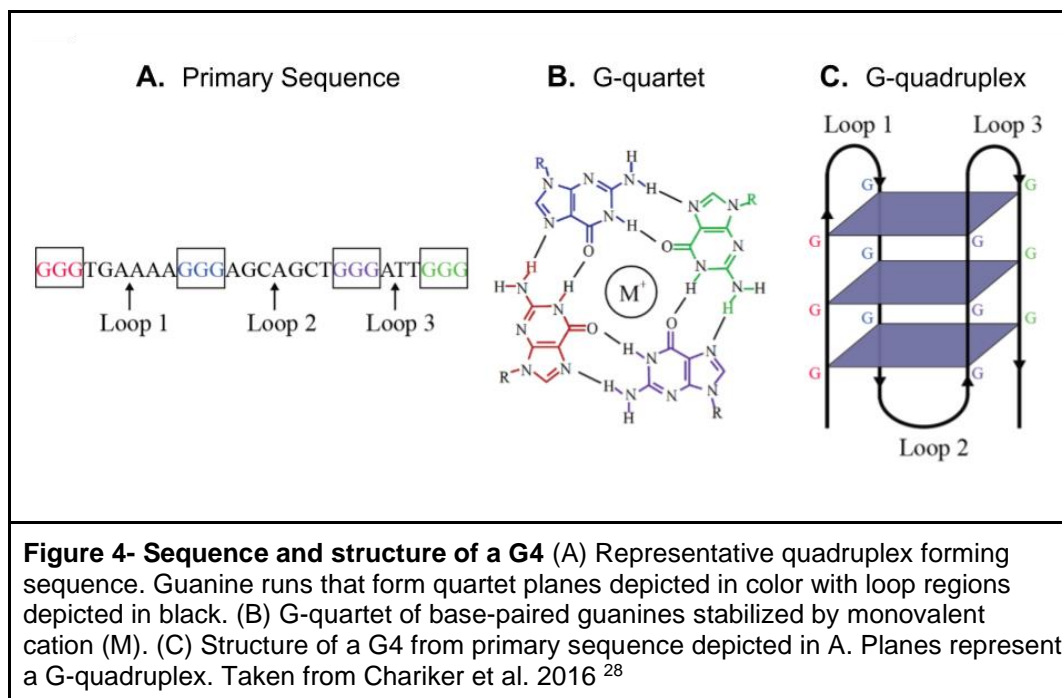
associated with loss of sequence specificity. For example, a crystal structure of the NF- κ B TF in complex with RNA showed that the TF used a nearly identical interface to bind DNA and RNA, despite the RNA sequence bearing little resemblance to the TF's consensus DNA sequence.²²

Dual DNA and RNA-binding capabilities allow TFs to engage in orthogonal regulation of gene expression at multiple levels. In fact, TF-RNA binding activity is involved in a diverse subset of RNA regulatory and processing roles. These include repression of TF activity through RNA mimicry of DNA substrates,²³ splicing,²⁴ genomic localization and activation of TF activity through association with nascent RNA transcripts.²⁵ The wide range of regulatory functions mediated by TF-RNA interactions indicates that these interactions play a large and underappreciated role in gene regulation by TFs.

1.2.3 Transcription Factor interactions with G-quadruplexes

As mentioned above, G-quadruplexes (G4s) are another major class of noncanonical TF targets. A G4 is a unique secondary structural element that can form in guanine-rich nucleic acids. Guanine nucleotides are capable of forming a “G-quartet” in which 4 guanines base pair on both genetic and Hoogsteen faces to form a ring²⁶ (Fig. 4B). The G4 is composed of multiple stacked G-quartets that form a multi-planed, shelf-like structure in nucleic acids (Fig. 4C). From this, a potential quadruplex-forming sequence (PQS) has been identified as “G₃₊N₁₋₇G₃₊N₁₋₇G₃₊N₁₋₇G₃₊” where ‘N’ refers to any nucleotide and correlates to the unpaired loop regions of the quadruplex²⁷ (Fig. 4A). This structure usually forms intramolecularly where a single stranded region of a nucleic

acid base-pairs with itself to fold into a quadruplex. However, bimolecular G4s and, in rare cases, tetramolecular G4s have also been observed. These intermolecular quadruplexes form quartets with guanines from two or more nucleic acid strands, giving them the unique ability to form hybrid DNA-RNA G4s.



G4 formation is highly dependent on the presence of monovalent cations. The positive charge is required to neutralize the electrostatic repulsion of the carbonyl groups on the guanines that are in close contact due to G-quartet formation. To counteract this electrostatic repulsion, the cation sits in the middle of the ring formed by the G-quartet (Fig. 4B). Due to their size, potassium ions are best suited to stabilize this secondary structure as they can be positioned between adjacent quartets.²⁹ Sodium ions can also stabilize this structure but to a lesser extent as the size of the sodium ion makes it such that they are not regularly positioned in the quadruplex.³⁰ Lithium ions,

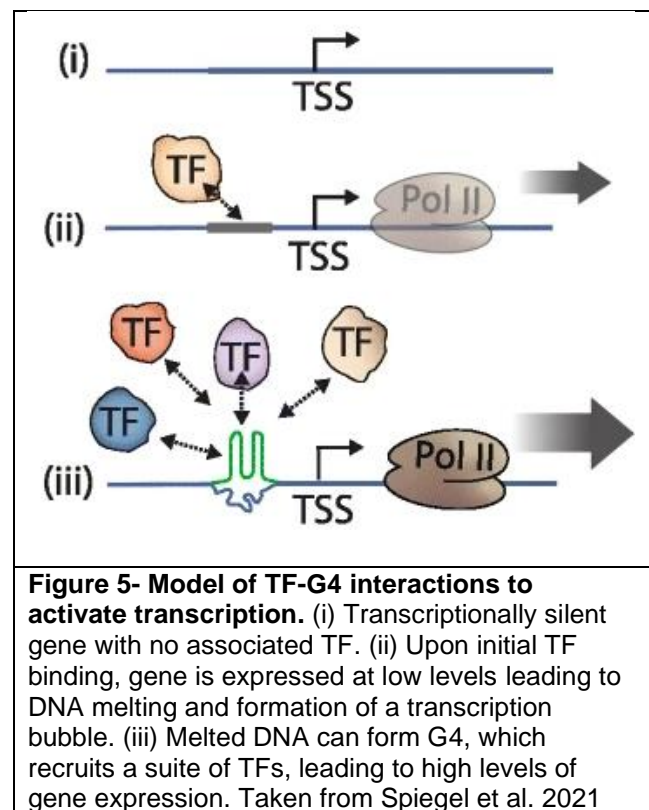
however, are too small to effectively stabilize the formation of the quartet and, as such, the presence of lithium does not favor quadruplex formation.²⁹

G4s were initially characterized *in vitro*²⁶, but initial studies did not clearly demonstrate if this structure was biologically relevant. However, RNA and DNA G4s have been increasingly identified *in vivo* and are now thought to mediate a variety of regulatory roles including pre-mRNA splicing, telomere maintenance, chromatin remodeling, replication, transcription, and translation.²⁹ Additionally, bioinformatic surveys indicate that G4s are enriched in promoter regions of the genome. During transcription, the superhelical stress caused by generating a transcription bubble causes the promoter region and other regulatory elements upstream of the transcription start site to become unpaired.³¹ The formation of the transcription bubble creates the ideal environment for G4 formation as it disrupts the competing double-stranded conformation. As a result, G4s generally occur in regions of active transcription. This model is further supported by the observation that G4s form in nucleosome-deplete regions.^{31,32} Furthermore, promoter G4s are associated with genes with high transcription levels across multiple cell lines and have been attributed to help form the cell-type specific transcriptome.³³ Notably, G4s have been documented in the promoters of numerous oncogenes, including MYC, KRAS, and KIT.³⁴

The evidence for the biological relevance of G4s was bolstered by the discovery of a number of proteins that exhibit G4-binding specificity. The first of these were a set of helicases, including BLM and WRN, that were found to possess G4-unwinding capabilities.^{35,36} G4-specific helicases are thought to be important to prevent polymerase or ribosome stalling in the presence of quadruplexes.³⁷ Another interesting

subclass of G4-binding proteins that have emerged are TFs. Specifically, there appears to be a suite of TF's that interact directly with G4s. Several TFs, such as SP1, YY1, and MAZ have been found to exhibit G4 specific binding *in vitro*.^{38–40} Furthermore, these proteins have also been shown to associate with G4 structures *in vivo*.^{38,41} An *in vivo* chemical profiling study of DNA G4-interacting proteins identified 201 candidate G4-interacting proteins.⁴² 24.4% (49/201) of which are known TFs.² Independently, a study of TF genomic binding sites found that *in vivo* DNA G4s are enriched binding sites for many different TFs⁴³ (Fig. 5). Put together, these data suggest a mode of gene regulation that relies on productive interactions between G4s and TFs.

There are several proposed functions of TF-G4 interactions in gene regulation. Given the enrichment of active G4's in TF binding-sites, one proposed model posits that G4's form in the promoters of actively transcribed genes and recruit multiple TFs and chromatin remodeling proteins to further activate transcription^{33,43} (Fig. 5). It has also been observed that G4s stabilize long-range enhancer-promoter DNA looping interactions.⁴⁴ G4s are likely to act in concert with TFs to form and stabilize these loops.⁴⁵ A model of this phenomenon is YY1, a TF that is not only known to



stabilize DNA loops, but also has been observed to interact with G4s. Disruption of YY1-G4 interactions *in vivo* have been shown to disrupt DNA-looping.³⁸

1.3 High-Mobility Group Box Proteins

One family that has repeatedly emerged in studies of TF interactions with non-B-form nucleic acids are High-Mobility Group Box (HMGB) proteins. HMGB proteins form a diverse family of TFs and chromatin remodeling proteins, all of which are defined by the presence of a highly conserved DNA-binding domain called the HMGB domain. This family can be further divided into seven subfamilies based on sequence homology of the HMGB domain.⁴⁶

The HMGB domain consists of three alpha helices that fold into an L-shaped conformation⁴⁷ (Fig. 6 shown in red). This domain intercalates into the minor groove of B-form DNA, underwinding the helix and inducing a dramatic 60-120° bend at the binding site.^{48–53}

The majority of HMGB proteins

contain one HMGB domain that binds a consensus DNA sequence. However, a select subpopulation of HMGB domains bind nucleic acids without sequence specificity. Non-sequence specific domains are usually found in proteins with multiple HMGB domains (e.g. TFAM and HMGB1-4).⁴⁷ These proteins generally act as chromatin remodelers,

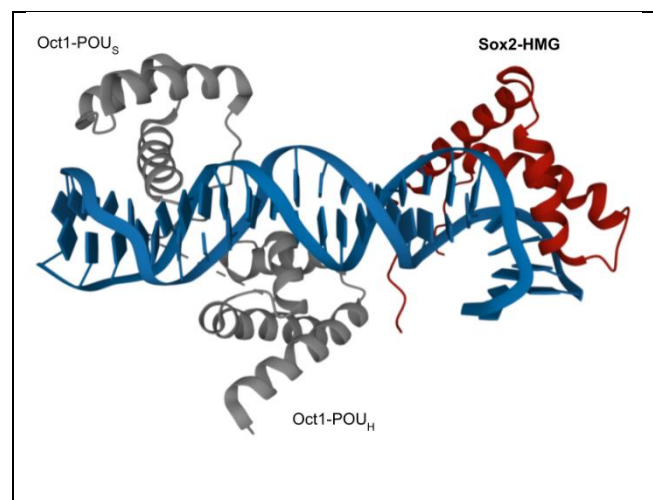


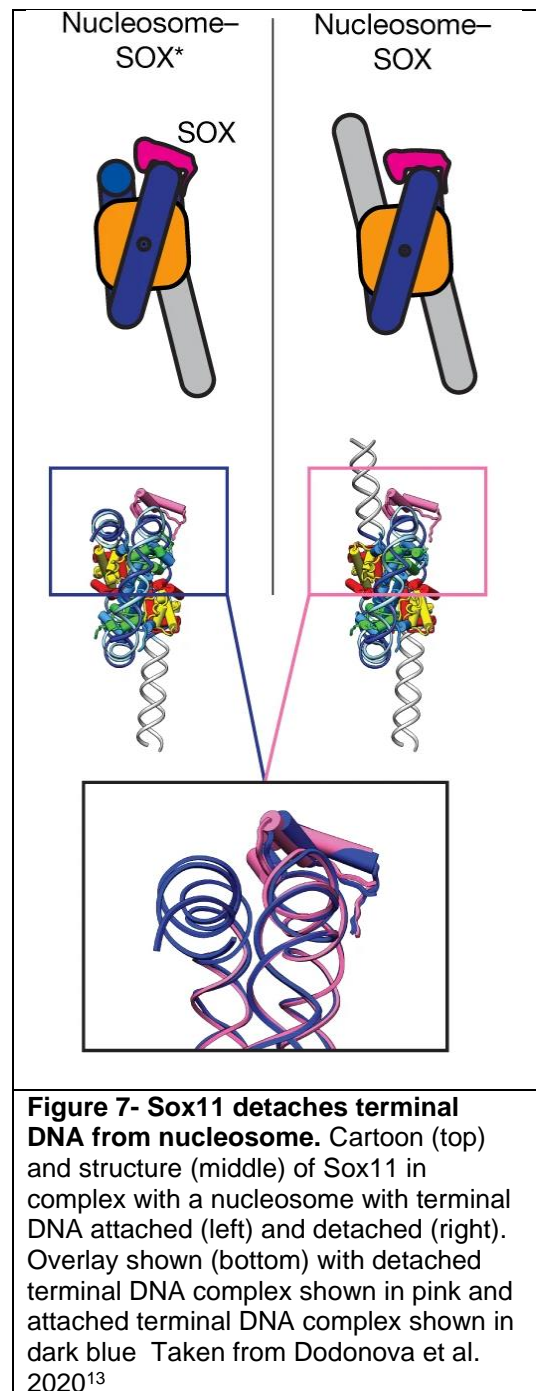
Figure 6- Sox2/Oct1/FGF4 crystal structure. Structure depicts HMG domain of Sox2 in complex with a DNA segment from the FGF4 enhancer and with the POU_S and POU_H domains of binding partner, Oct1. PDB ID: 1GT0

whereas the sequence specific HMGB proteins tend to act as TFs.⁴⁷ This study focuses on two families of sequence specific HMGB proteins, the Sox and TCF families of TFs, and specifically on the representative proteins Sox2 and Lef1 from each respective family.

1.3.1 Sox2 and Sox Family Proteins

The Sox, or SRY-related HMG-box, family of TFs are defined by an HMGB domain 50% sequence similarity to the HMGB domain of SRY, a TF that plays a critical role in sex determination. This family can further be divided into nine groups (A, B1, B2, C, D, E, F, G, H), with at least 80% sequence homology within each group.⁵⁴ Sox proteins recognize a 5'-TTGT core consensus DNA motif.⁵⁵ Within this family, Sox2 is the best studied and will be the focus of this work.

Sox proteins are master regulators of gene expression and are active during early development to regulate cellular lineage commitment. In particular, Sox TFs regulate neurogenesis and the differentiation of neural



precursor cells,⁵⁶ but are also involved in the differentiation of a wide range of cellular























SOX	Partner	Cell-type specification
		ES, Inner cell mass
		Neural progenitor
		Neural progenitor
		Retina, Lens
		Melanocyte
		Melanocyte
		Schwann cell
		Schwann cell
		Chondrocyte
		Chondrocyte
		Vascular endothelium

Figure 8- Sox proteins associate with binding partners to specify cell fate. Sox family protein (left) depicted with known binding partners (middle) and lineage specified by their interactions (right). Taken from Kondoh & Kamachi 2010

lineages.⁵⁷ This capability stems from the ability of Sox proteins to act as pioneer TFs.⁵⁸ Sox2 and Sox11 have been shown to loosen nucleosomal DNA from the histone octamer, allowing them to make condensed chromatin transcriptionally active¹³ (Fig. 7). To this end, Sox proteins frequently interact with other TFs as binding partners to specify lineage commitment^{57,59} (Fig. 8).

There is a growing body of evidence that points to Sox family proteins as a family of RNA-binding TFs. Sox2 has been found to bind RNA in vitro and directly associate with RNA in vivo through its HMGB domain.⁶⁰ Sox proteins have also been found to directly bind RNA during splicing and it has been suggested that this behavior may be

important for bending pre-mRNA to allow splicing to occur, although this claim has not yet been confirmed.²⁴ Furthermore, Sox2 has been found to associate with a number of lncRNAs, such as LincQ, RMST, and Evf2. These associations have been found to play a role in neurogenesis and the maintenance of pluripotency, and may be important for Sox2 localization to genomic target sites.^{61–63}

1.3.2 LEF1 and TCF Family Proteins

The T-cell factor (TCF)/lymphoid enhancer factor (LEF) family of proteins is another sequence specific family of HMGB TFs. This family recognizes a 5'-TCAAAG consensus motif.^{64,65} In addition to the HMGB domain, TCF/LEF proteins also contain a conserved basic domain at their C-terminus. This domain both enhances nucleic acid affinity through electrostatic interactions and allows this family to bind β -catenin.⁶⁶ β -catenin binding ablates TCF/LEF proteins to modulate the activity of Wnt target genes. These proteins have been observed to act as repressors of Wnt target genes in the absence of β -catenin and transcriptional activators of these genes when β -catenin is present.^{66,67} Wnt signaling is important for the regulation of stem cells through mediating such functions as stem cell self-renewal.^{68,69} As a result, TCF/LEF proteins are, like Sox proteins, very active in embryonic cells during early development. This family has also been observed to play a role in lineage differentiation of multipotent cells and are frequently mis-regulated in cancers.^{67,70}

1.4 Summary

Transcription factors are a deeply diverse class of proteins that have been can directly interact with a broad range of nucleic acids distinct from B-form DNA such as with RNA and G4s. These interactions play roles in the regulation of gene expression that remain poorly understood to date. HMGB proteins are one such family of transcription factors that have been repeatedly implicated in alternative regulatory roles via their ability to associate with a diverse set of nucleic acid structures. We hypothesize that these transcription factors may be acting through structure specific recognition of

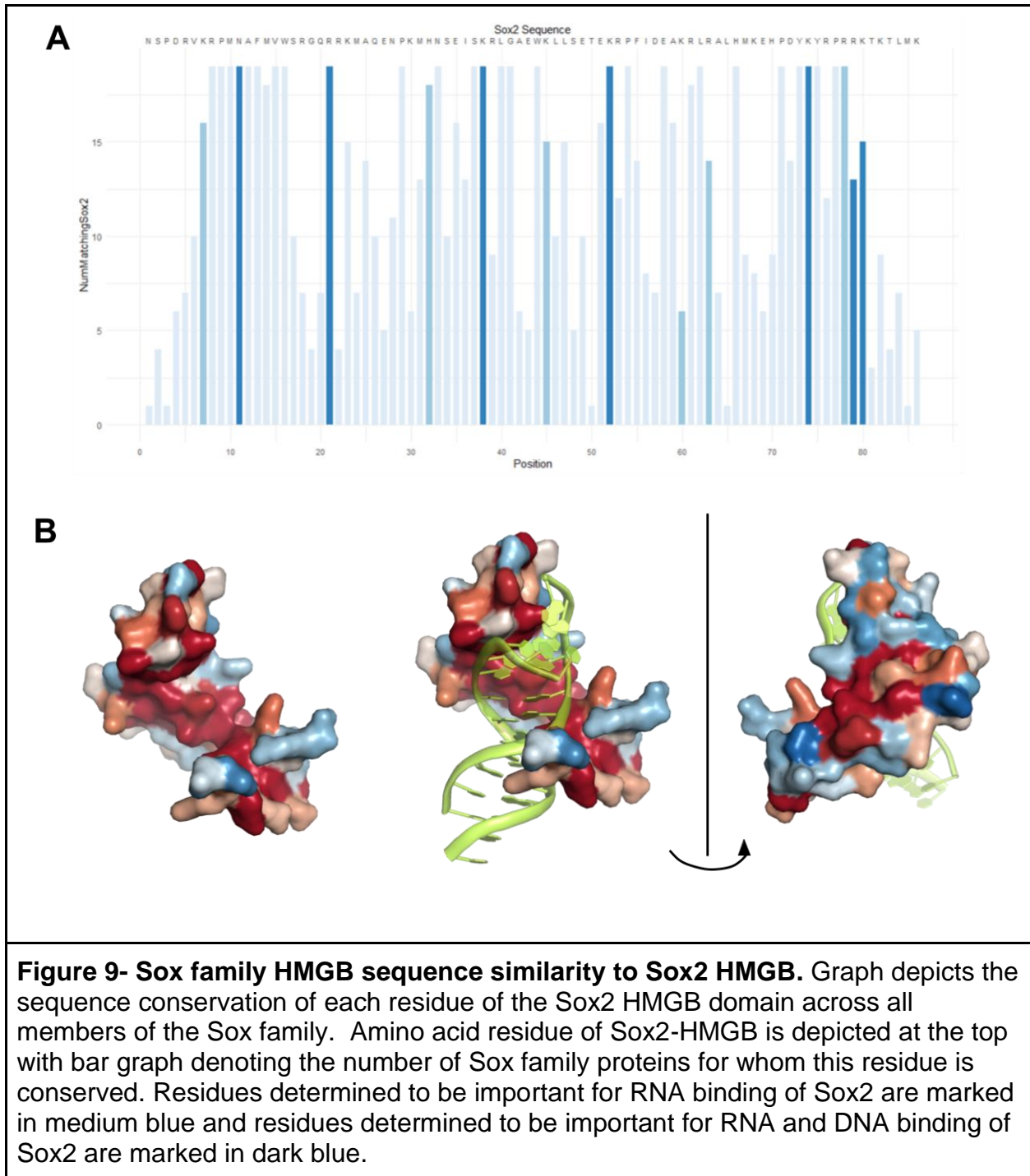
substrates. To this point, we attempt to solve a crystal structure of Sox2 in complex with a potential regulatory target: bulged RNA hairpins. In parallel, we investigate the possibility of Sox2 and LEF1 directly targeting G-quadruplex structures in the genome. We find that, both, Sox2 and LEF1 associate with G-quadruplexes in the genome and that these interactions occur significantly in regulatory regions, likely making these associations relevant for biology. Furthermore, we find that Sox2 binds genomic G-quadruplexes with high affinity *in vitro*, and develop the framework for an *in vitro* workflow to thoroughly assess these interactions.

2. Attempting to Determine the Structure of Sox2 in Complex with RNA through X-Ray Crystallography

2.1 Introduction

As described previously, it has been repeatedly verified that Sox2 directly interacts with RNA *in vivo*. The mechanism by which RNA recognition occurs remains largely unclear. Sox2 binds dsRNA through its DNA-binding HMGB domain.⁶⁰ However, it does not appear to exhibit sequence specificity when bound to RNA.⁶⁰ This indicates that RNA recognition by Sox2 HMGB either occurs through nonspecific or structure specific binding. It has also been suggested that the Sox2 TF has an RNA-binding domain upstream of its DNA-binding domain that permits concurrent binding of DNA and RNA.⁷¹ This upstream domain also may imbue some sequence specificity.

Previous data has indicated that Sox2 is able to adopt a variety of conformations to bind different nucleic acid structures.⁶⁰ This versatility could permit Sox2 to recognize non-B nucleic acids in a structure specific manner and could allow the wide spectrum of nucleic acid binding activities that is observed. However, no structure exists to date showing a different Sox2 binding interface than that used to bind its consensus DNA. A high-resolution structure of the Sox2 HMGB domain in complex with an RNA substrate would provide invaluable insight into the mechanism of Sox2 recognition of RNA and into recognition of non B-form nucleic acids. Specifically, this structure could elucidate whether Sox2 induces a bend into RNA, as it does with B-form DNA. Furthermore, it would allow the RNA-binding interface of Sox2 to be clearly defined.



Additionally, a mutagenic alanine scan of the Sox2 HMGB domain revealed certain important residues of the Sox2 HMGB domain for RNA-binding.⁶⁰ Interestingly, these residues are highly conserved across the Sox family (Fig. 9). This indicates that the

interface used to bind RNA by Sox2 may be representative of the RNA-binding capabilities of the Sox family as a whole.

Sox2 has been shown to bind RNA duplexes and hairpins with high affinity *in vitro*.⁶⁰

Additionally, a SELEX experiment has been completed in a previous study to evolve RNAs with high Sox2 affinity.⁷¹ This experiment also identified a number of hairpin structures, many of which had internal bulges. Considering these data, and because

hairpins and internal bulges are some of the most common RNA structural features, we recently chose to examine the affinity of Sox2 for RNA hairpins relative to internal bulge size (Hamilton et al. manuscript under review, data collected by Abigail Hein). Five RNA hairpins were selected to examine this relationship: a fully base paired hairpin and hairpins with internal bulges of 0+1, 1+2, 2+3, and 3+4 unpaired nucleotides respectively (Appendix 9).

Sox2 HMGB binding affinity for these

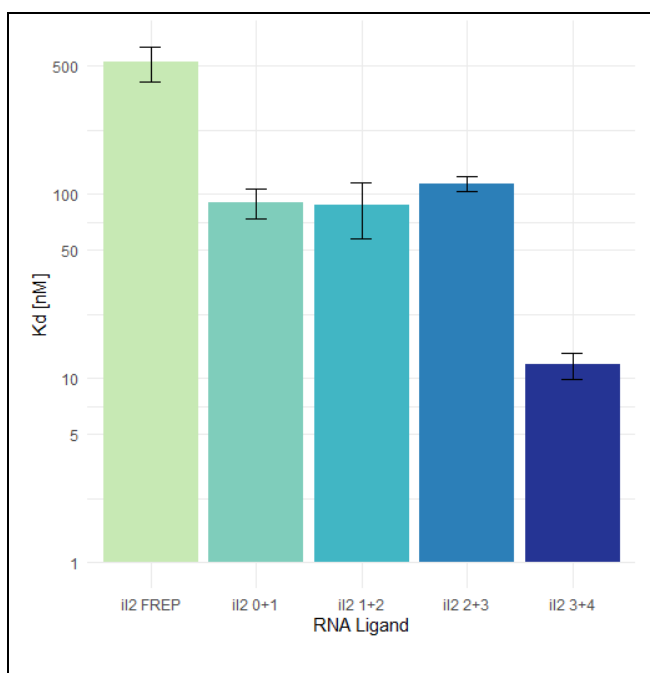


Figure 10- Sox2 binding affinity for internally bulged RNA hairpins. $K_{D,app}$ [nM] of Sox2-HMGB bound to fully base paired hairpin (iL2 FREP) and hairpins with internal bulges of 0+1 (iL2 0+1), 1+2 (iL2 1+2), 2+3 (iL2 2+3), and 3+4 (iL2 3+4) unpaired nucleotides. Y-axis is log scaled and error bars represent standard deviation of $K_{D,app}$

RNAs was measured through fluorescence anisotropy (FA) binding assay. We observed that the Sox2 HMGB exhibits significantly higher affinity for all hairpins with an internal bulge over the fully base-paired hairpin (Fig. 10). Furthermore, Sox2 binds significantly

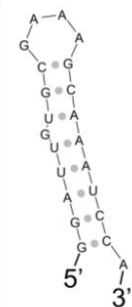

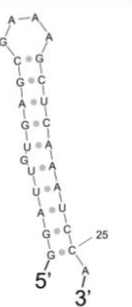
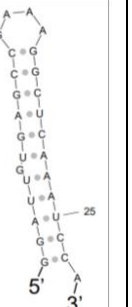
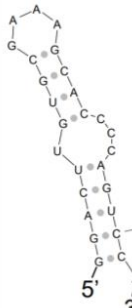

tighter to the hairpin with a 3+4 internal bulge ($K_{d,app} = 12 \pm 2$ nM) than to all other hairpins.

The preferential binding of internally bulged RNAs indicates that Sox2 may have the capability to discriminate between RNA ligands. These data also indicate that internal bulge size may dictate Sox2 affinity, as the 3+4 internal bulge RNA was bound with significantly greater affinity than all other bulged RNAs. However, we did not examine internal bulges of the same size with different sequences, so we cannot eliminate the possibility that the sequence of the RNA contributed to high affinity seen. Put together, this hints that Sox2 may recognize RNAs with structural selectivity. A crystal structure would be an important piece of the picture in determining if base dependent readout plays a role in Sox2 binding or if recognition is purely structural. These observations therefore informed my efforts to attempt to solve the structure of Sox2 in complex with an internally bulged RNA hairpin through X-ray crystallography.

2.2 Results

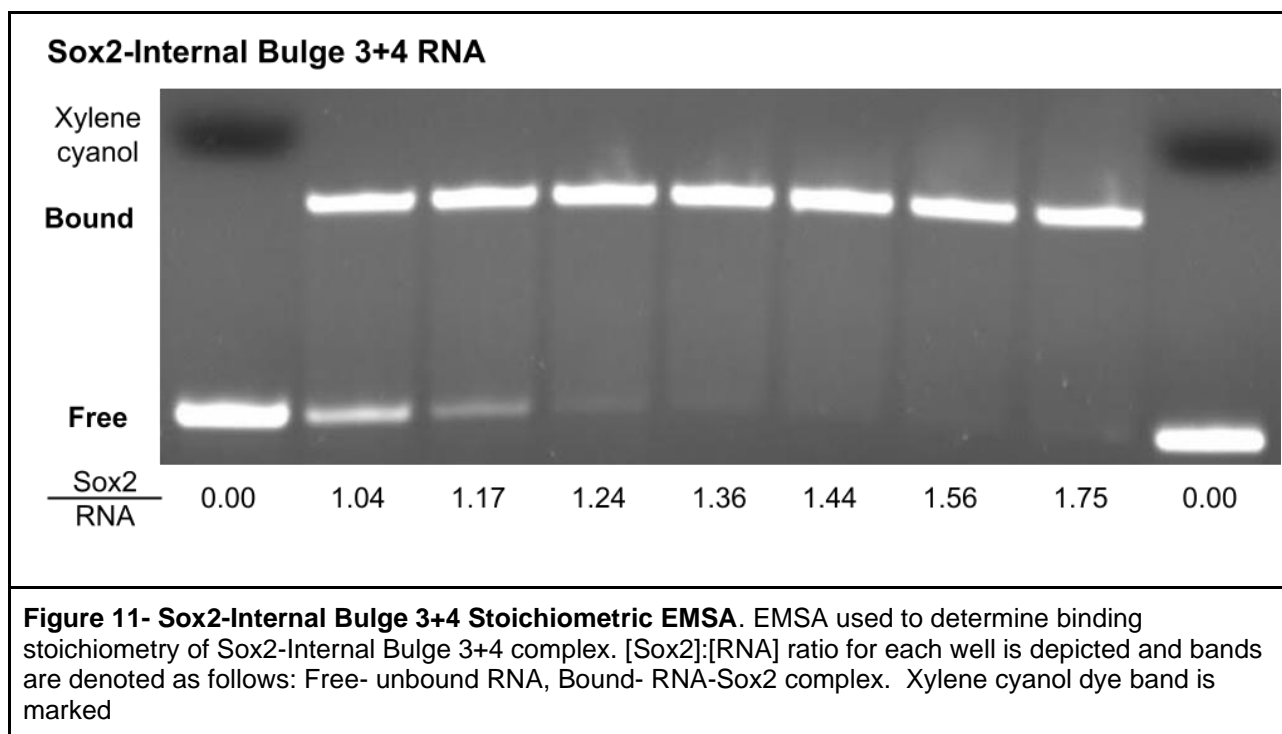
2.2.1 Design of RNA Library

To further evaluate the mechanism of RNA recognition by Sox2, we set out to solve the structure of the Sox2 HMGB domain bound to an RNA hairpin. This required the design of a robust library of RNAs for crystallography trials. Based on the Sox2 preference for internally bulged RNAs, we designed a set of hairpins with internal bulges of sizes 0+1, 1+3, and 3+4 and of varying lengths (Table 1).

Crystallographic RNA Hairpins			
1+0 Internal Bulge			
			
Table 1- RNA crystallography constructs			

All hairpins were designed with two additional features to favor crystallization. First, all RNAs were engineered with a GAAA tetraloop feature, which has been shown to mediate crystal contacts in structured RNAs through loop-loop interactions.⁷² Second, all constructs were designed with an overhanging adenosine nucleotide on the 3' end of the hairpin. Overhanging nucleotides have been repeatedly shown to increase diffraction quality of nucleic acid crystals through interactions with the major or minor groove.^{73–75} Secondary structure was verified by plugging sequences into the Sfold RNA structure prediction software.⁷⁶

The Sox2 protein construct used for crystallographic trials was the minimal HMG domain as defined by UniProt and is the same construct used in previous crystallographic studies.^{77,78} The binding stoichiometry of each Sox2-RNA complex was assessed through electrophoretic mobility gel shift assay (EMSA). This verified that each RNA-protein complex formed a single bound state and that the ratio stoichiometry of each complex was roughly 1:1 Sox2:RNA (Fig. 11, Appendix 1).



2.2.2 Surveying Sox2-RNA Crystallographic Conditions

Crystallization trials were performed for each Sox2-RNA complex using a number of different sparse matrices to broadly sample condition space. This was with the goal of obtaining promising crystal hits that could be further optimized. Specifically, we used the Hampton Nucleic Acid Mini Screen and the NeXtal Nuclix screen, both of which survey conditions that are well documented to favor nucleic acid crystallization.⁷⁶ We also used the Hampton PEG/ion screen and the

NeXtal PEGs II suite because Sox family proteins and Sox-DNA complexes are often crystallized in PEG conditions.^{13,77,79–81}

However, sparse matrix crystallization trials up to this point have yielded very few feasible crystal hits. A recurring issue that has emerged from these complexes is the formation of crystals in conditions that force dissociation of the Sox2-RNA complex. This results in crystals that contain only one component of the complex (Fig. 12).

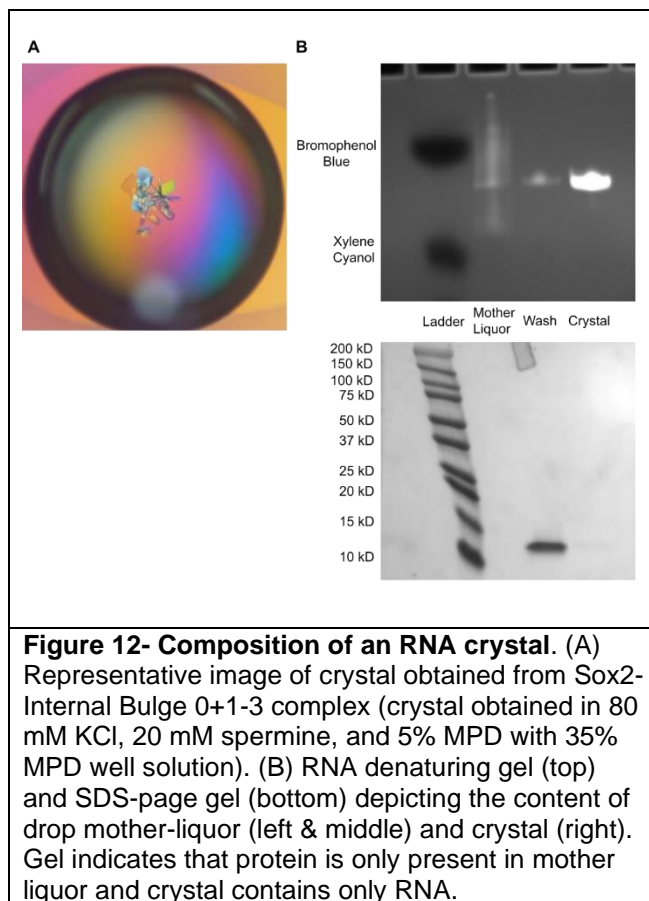


Figure 12- Composition of an RNA crystal. (A) Representative image of crystal obtained from Sox2-Internal Bulge 0+1-3 complex (crystal obtained in 80 mM KCl, 20 mM spermine, and 5% MPD with 35% MPD well solution). (B) RNA denaturing gel (top) and SDS-page gel (bottom) depicting the content of drop mother-liquor (left & middle) and crystal (right). Gel indicates that protein is only present in mother liquor and crystal contains only RNA.

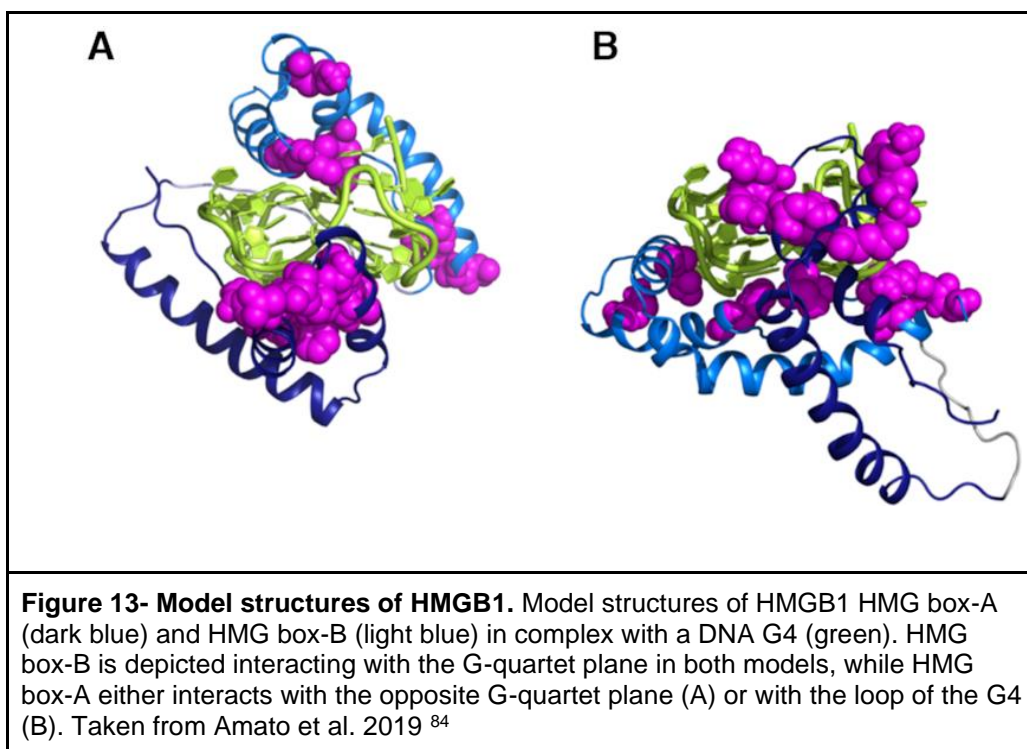
2.3 Discussion and Future Directions

The lack of success in crystallization of a Sox2-RNA complex may be reflective of a number of causative mechanisms, none of which are possible to elucidate from the results of these crystallographic trials alone. For example, it is possible that the RNA constructs designed for this study have multiple binding sites (e.g., the internal bulge and the hairpin loop) that migrate at the same speed on a gel. If this were the case, we would not be able to see this behavior via the analysis we performed. To obtain a crystal structure, in the future it would be necessary to survey more condition space through trials with a greater number of RNA ligands.

3. Developing an *In Silico* and *In Vitro* Pipeline for the Identification of G4-Binding HMGB Proteins

3.1 Introduction

Recently, a number of HMGB proteins have been found to bind G4s, the most prominent examples being HMGB1,-2, and -3. Specifically, HMGB1-3 have been identified as G4 interacting proteins in two independent high-throughput surveys. The first of these studies identified G4 interactions *in vivo* by crosslinking G4-protein complexes,⁴² whereas the second labeled G4 proximal proteins via peroxidase activity with a G4 probe in cell lysate.⁸² These two approaches indicate that HMGB1-3 associations with G4s are likely biologically relevant and not just *in vitro* artifacts. Furthermore, HMGB1-3 genomic binding sites, as identified by chromatin immunoprecipitation sequencing (ChIP-seq), overlap significantly with G4s in the genome (identified by G4 ChIP-seq).^{43,83} *In vitro*, HMGB1 has been shown to bind and stabilize DNA G4s through its two HMG domains.^{84,85} Structural models of these have suggested two possible binding modes for these interactions, in which one HMG box interacts with the quartet plane while the other either (1) interacts with the quartet plane on the opposite end of the quadruplex (Fig. 12A) or (2) with the loops of the quadruplex (Fig. 12B).⁸⁴

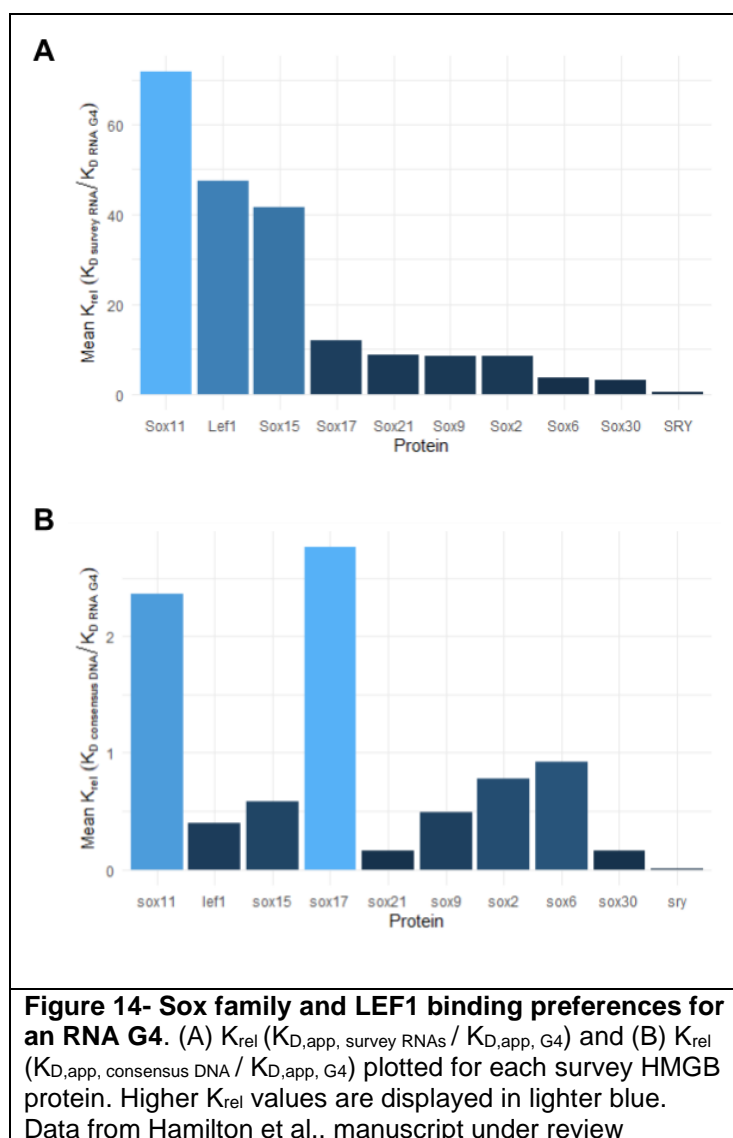


Beyond HMGB1-3, a set of other studies suggest broader interactions of HMGB domain containing proteins and G4s. TFAM, a mitochondrial transcription factor, with two HMG boxes⁸⁶ was found to bind G4s with similar affinity to double-stranded DNA,⁸⁷ and was also identified as a G4 proximal protein in cell lysate by peroxidase labeling.⁸²

In addition, both Sox6 and Lef1 were found to be in the top half of candidate G4-binding TFs in a survey of roughly 500 TFs.⁴³ Put together, these data hint that G4-binding may be an unrecognized behavior of the HMGB family.

Recently, we conducted a survey of the binding affinities of 10 HMGB family proteins (SRY, Sox2, Sox6, Sox9, Sox11, Sox15, Sox17, Sox21, Sox30, and Lef1) in complex with a number of nucleic acid ligands, including a consensus DNA, a nonspecific DNA, an internally bulged RNA hairpin, a segment of a lncRNA, an RNA 4-way junction, and a model RNA G4 (Hamilton et al. manuscript under review, data

collected by Desmond Hamilton). Interestingly, we found that the G4 was the tightest bound RNA structure for 9 out of the 10 of the proteins surveyed.



To further examine G4 selectivity of the HMGB proteins surveyed, we calculated the binding affinity for the RNA G4 relative to the mean binding affinity for all RNAs surveyed (Fig. 14A) and the binding affinity for the G4 relative to the binding affinity for the protein's consensus DNA (Fig. 14B). This analysis revealed a preference for G4 structures by a number of proteins. Sox11, LEF1, and Sox15 all bound the G4 much

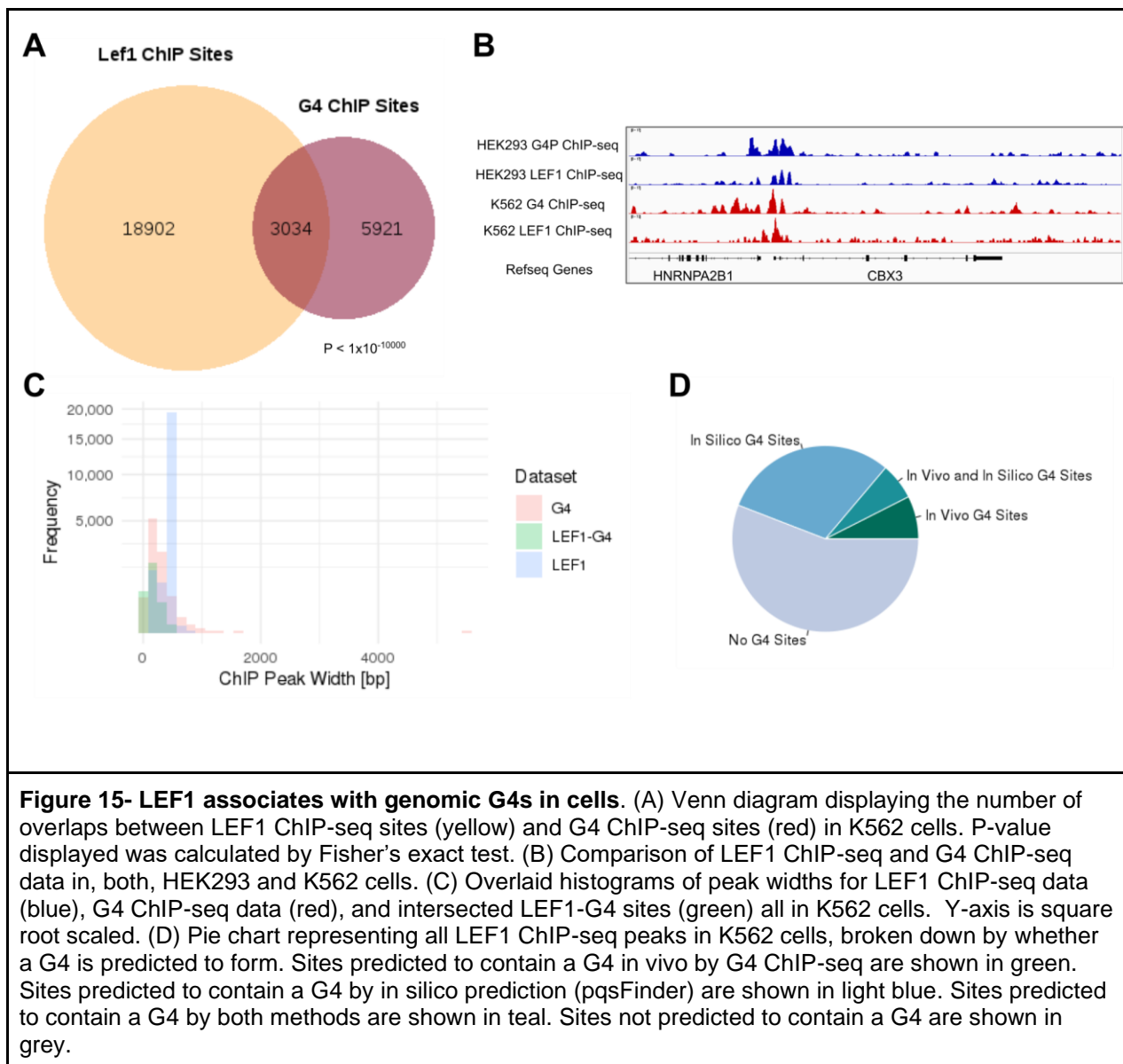
tighter than all other survey RNAs, and Sox11 and Sox17 bound the G4 with higher affinity than their consensus DNA. Given that there is a growing body of evidence for G4-TF interactions in gene regulation, we hypothesized that G4s are an important regulatory target of HMGB proteins.

Based on the identification correlation of LEF1 binding sites with genomic G4s, and the heavy binding preference of LEF1 for a G4 over other RNA ligands, we selected LEF1 for further investigation as a G4-binding protein. We elected to conduct an in-depth examination of genomic DNA G4s in LEF1 binding sites to determine if these structures are a likely regulatory target of LEF1. Supplementarily to our LEF1 analysis, we also examine the interactions of genomic G4s with Sox2 in embryonic stem cells to examine how HMGB transcription factors may interact with genomic G4s in a biologically relevant cell line.

3.2 Results

3.2.1 LEF1 Associates with G4s *In Vivo*

To examine the association of the LEF1 TF with genomic G4s, we took publicly available LEF1 ChIP-seq data (ENCFF659WAF) and G4 ChIP-seq data (GSE107690) in K562 cells. We observed that 33.9% (3034/8955) of identified G4 ChIP -seq sites overlapped with LEF1 binding sites, conversely 13.8% (3034/21936) of LEF1 ChIP-seq sites overlapped with G4 sites in the genome (Fig. 15A, B). This intersection was determined to be significant to $P < 1 \times 10^{-10000}$ by Fisher's exact test. These intervals of overlap between G4 ChIP-seq peaks and LEF1 ChIP-seq peaks are hereafter referred to as LEF1-G4 sites.



The size of the interval regions, or regions of overlap between LEF1 ChIP-seq peaks and G4 ChIP-seq peaks, were then examined to ensure that the overlap of the intervals did not occur only at the edges of peaks. The mean peak width of LEF1-G4 sites was found to be 148.3 base pairs, with 96.67% (2933/3034) of sites larger than 15 base pairs. 15 base pairs were selected as a benchmark because it applies to the minimal size required to form a quadruplex, as defined by the putative quadruplex

forming sequence (“G₃₊N₁₋₇G₃₊N₁₋₇G₃₊N₁₋₇G₃₊”). Furthermore, 92.58% (2809/3034) of LEF1-G4 sites were wider than 30 base pairs. The distribution of peak widths of LEF1-G4 sites was also observed to be similar to the distributions of LEF1 and G4 peak widths (Fig. 15C).

This analysis was repeated in HEK293 cells to verify that these trends were not specific to K562 cells. We obtained very similar results in both cell lines (Fig. 1B, Appendix 2). In HEK293 cells, 2213 LEF1-G4 sites were identified out of 40790 G4 sites (GSE133379) and 3527 LEF1 sites (ENCFF333UCS) in this cell line ($P < 1 \times 10^{-8000}$).

If LEF1 is in fact binding G4 structures in the genome, we might expect LEF1 to localize to its putative G4 binding sites through recognition of G4 structure rather than through recognition of the protein’s consensus binding motif. If this were the case, we would consequently expect LEF1-G4 sites to be significantly depleted for the LEF1 consensus sequence. In the K562 cell line, 28.7% of LEF1 ChIP-seq sites were found to have the core 5'-TCAAAG LEF1 consensus sequence, whereas only 4.05% of LEF1-G4 sites were found to have the consensus sequence. This difference is significant to $P < 2.2 \times 10^{-16}$, indicating that LEF1 is likely binding these G4 sites through an alternative recognition mechanism than its consensus binding activity. This trend is also conserved in HEK293 cells (Appendix 2C).

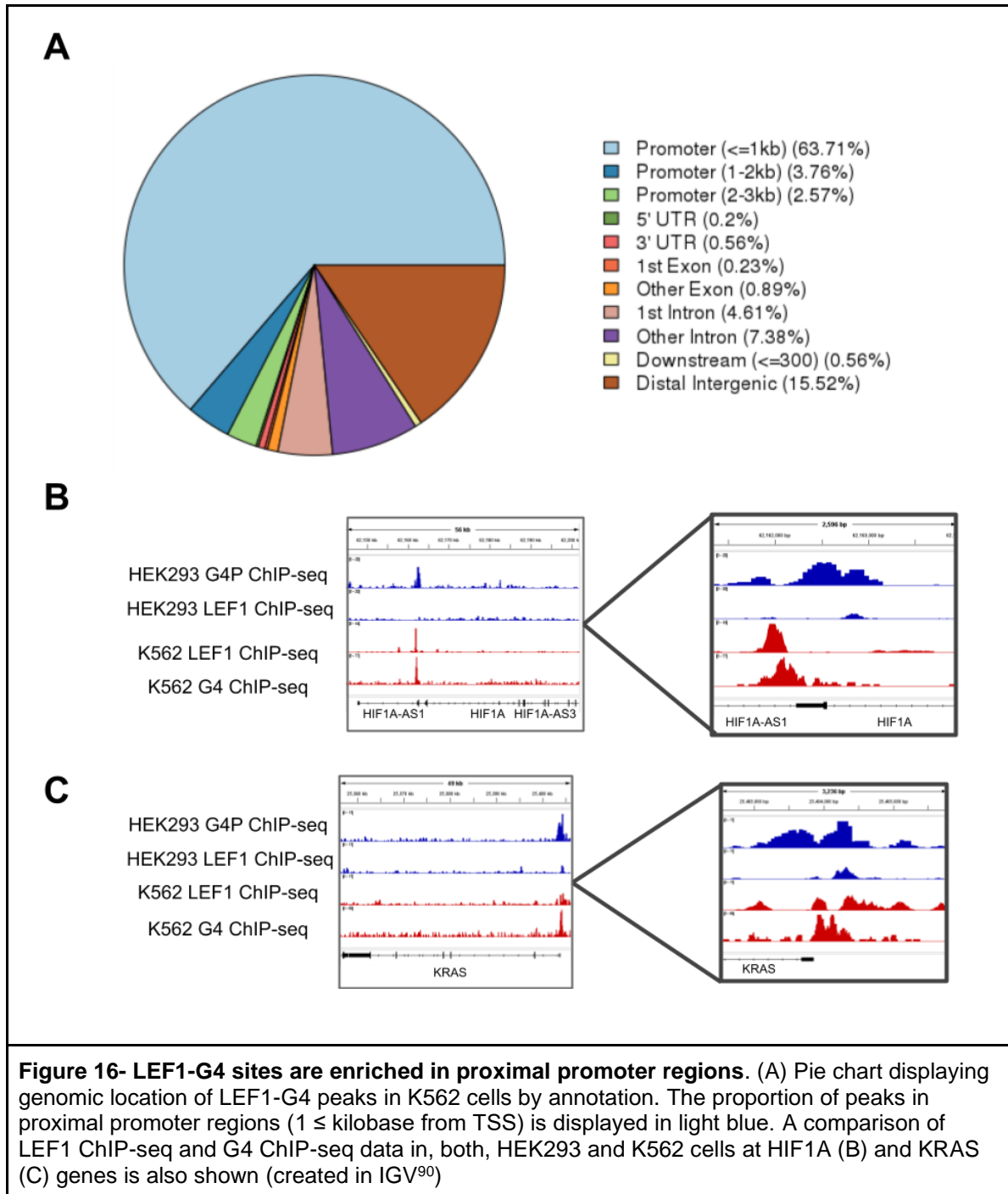
In order to increase confidence in the Lef1-G4 sites, we ran the LEF1 datasets through pqsFinder,⁸⁸ a G4 sequence prediction software. The intention of this step was to eliminate any LEF1-G4 hits that are an artifact of incidental localization of LEF1 to regions of active transcription where G4s happen to be enriched. This analysis predicted at least one high confidence quadruplex in 36.6% (8040/21936) of LEF1

ChIP-seq sites in K562 cells. When the sequences of the LEF1 ChIP-seq sites were shuffled, the number of sites that were predicted to have a G4 dropped by over 50% to 3612/21936. This indicates that the correlation between LEF1 binding and G4 presence is not just a function of G-richness in binding sites. When looking at K562 cells, 46.0% (1395/3034) of LEF1-G4 sites were predicted to have a quadruplex by pqsFinder (Fig 15D). It should, however, be noted that computational G4 prediction often fails to identify imperfect G4s, such as G4s with long loops, mismatches in the G-quartet, or bulges in between quartets.^{88,89} Nevertheless, imperfect G4s have been observed *in vivo*.^{88,89} pqsFinder is designed to tolerate imperfect G4s, however it likely still suffers from these shortcomings. These data are consistent in both cell lines (Appendix 2D and E).

3.2.2 LEF1 Associations with G4s are Enriched in Functional Regions of the Genome

Next, we sought to determine if LEF1 binding to G4s in the genome could have functional implications. To address this, we annotated the LEF1-G4 dataset to elucidate where these interactions occurred in the genome. Given that G4s have been found to be enriched in promoter regions,³² we would expect LEF1-G4 putative binding events to be enriched in promoter regions of the genome. This is consistent with our observations. In K562 cells, we found that 24.2% and 76.0% of LEF1 ChIP-seq and G4 ChIP-seq peaks, respectively, localized to proximal promoters in the genome (proximal promoters here refer to regions ≤ 1 Kilobase from the transcription start site (TSS)). From this we would expect 39.2% of LEF1-G4 sites to occur in these regions by chance

alone. Strikingly, we observe that 63.7% of LEF1-G4 sites are observed in proximal promoters, which is 16% more sites than we would expect to see by chance alone ($P < 2.2 \times 10^{-16}$) (Fig. 16A). This enrichment is also consistent in the HEK293 cell line, in



which 84.68% of LEF1-G4 sites are found in proximal promoters ($P < 2.2 \times 10^{-16}$ by proportion test).

We then proceeded by investigating which genes are regulated by promoters that contain LEF1-G4s. We constrained our analysis to LEF1-G4 sites that had been predicted *in vivo* and *in silico* (1395/3034 sites in K562 cells and 1619/2213 sites in HEK293 cells) to ensure we obtained the highest confidence hits possible. LEF1-G4s were found in the proximal promoter region of 1286 genes in K562 cells and 847 genes in HEK293 cells. There were 136 genes common to both datasets. Interestingly, we also identified LEF1 binding at established G4 sites in the genome. Specifically, a LEF1-G4 site was observed in the HIF1A promoter, which has been implicated as a promoter that forms G4s *in vivo*⁹¹ (Fig. 16B). There is also evidence to suggest that LEF1 regulates the expression of the HIF1A antisense-1 lncRNA (29369172), indicating that LEF1-G4 associations in this promoter may be important for gene regulation. LEF1 was also found to bind the region that is known to form three G4s in the KRAS promoter⁹² (Fig. 16C).

3.2.3 Verification of LEF1-G4 association by G4-seq

One significant caveat of using G4 ChIP-seq data is that this protocol uses a G4 specific antibody (BG4) to pull down G4s in the genome.⁹³ This antibody displays high affinity for G4s ($K_d \leq 2$ nM) and displays selectivity for G4 structure over other nucleic acid structures.⁹⁴ Therefore, it is possible that BG4 induces G4 structure from G-rich sequences that would not form a G4 in native conditions, and are therefore not biologically relevant. We elected to use G4 ChIP-seq datasets because it is the most

common G4 sequencing method, and therefore has the most extensive body of publicly available data. However, to verify that the LEF1-G4 associations we see are not artefacts of the BG4 antibody, we compared the LEF1-G4 associations seen in HEK293 cells to G4s mapped in the same cell line by G4-seq (GSM3003539). G4 structure is known to cause polymerase stalling, therefore G4-seq maps sites in the genome where polymerase stalling occurs in the presence of potassium ions (G4 stabilizing), but not in the presence of lithium ions (not G4 stabilizing).⁹⁵ This method uses biologically relevant G4 stabilizers (potassium) and as such it is less likely that it induces G4s that do not form in the biological context. We observe that 24.2% (855/3527) of LEF1 sites are overlap with a G4-seq site. This correlation is smaller than that seen between G4s identified by G4 ChIP-seq, but the association is still significant to $P < 1 \times 10^{-243}$ (Appendix 3). Furthermore, 75.7% of LEF1-G4 sites predicted by G4-seq were also predicted by G4 ChIP-seq (Appendix 3). The G4 ChIP-seq and G4-seq datasets used for this analysis were correlated to $P < 1 \times 10^{-29105}$, indicating that G4s mapped in both datasets are consistent. It should be noted that the G4-seq method does not account for chromatin state, and as such likely includes many putative G4s that would not form in vivo.⁹³ This likely accounts for the large number of putative G4 sites observed by G4-seq (Appendix 3). For this reason, we only used G4-seq data to verify the trends seen with G4 ChIP-seq.

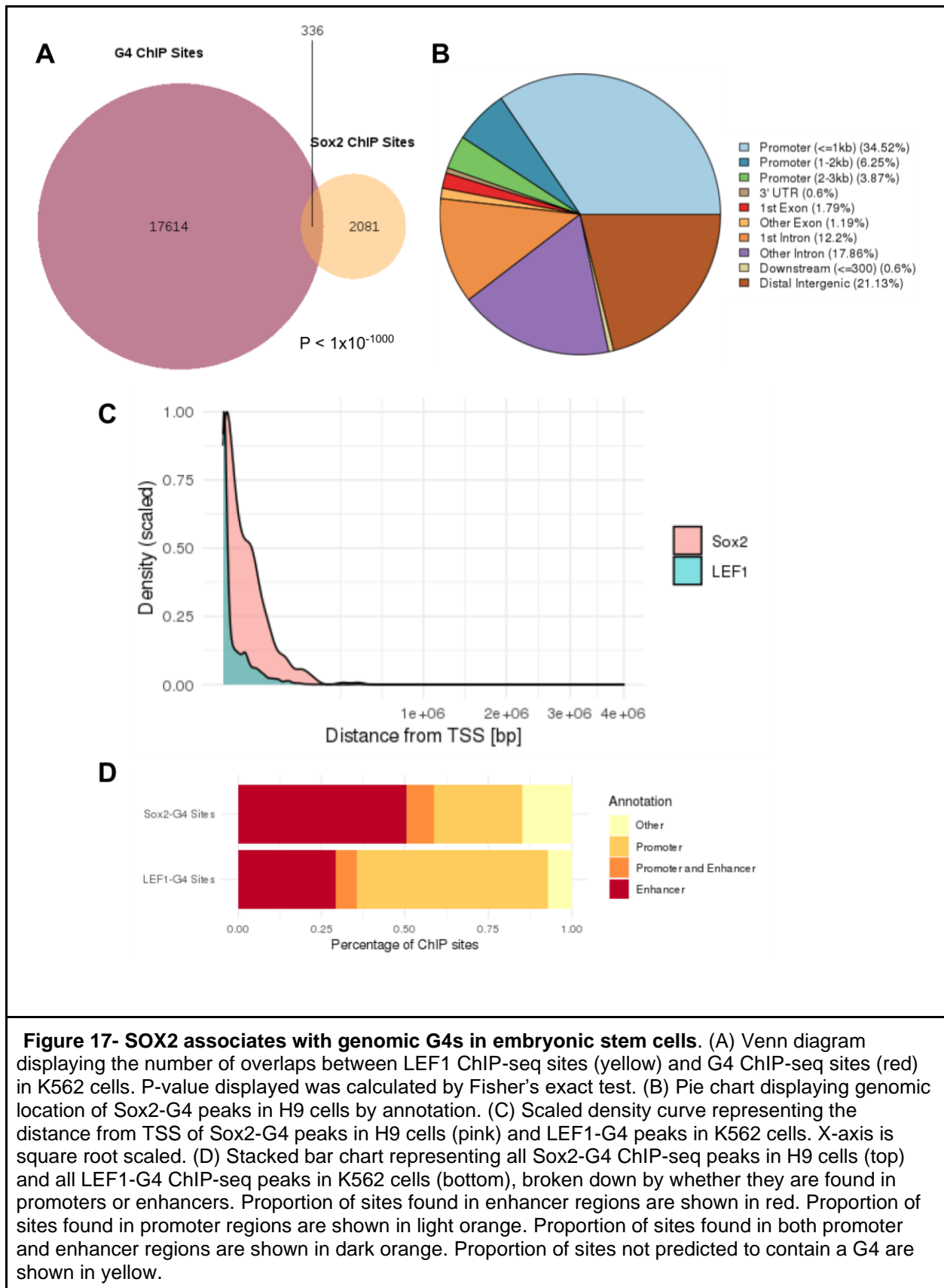
3.2.4 Sox2 Associates with G4s in Embryonic Stem Cells

We next sought to examine if there was any evidence to suggest that HMGB TFs associate with G4s in stem cells. Stem cells have a radically different epigenetic

landscape relative to differentiated cells and express a unique set of TFs.^{96,97} As previously mentioned, the Sox and TCF/LEF families of HMGB proteins play important roles in stem cell differentiation and in maintaining pluripotency.^{70,98,99} Therefore, examining the interaction between G4s and HMGB proteins in stem cells may provide more biologically relevant insight into this family's interactions with G4s in the genome.

For this line of inquiry, we again used publicly available G4 ChIP data (GSE161531) from a human embryonic stem cell line (H9 cells) and Sox2 ChIP data in the same cell line (Cistrome ID: 44233). Sox2 was selected because its activity in embryonic stem cells is well documented,¹⁰⁰ and this protein had a pre-existing ChIP dataset in H9 cells. We completed the same analysis detailed previously with these datasets to characterize putative Sox2-G4 binding sites (Fig. 17A, Appendix 4).

This pipeline revealed that Sox2-G4 sites in the genome are much less prevalent than LEF1-G4 sites. Only 336 overlapping sites were identified between G4 and Sox2 datasets. Therefore, G4s only account for 13.9% (336/2417) of Sox2 binding events (Fig. 17A). However, this correlation is still highly significant to $P < 1 \times 10^{-1000}$ by Fisher's exact test.

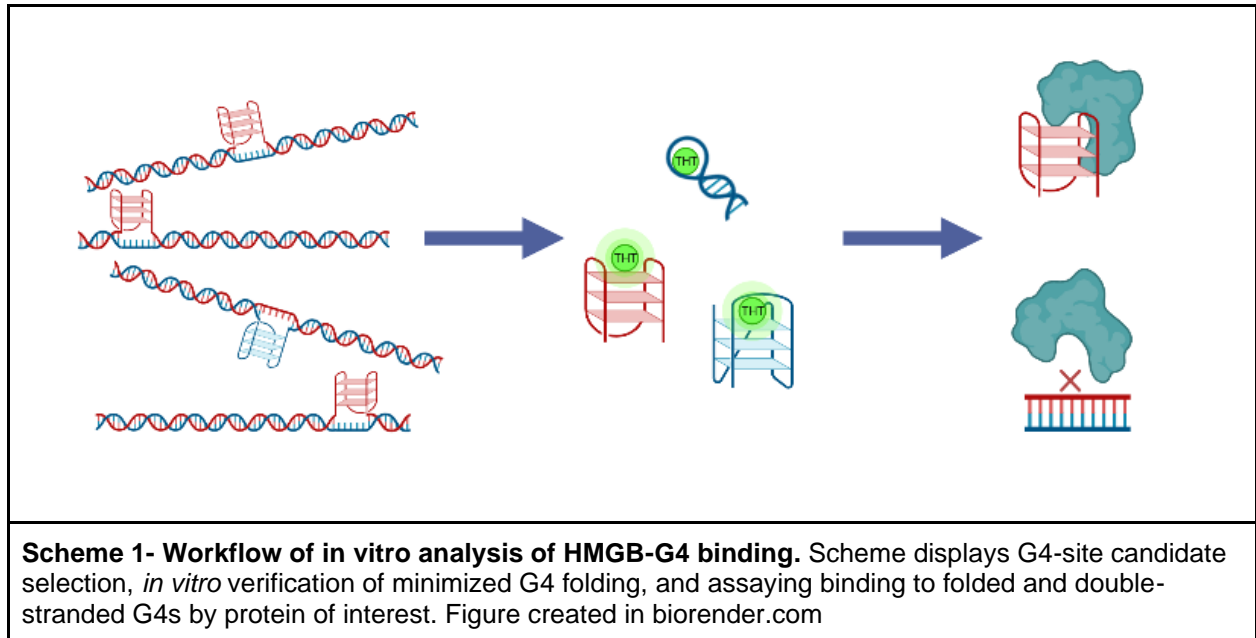


In sharp contrast to the behavior observed in LEF1-G4 sites, Sox2-G4 sites are not significantly enriched in promoter regions. Rather, the number of Sox2-G4 sites found in proximal promoter regions (34.5% of Sox2-G4 sites) is depleted from the expected proportion of 59.7% ($P=1.3 \times 10^{-10}$ by proportion test). The majority of Sox2-G4 sites are, instead, found in intergenic regions, or within introns (Fig. 16B). Consistent with this observation, Sox2-G4 sites are distributed further from the TSS than LEF1-G4 sites (Fig. 17C). This difference in Sox2-G4 vs LEF1-G4 site location is more pronounced than the general population of Sox2 vs LEF1 binding sites (Appendix 5). Sox2 is known to localize to enhancer regions in the genome,^{101–103} so we then investigated if Sox2-G4 interactions are found in enhancer regions rather than promoters. To this end, we calculated the proportion of Sox2-G4s that are found in enhancers in H9 cells. Our H9 enhancer dataset was obtained from Enhancer Atlas 2.0.¹⁰⁴ From this, we determined that 58.6% (197/336) of Sox2-G4 sites are found in enhancers, whereas less than 5% of LEF1-G4 sites were found in enhancers. These data further hint that different HMGB TFs may interact with a specific, distinct subset of G4s in the genome that are important for TF function.

3.2.5 Sox2 binds Genomic G4 Sites *In Vitro*

To fully understand the nature of HMGB protein interactions with G4s, it is necessary to characterize protein-G4 interactions *in vitro*. To this end, we piloted a set of experiments designed to validate the results of our bioinformatic analysis (Scheme 1). It should be noted that these experiments were conducted before the results of the

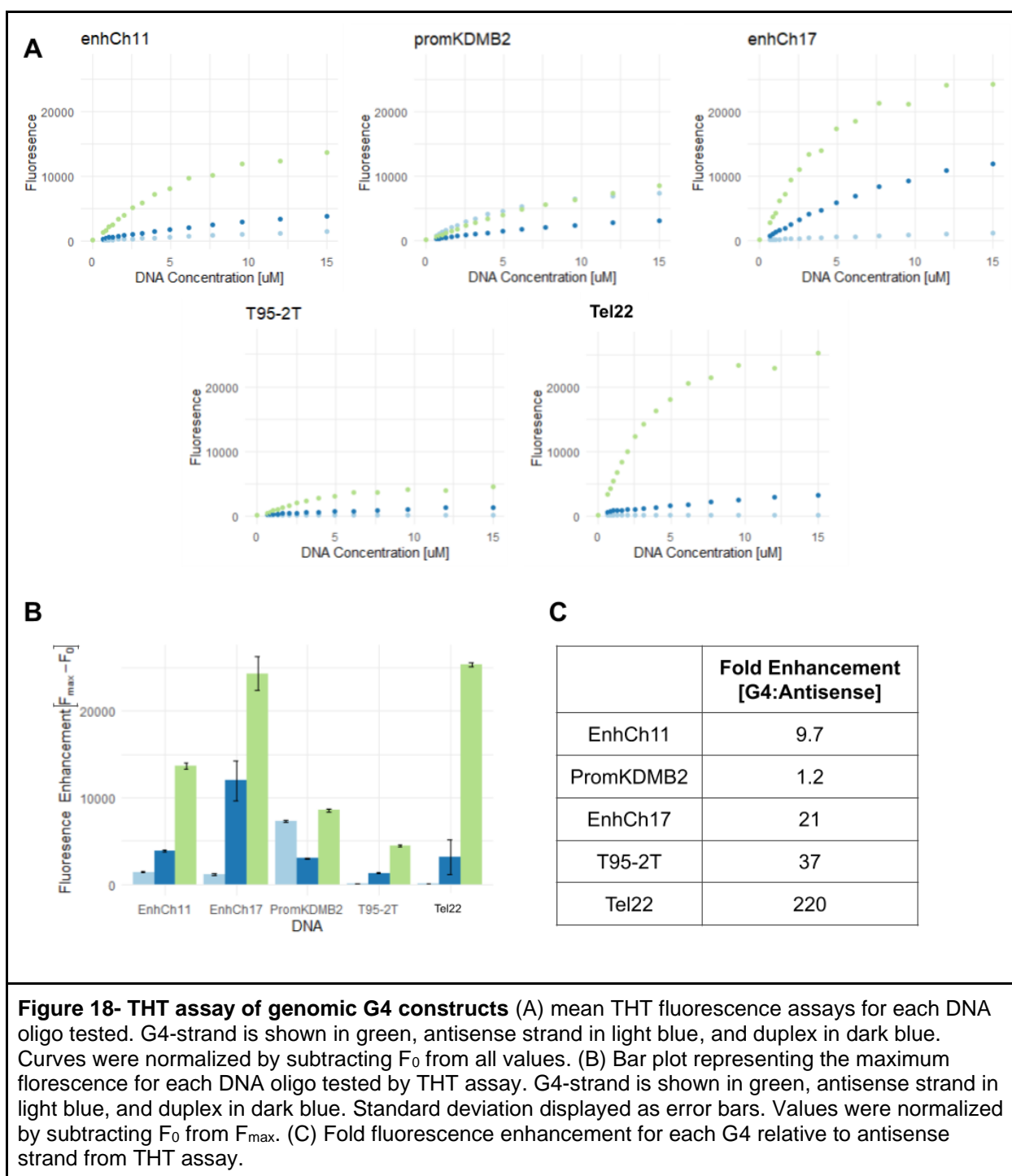
bioinformatic analysis had come to full fruition. As such these experiments are intended to serve as a proof of concept for the *in vitro* analysis that could be completed using the results from sections 3.2.1-3.2.3.



To assess the capability of Sox2 to bind G4s, we selected a curated suite of high quality G4s from Sox2-G4 sites in the human embryonic H1 cell line. At the time of experiment, the embryonic stem cell G4 ChIP-seq dataset analyzed in section 3.2.3 had not been published. Rather, to select a preliminary dataset for experimentation, we used DeepG4,¹⁰⁵ a computational, deep learning tool that predicts cell-type specific active G4s, as a proxy for a G4 ChIP-seq dataset in embryonic stem cells. We selected three Sox2-DeepG4 sites for *in vitro* characterization (Appendix 6). Two of these sites were in enhancer regions in H1 cells (hereafter referred to as enhCh11 and enhCh17).¹⁰⁴ The third Sox2-DeepG4 site was in the KDMB2 promoter, a gene that has been shown to be directly regulated by Sox2 (hereafter referred to as promKDMB2).¹⁰⁶ Each site was selected based on the presence of a high quality G4 predicted by pqsFinder (quality

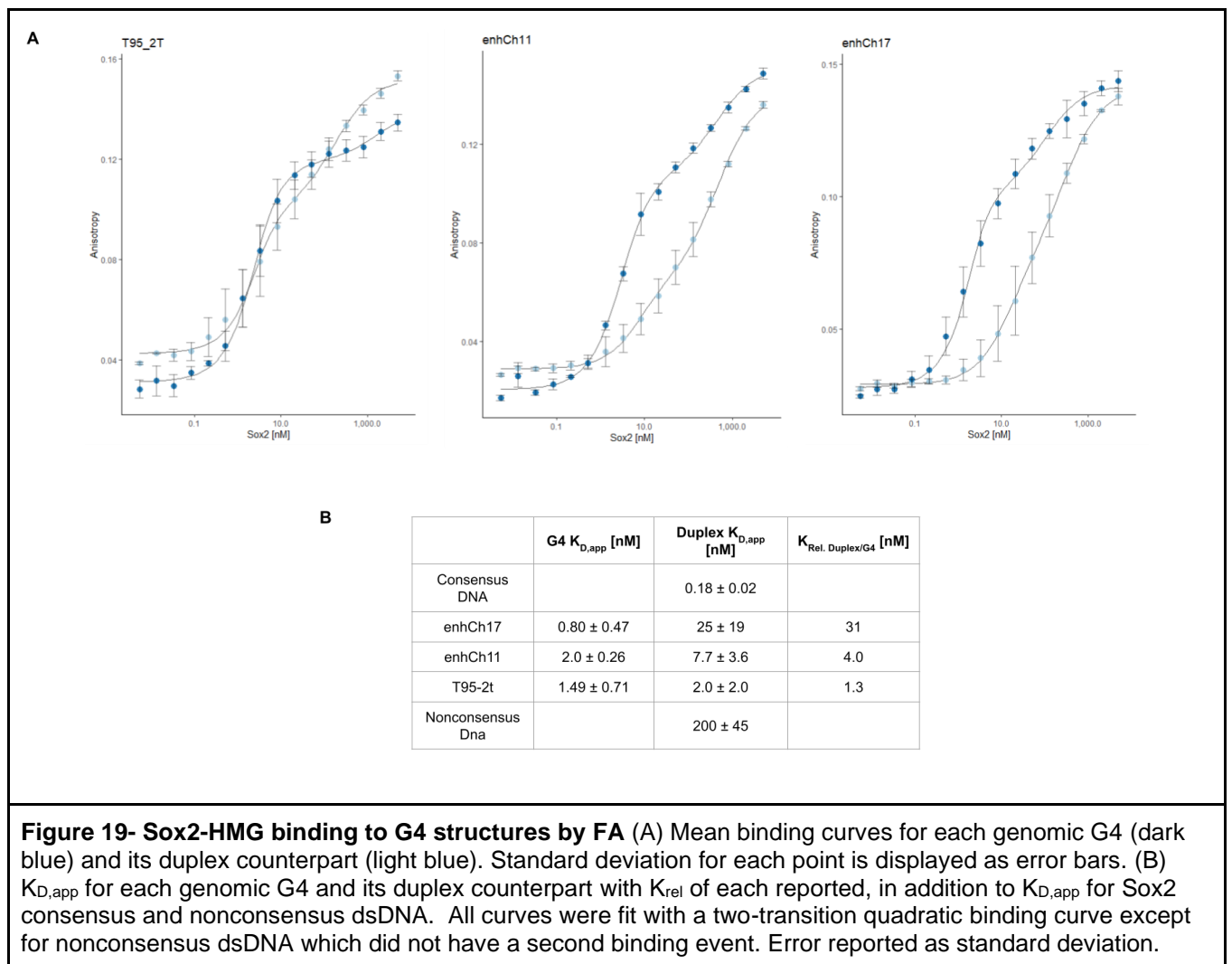
was determined by the score given to the putative quadruplex-forming sequence by pqsFinder). We then generated minimized DNA constructs that encapsulated each predicted G4 (Appendix 6). All G4s used for *in vitro* characterization had a pqs score ≥ 70 .

To verify that our minimalized genomic G4 selections form a G4 *in vitro*, we conducted a Thioflavin T (THT) fluorescence assay. THT has been shown to act as a G4 specific fluorescent probe for DNA and RNA G4s.^{107–109} The THT assay was conducted with the three genomic G4s and two well characterized model DNA G4s. The model G4s used in this study were d[TT(GGGT)₄] (hereafter referred to as T95-2T) and d[A(GGGTTA)₄] (hereafter referred to as tel22), both of these were selected because they are well characterized G4s that have been extensively studied *in vitro*.^{110,111} We performed a THT titration to measure the fluorescence of the sense, antisense, and double-stranded variants of each construct (Fig. 18A). To determine if the construct formed a G4, we calculated the fold enhancement of fluorescence of the sense strand relative to the antisense strand (Fig. 18B,C). Constructs with a fold enhancement less than 5 were eliminated from consideration. From this analysis the promKDMB2 genomic G4 construct was determined not to form a viable G4 *in vitro*. This is likely due to the high GC-content of the construct, as this makes it more likely to form hairpins or self-dimers.



We then sought to determine if Sox2 displayed G4-binding structural selectivity. To assess this, we performed a fluorescence anisotropy binding assay of the Sox2 HMGB with several nucleic acid ligands. We tested binding with FAM-labeled T95-2T,

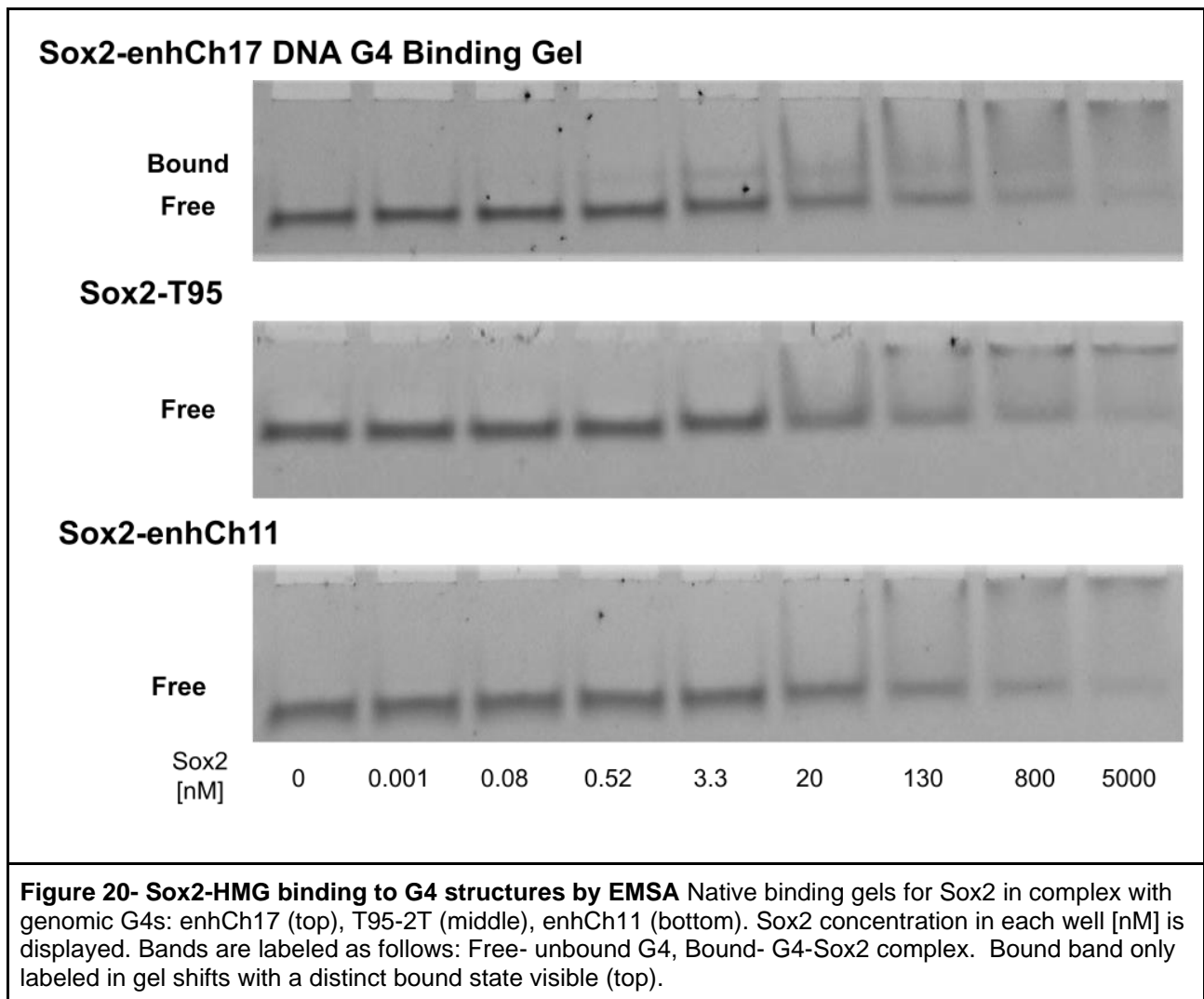
enhCh11, enhCh17 G4s and with Sox2 consensus and non-consensus dsDNAs. Sox2 appeared to exhibit multiple binding events to all three G4s, as binding curves appear to have two distinct transition states (Fig. 19A). Binding affinity of Sox2 for DNA G4 structures was found to rival affinity for its consensus DNA (Fig. 19B). The $K_{D,app}$ of the first transition of all three G4s measured ranged from 0.8 to 2.0 nM. While these affinities are higher than the $K_{D,app}$ observed for consensus dsDNA (0.18 ± 0.02 nM), they are all magnitudes lower than the $K_{D,app}$ observed for non-consensus dsDNA (200 ± 45 nM). These affinities are well within the range of biological relevance.



Sox2-G4 binding assays for G4 constructs were also conducted by native EMSA to validate results observed by FA. While a clear bound band did not appear on the gel for these constructs, we did observe that unbound G4 depleted as the concentration of Sox2 was increased in the titration (Fig. 20). We attribute the lack of emergence of a distinct bound band to the nature of EMSA binding assays, in that this assay is not an equilibrium assay, in that it requires the separation of bound and unbound states. As such there are a number of factors that may cause this effect, such as a high k_{off} for the protein-G4 complex. Notably, in stoichiometric conditions we do observe a slight bound band for the Sox2-T95-2T complex (Appendix 7)

To examine if binding affinities observed in FA were specific to G4 structure, we measured binding affinity for the corresponding double stranded DNA construct for each G4 surveyed. In an attempt to eliminate all G4 formation, unlabeled antisense DNA was added in 1.5-fold excess to the labeled G4 strand. The complex was monitored on a native gel to ensure the formation of duplex and elimination of G4 at this concentration (Appendix 8). It should be noted that, even at 1:1.5 [G4:antisense], G4 still formed for the T95-2T construct. We observe a marked shift in the binding curves for dsG4 constructs of enhCh17 and enhCh11, indicating that Sox2 displays lower affinity for these dsDNA over G4-DNA (Fig. 19). However, the exact magnitude of this shift in affinity is impossible to determine from these data, as the duplex binding concentrations do not appear to fully saturate in the window of measurement. We observed next to no drop in affinity for dsT95-2T, however this may be due to the difficulty in eliminating G4 formation for this construct (Appendix 8). Ultimately, Sox2 does appear to bind G4 structures with some degree of structural selectivity, however it appears that Sox2 is still

able to effectively bind the duplex counterparts of the G4. If this observation is accurate, and not an artefact of any G4 that has not base paired into duplex form, it could indicate that G4 primary sequence allows for Sox2 binding in the duplex state and G4 structure allows for Sox2 binding in the folded state.



3.3 Discussion

Together, these data support the narrative of G4s as functionally relevant targets of HMGB TFs. From our bioinformatic analysis, we find that LEF1 binding sites across multiple cell lines correlate significantly with G4s in the genome, indicating that LEF1

likely associates with G4s in cells and that G4s may represent a distinct subpopulation of LEF1 binding sites in the genome. However, these data were collected in two immortalized cell lines, as such it is possible that observed LEF1-G4 sites are not entirely reflective of the LEF1-G4 interactions we would see in native context. For example, HEK293 cells are derived from kidney cells, but LEF1 is not detected in kidney tissue in the native context (determined in proteinatlas.org).¹¹² Therefore, LEF1 is likely overexpressed in this cell line and could be binding irrelevant sites in the genome.

A more biologically relevant cell type to explore HMGB-G4 interactions would be to examine these interactions in cells that are not terminally differentiated. To this point, the G4 landscape in the genome has been found to be very different in stem cells.¹¹³ Sox2 is observed to significantly associate with G4s in these embryonic stem cells, though the association is markedly less than LEF1-G4s. Additionally, the apparent preference of Sox2 to bind G4s in enhancers relative to LEF1 seems to indicate that TFs may localize to specific subsets of genomic G4s.

We posit that LEF1 and Sox2 interact with G4s in regulatory regions of the genome, and that these interactions are likely mediated by direct binding of G4 structures by the HMG box of these proteins. Furthermore, our observation that Sox2 preferentially binds G4 structure, but still has high affinity for duplex G4 DNA could hint at a mechanism of Sox2 binding that allows Sox2 to remain at a target site after the dsDNA has melted to form a transcription bubble, leading to high levels of gene expression. This mechanism would be consistent with the observation that G4 formation in regulatory regions is correlated with high gene expression.³³

However, while these data offer interesting preliminary evidence of the G4 binding potential of HMGB proteins, these data are still extremely preliminary. The bioinformatic analysis detailed here shows correlation between G4s and LEF1 binding sites. There is no definitive proof of causation. Furthermore, binding assays revealed that Sox2 still appeared to exhibit high affinity for the G4 sequence as a duplex indicating the possibility that these proteins are associating with G-rich sequence rather than G4 structure.

4. Conclusion and Future Directions

The work detailed in this thesis provide promising evidence that suggests that G4 binding may be a conserved function across multiple families of HMGB proteins. The TFs studied in this work, Sox2 and LEF1, are not considered to be G4 binding proteins, making our findings highly novel. The difference in G4 associations between Sox2 and LEF1 indicate that there is some factor that makes LEF1 localize more significantly to G4s in the genome. We hypothesize that the HMGB domains as a whole bind many diverse nucleic acid structures with high affinity, but individual HMGB proteins (or HMGB subfamilies) are perhaps tuned to exhibit greater selectivity for a sub-portion of these structures, such as G4s.

This work provides illuminates several avenues for further investigation of TCF/LEF and Sox family interactions with G4s. From a bioinformatic angle, the G4 association is wholly undocumented for the rest of the TCF family. Performing the analysis detailed here with available TCF ChIP data would provide powerful insight into the G4-TCF/LEF family associations that cannot possibly be gleaned from examining LEF1 on its own. Also, as discussed previously, performing a LEF1 ChIP-seq experiment in a non-immortalized cell line (e.g. embryonic stem cells) would provide more realistic insight into LEF1-G4 associations. Additionally, this would allow for direct comparison to Sox2-G4 interactions. Furthermore, another major caveat to the analysis described here is that it can only show correlation between G4s and TF binding. A more robust approach would require repeating ChIP-seq with some type of knock-down of G4-TF interactions. This could be done through the use of a G4 stabilizing ligand that prevents TF binding to the G4 (e.g. PDS³⁸).

Finally, fully characterizing Sox2 or LEF1 interactions with G4s would require a far more robust *in vitro* analysis than the approach used here. Specifically, it would require an orthogonal approach to verify G4 formation such as circular dichroism or a foot printing experiment. Furthermore, to verify that the protein is binding in a G4-specific manner, all binding G4 assays should be carried out in Li and K in parallel. As G4s are not stabilized by Li²⁹, these assays should have reduced binding affinity. Duplex formation would also need to be verified more quantitatively. The binding analysis completed in this thesis was completed under heavy time constraint and was intended to serve as proof of concept for more in-depth experimentation. The results of the bioinformatic work can be easily curated for an *in vitro* survey with high biological relevance for this purpose.

5. Materials and Methods

5.1 Protein Purification

Sox2 HMGB domains with a His8-MBP affinity tag fused at the N-terminus (Appendix 10) were cloned into pET30b cells via circular polymerase extension cloning (CPEC).¹¹⁴ Sequence verified vectors were transformed into Rosetta(DE3)/pLysS E. coli cells and grown to saturation for 12-18 hours in 2xyt media. Cells were then inoculated in LB media with 50 µg/mL kanamycin and 34 µg/mL chloramphenicol and grown at 37 °C to an OD600 of 0.4-0.6. Vector expression was induced by the addition of 1 mM IPTG for 4 hours at 37 °C followed by centrifugation of cells at 4000 xg at 4 °C for 30 minutes. The resultant cell pellet was resuspended in lysis buffer (1.5 M NaCl, 50 mM NaH₂PO₄, 50 mM imidazole, pH 8, 10% glycerol, 1x protease inhibitor cocktail (Roche, cat. #04 693 132 001)) and lysed in a C3 cell homogenizer. Cell lysate was centrifuged at 15,000 xg at 4 °C for 30 minutes. Supernatant was subsequently mixed with lysis buffer equilibrated Ni-NTA agarose resin (Thermo Scientific) for 30 minutes at 4 °C. The resin was washed thrice with wash buffer (1.5 M NaCl, 50 mM NaH₂PO₄, 50 mM imidazole, pH 8, 10% glycerol). Tagged proteins were eluted off the Ni column with 300 mM NaCl, 50 mM NaH₂PO₄, 300 mM imidazole, pH 8, and 10% glycerol at 4°C. To remove the MBP tag, eluent was buffer exchanged into cleavage buffer (150 mM NaCl, 50 mM NaH₂PO₄, pH 7.5, 10% glycerol, 1 mM DTT) and incubated with His8-MBP tagged 3C protease for 12-16 hours at 4 °C. The cleaved protein solution was added to wash buffer equilibrated Ni-NTA agarose resin, mixed gently for 15 minutes and subsequently eluted with wash buffer. The HMGB domain was then buffer exchanged into storage buffer (10 mM Tris-HEPES, 135 mM KCl, 15 mM NaCl, pH 7.5, 10% glycerol) in order to

run the protein through a size exclusion column (HiLoad 16/600 Superdex G-75 prep grade column (GE Life Science)). Fractions containing the Sox2-HMGB were collected, pooled, and stored in storage buffer 2 (10 mM Tris-HEPES, 135 mM KCl, 15 mM NaCl, pH 7.5, 45% glycerol) at -20 °C. Protein purification success was validated by SDS-Page and Fluorescence anisotropy binding assay with consensus DNA. Representative SDS-PAGE gel of a Sox2 purification displayed below (Fig. 21)

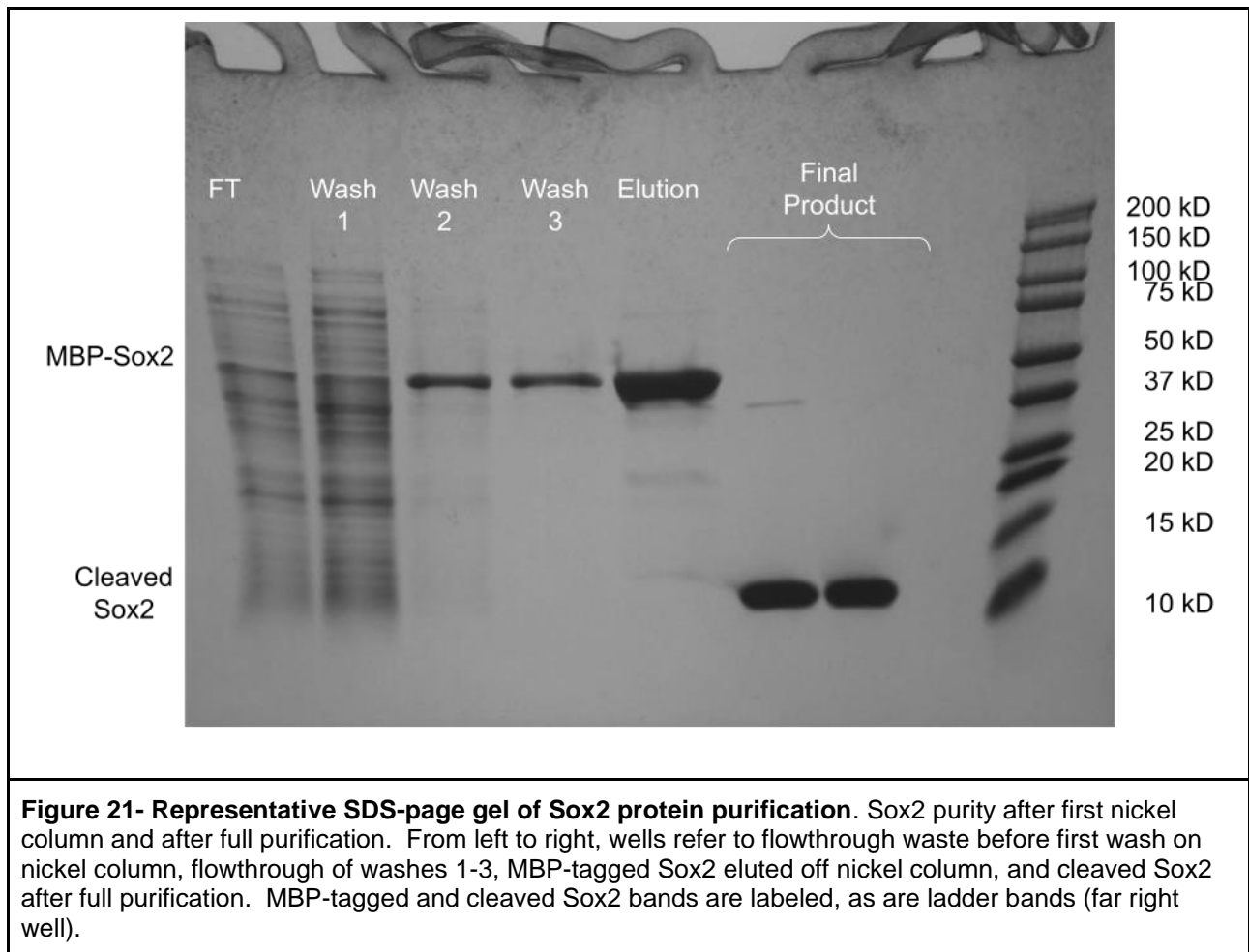


Figure 21- Representative SDS-page gel of Sox2 protein purification. Sox2 purity after first nickel column and after full purification. From left to right, wells refer to flowthrough waste before first wash on nickel column, flowthrough of washes 1-3, MBP-tagged Sox2 eluted off nickel column, and cleaved Sox2 after full purification. MBP-tagged and cleaved Sox2 bands are labeled, as are ladder bands (far right well).

5.2 RNA Purification

All RNA constructs used in this work are listed in appendix 9. DNA gBlocks™ encoding RNAs of interest were designed with the T7 promoter at the 5' end of the RNA

sequence and a 3' HDV ribozyme to prevent the incorporation of an n+1 nucleotide at the 3' end of the RNA upon T7 transcription. DNA templates were ordered Integrated DNA Technologies (IDT) (Coralville, IA, USA) with standard desalting. PCR amplification of template and subsequent in vitro T7 transcription were performed by standard protocols.¹¹⁵ For RNAs 27 nucleotides or smaller in length RNA solution were buffer exchanged into 10 mM H₂O₄PNa, 100 mM NaCl, pH 6.5, concentrated, and subsequently run through a size exclusion column (HiLoad 16/600 Superdex G-75 prep grade column (GE Life Science)) (10.1261/rna.342607). Fractions with desired RNA product were collected, buffer exchanged into 0.5x TE (5 mM Tris base, pH 8.0, 0.5 mM EDTA), and stored at -20 °C. Representative chromatograms of size-exclusion purified RNAs are shown below (Fig. 22). RNAs over 27 nucleotides in length were purified using standard denaturing polyacrylamide gel electrophoresis purification,¹¹⁵ as RNAs above this threshold were found to not be adequately separated from the ribozyme during size exclusion chromatography. RNA purity was verified on a 15% denaturing polyacrylamide gel (Fig. 23).

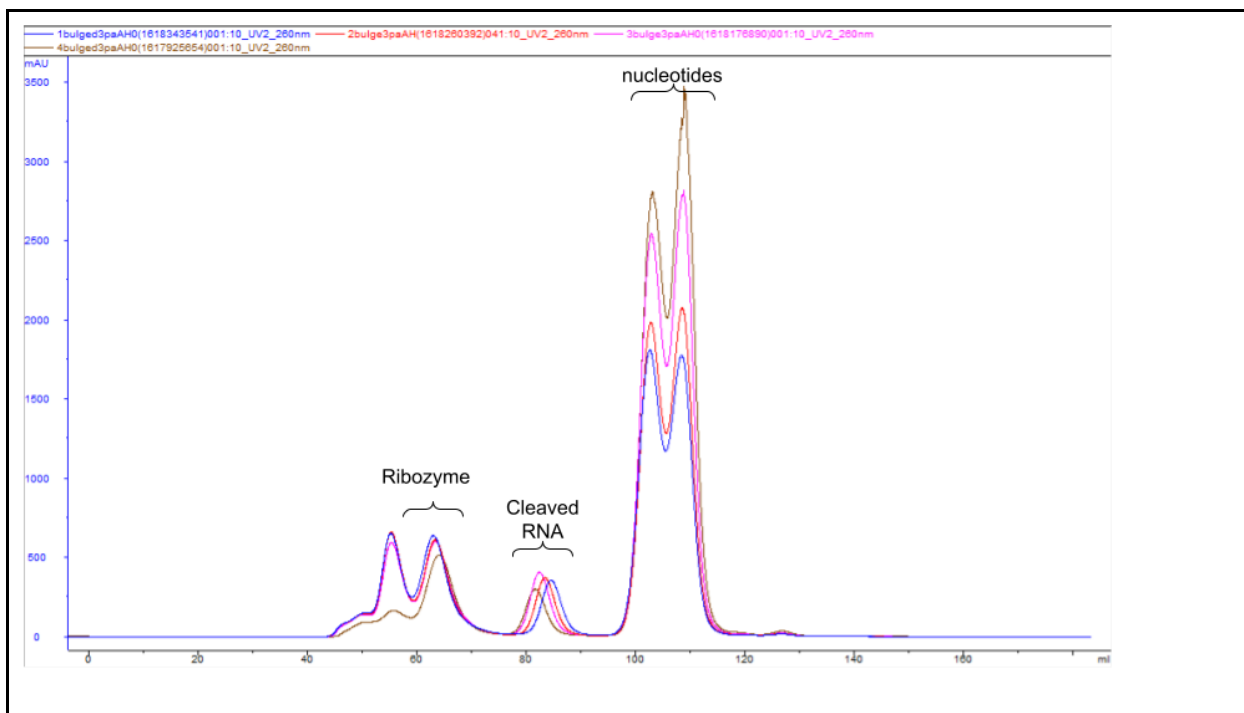


Figure 22- Representative chromatogram of SEC-purified RNAs. Chromatogram from 1.25 mL *in vitro* transcriptions of Internal bulge 0+1-1 (blue), Internal bulge 0+1 -2 (red), Internal bulge 0+1-3 (pink), and Internal bulge 0+1-4 (brown) RNAs. 260 nm absorption vs mL eluted shown. Known peaks are labeled.

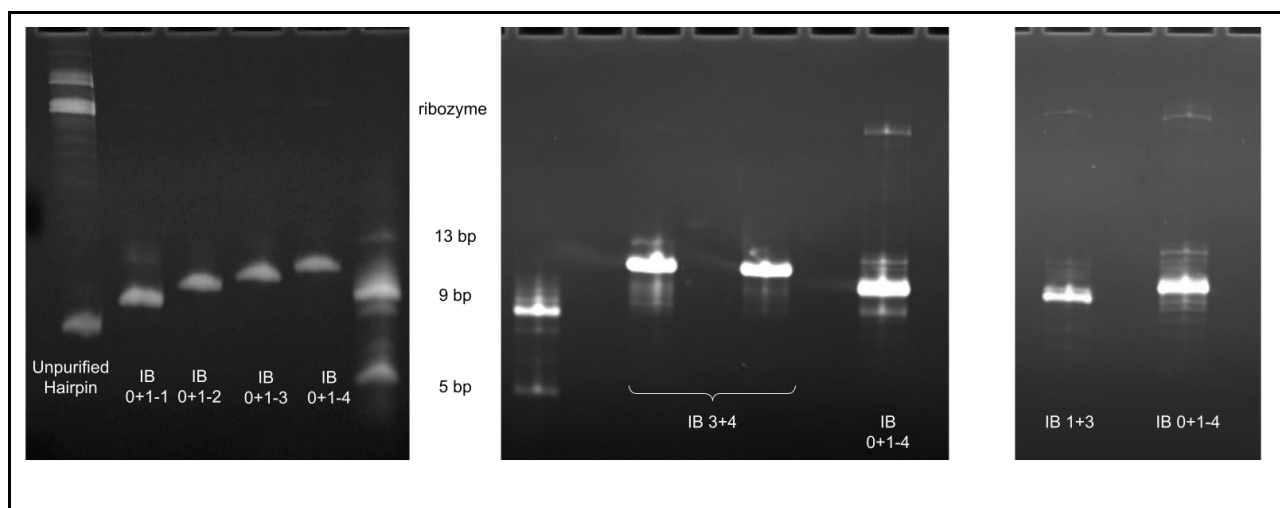


Figure 23- Representative denaturing gels of final purified RNA products. Ladder bands and cleaved ribozyme bands are labeled.

5.3 Nucleic Acid preparation

DNA ligands were ordered from IDT with standard desalting and resuspended in 0.5x TE for storage at -20 °C. dsDNA ligands were annealed by heating to 95 °C and cooling to 4 °C at a rate of -2 °C/min. Single-stranded DNA and RNA were folded by heating to 95 °C followed by rapid cooling in an ice bath for 10 minutes. All nucleic acids were then allowed to equilibrate for 30 minutes at 37 °C directly before use.

5.4 Fluorescence Anisotropy Binding Assays

FA binding assays were conducted as previously described in Holmes et al. 2020.⁶⁰ 5' 6-carboxyfluorescein (FAM)-labeled nucleic acids were prepared as previously described. Protein was incubated with 2 nM nucleic acid ligand in binding buffer (10 mM Tris-HEPES, pH 7.5, 8% Ficoll, 0.05% NP-40, 135 mM KCl, 15 mM NaCl, 1 mM DTT, 0.1 mg/mL non-acetylated BSA), centrifuged briefly with a microplate handyfuge, mixed by gentle shaking for 100 seconds, and allowed to equilibrate for 1 hour prior to FA measurements. Fluorescent anisotropic data was collected using a BMG Labtech CLARIOstar Plus microplate reader. Excitation wavelength was 482±8 nm and emission spectra were collected at 530±20 nm. Binding curves were either fit to the two-transition quadratic binding equation (1) or with the one-transition quadratic binding equation if only one transition occurred (2):

$$(1) A = A_0 + (A_1 - A_0) * \frac{(2+[L]_t+K_{D1})-\sqrt{(2+[L]_t+K_{D1})^2-(4*2*[L]_t)}}{2*2} + (A_2 - A_1) * \left(\frac{[L]_t}{[L]_t+K_{D2}}\right)$$

$$(2) A = A_0 + (A_1 - A_0) * \frac{(2+[L]_t+K_{D1})-\sqrt{(2+[L]_t+K_{D1})^2-(4*2*[L]_t)}}{2*2}$$

in which A = anisotropy, A_0 = lower anisotropy baseline, A_1 = upper anisotropy baseline of the first transition, $[L]_t$ = protein concentration, K_{D1} = apparent dissociation constant of the first transition, A_2 = upper anisotropy baseline of the second transition, and K_{D2} = apparent dissociation constant of the second transition. Fitting was completed in Kaleidagraph 4.1.1 for Macs, Synergy Software, Reading, PA, USA. www.synergy.com. All binding affinity measurements were performed in technical triplicate.

5.5 Electrophoretic Mobility-Shift Assay

To validate Sox2-G4 binding assays, Sox2 and pre-folded 5' FAM-labeled G4 DNA were mixed in EMSA binding buffer (10 mM Tris-HEPES, pH 7.5, 10% glycerol, 0.05% NP-40, 135 mM KCl, 15 mM NaCl, 1 mM DTT, 0.1 mg/mL non-acetylated BSA), left to equilibrate at room temperature for 30 minutes and at 4 °C for an additional 30 minutes. Samples were subsequently loaded on an 8% native polyacrylamide gel supplemented with 1x THE buffer (50 mM Tris base, 50 mM HEPES acid, pH 8.0, 1 mM EDTA) buffer and run at 5 W for 55 min at room temperature. Gels were imaged using a Typhoon PhosphorImager (Molecular Dynamics). For stoichiometric binding assays, the reactions were assembled in low-salt buffer (10 mM HEPES, pH 7, 100 mM KCl). Sox2-RNA stoichiometric binding assays were run on 10% native polyacrylamide gel for 55 minutes at 10 W, ethidium bromide stained, and imaged with an Alphamager (Alpha Innotech).

5.6 Crystal screening

To generate material for crystallography trials, hairpin-RNAs and Sox2 were exchanged into crystallography buffer (10 mM HEPES, pH 7, 50 mM KCl). Sox2 was

then added to 1.4-fold excess RNA for Internal Bulge 0+1 1-4 and Internal Bulge 1+3 constructs, or to 1.17-fold excess RNA for the Internal Bulge 3+4 construct (ratios were determined as being just below the Sox2-RNA titration point observed on stoichiometry gel (Appendix 1)). Final Sox2 concentration was 300 μ M. Crystal trials were performed by hanging-drop vapor diffusion with 3 μ L drops. Nucleic Acid Mini Screen (Hampton), Nuclix screen (NeXtal), PEG/ion screen (Hampton) screen, and PEGs II screen (NeXtal) were all used to survey crystallographic conditions.

5.7 Bioinformatic Analysis Pipeline

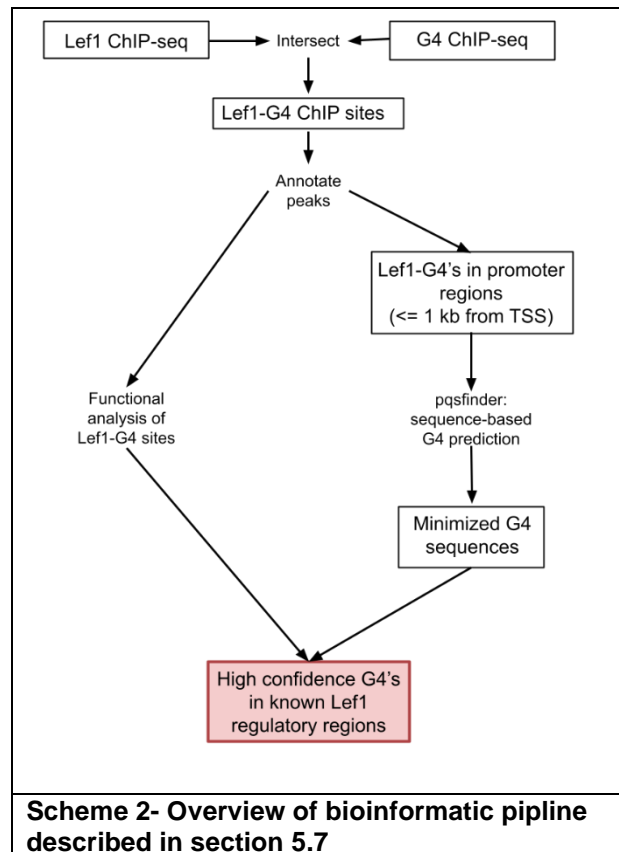
Bioinformatic analysis for this work was completed in R (3.6.0) and Bash (4.2.46). Other bioinformatic tools used for this analysis include BEDtools¹¹⁶ (v2.30.0) and LiftOver (UCSC). K562 and HEK293T LEF1 ChIP-seq datasets were obtained from (ENCFF659WAF) and (ENCFF333UCS), respectively. K562 and HEK293 G4 ChIP-seq datasets were obtained from (GSE107690) and (GSE133379), respectively. LEF1-G4 data was mapped to hg19. LEF1-G4 overlap was calculated using BEDtools Intersect, and significance determined by BEDtools Fisher. BED files were converted to FASTA files by BEDtools Getfasta for consensus sequence enrichment analysis and G4 sequence prediction. G4 sequence prediction was completed using pqsFinder (2.2.0) with a minimum score of 52 on both sense and antisense strands.⁸⁸ ChIP data was annotated using the ChIPseeker package¹¹⁷ (1.22.1)

and UCSC Known Genes¹¹⁸ (3.2.2).

Pathway analysis was completed with annotated dataset using ReactomePA¹¹⁹ (1.30.0). K562 and HEK293 enhancer datasets were obtained from EnhancerAtlas 2.0¹⁰⁴. General pipeline workflow is described in scheme 2.

Sox2-G4 analysis was performed as described above with minor modifications. H9 Sox2 ChIP data was obtained from Cistrome ID: 44233 and H9 G4 ChIP data was obtained from GSE161531. Sox2-G4

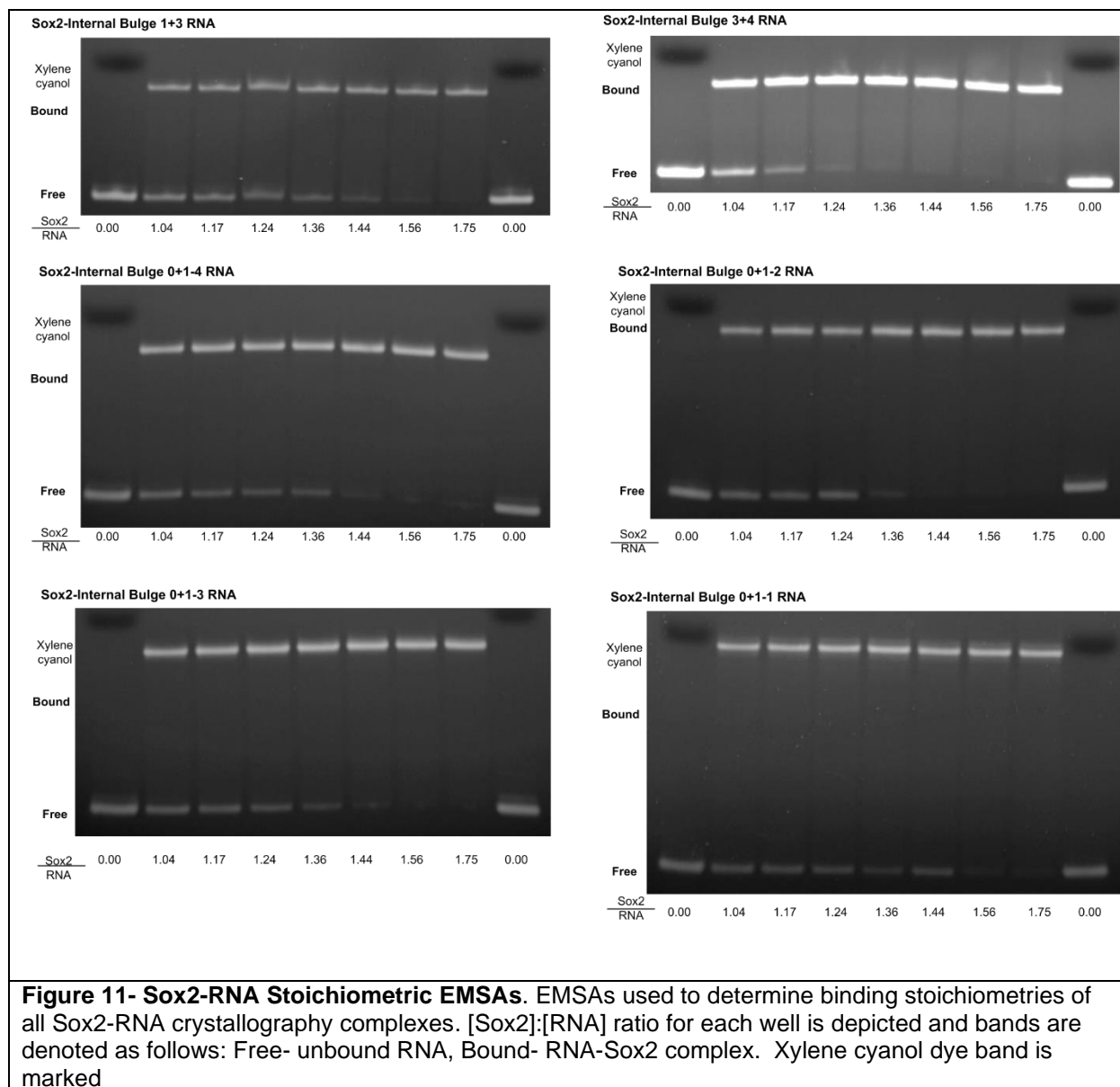
data was mapped to hg38. H9 enhancers dataset obtained from EnhancerAtlas 2.0¹⁰⁴ was converted to hg38 with LiftOver



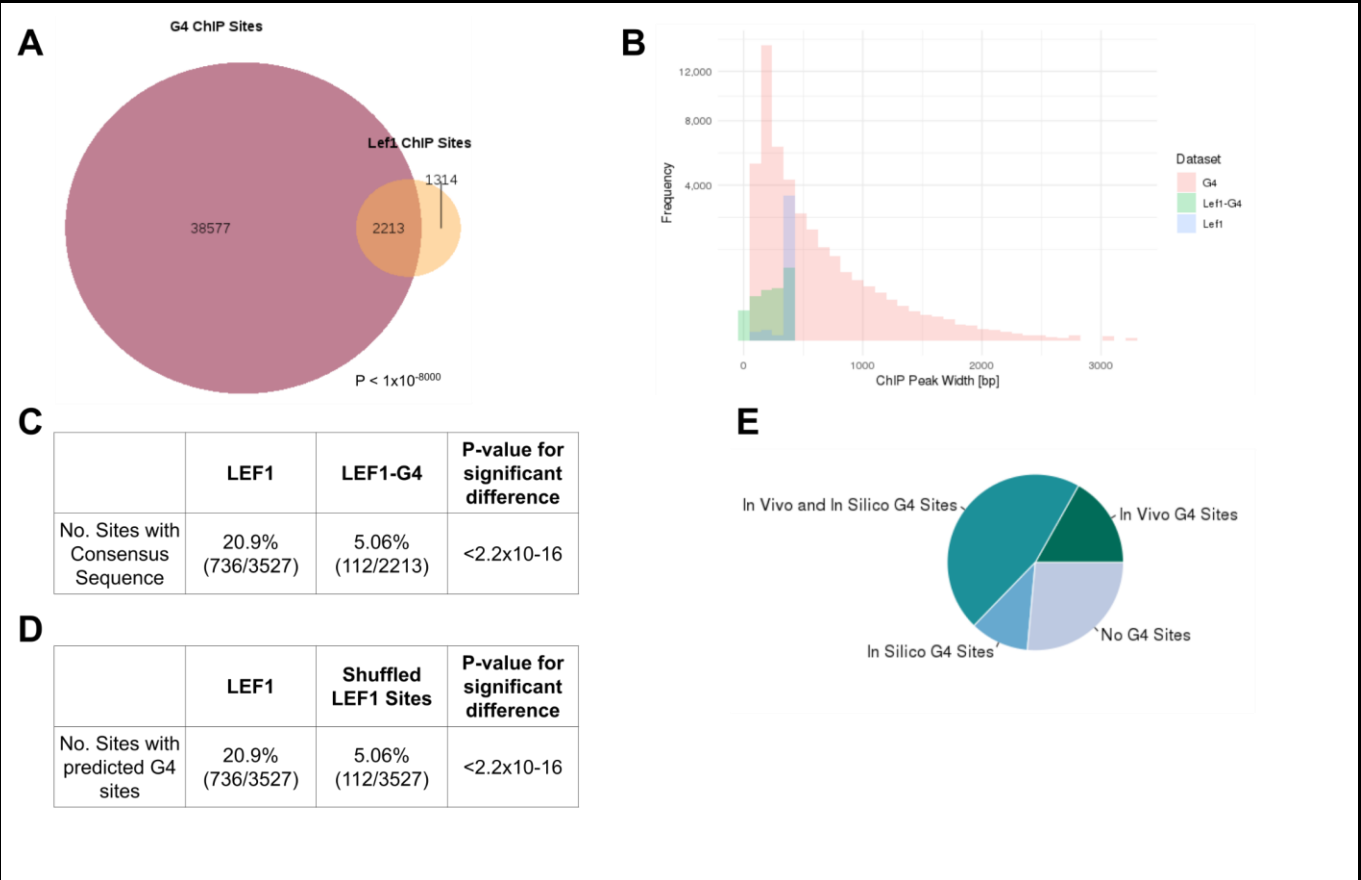
5.8 Thioflavin-T Fluorescence Assay

DNA ligands were prepared as described in section 5.3 and added to 3 μ M THT in THT binding buffer (50 mM Tris base, 50 mM, pH 7.2) mixed by gentle shaking for 90 seconds, centrifuged briefly with a microplate Handyluge, and allowed to equilibrate for 1 hour prior to measurement. Fluorescence data was collected using a BMG Labtech CLARIOstar Plus microplate reader. Excitation wavelength range was from 420 \pm 5 nm to 440 \pm 5 nm and emission spectra were collected at 490 \pm 8 nm. All THT assays were performed in technical triplicate.

Appendix 1: Sox2-RNA Hairpin Stoichiometric Binding Gels

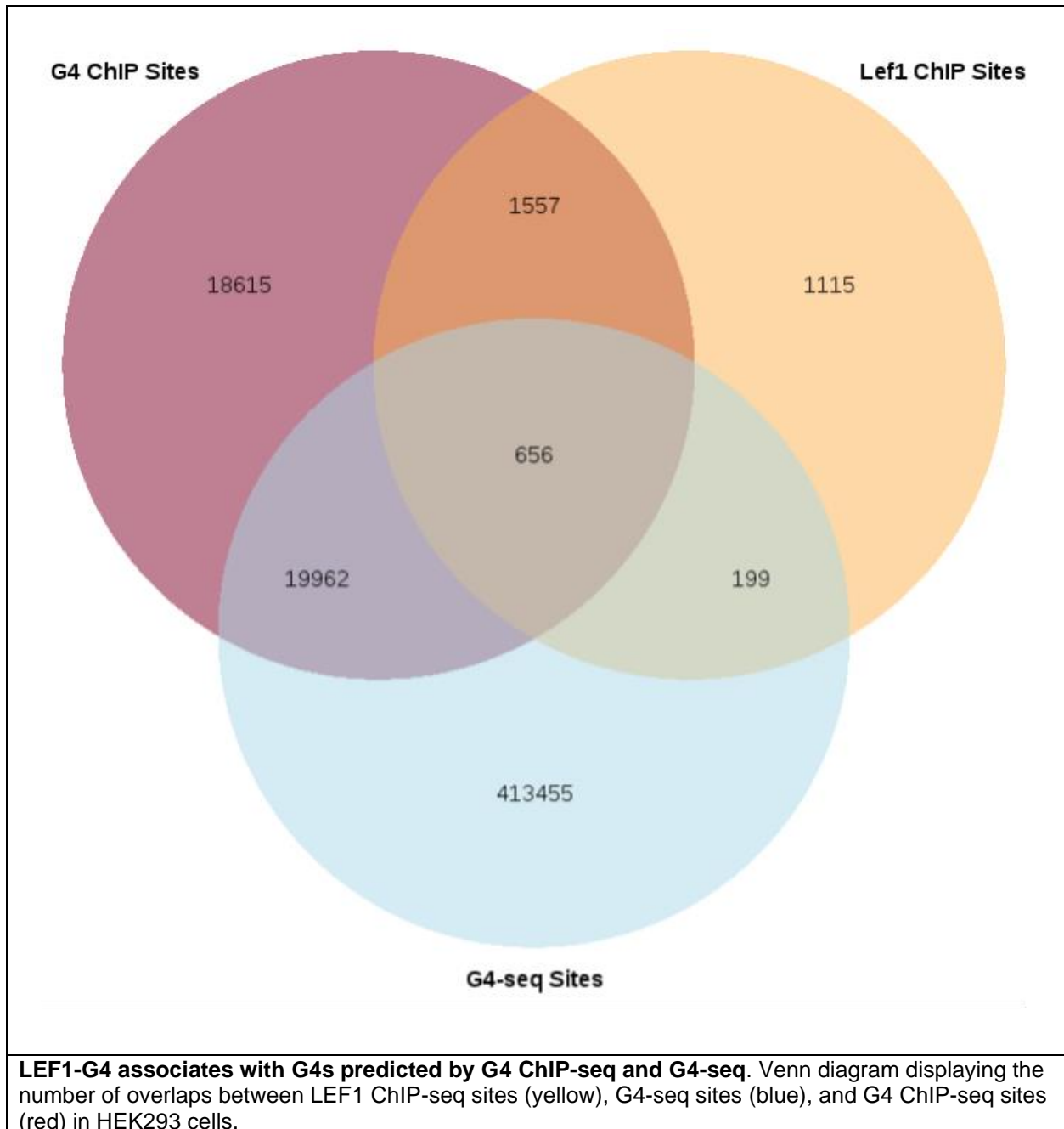


Appendix 2: LEF1-G4 genomic associations in HEK293 cells

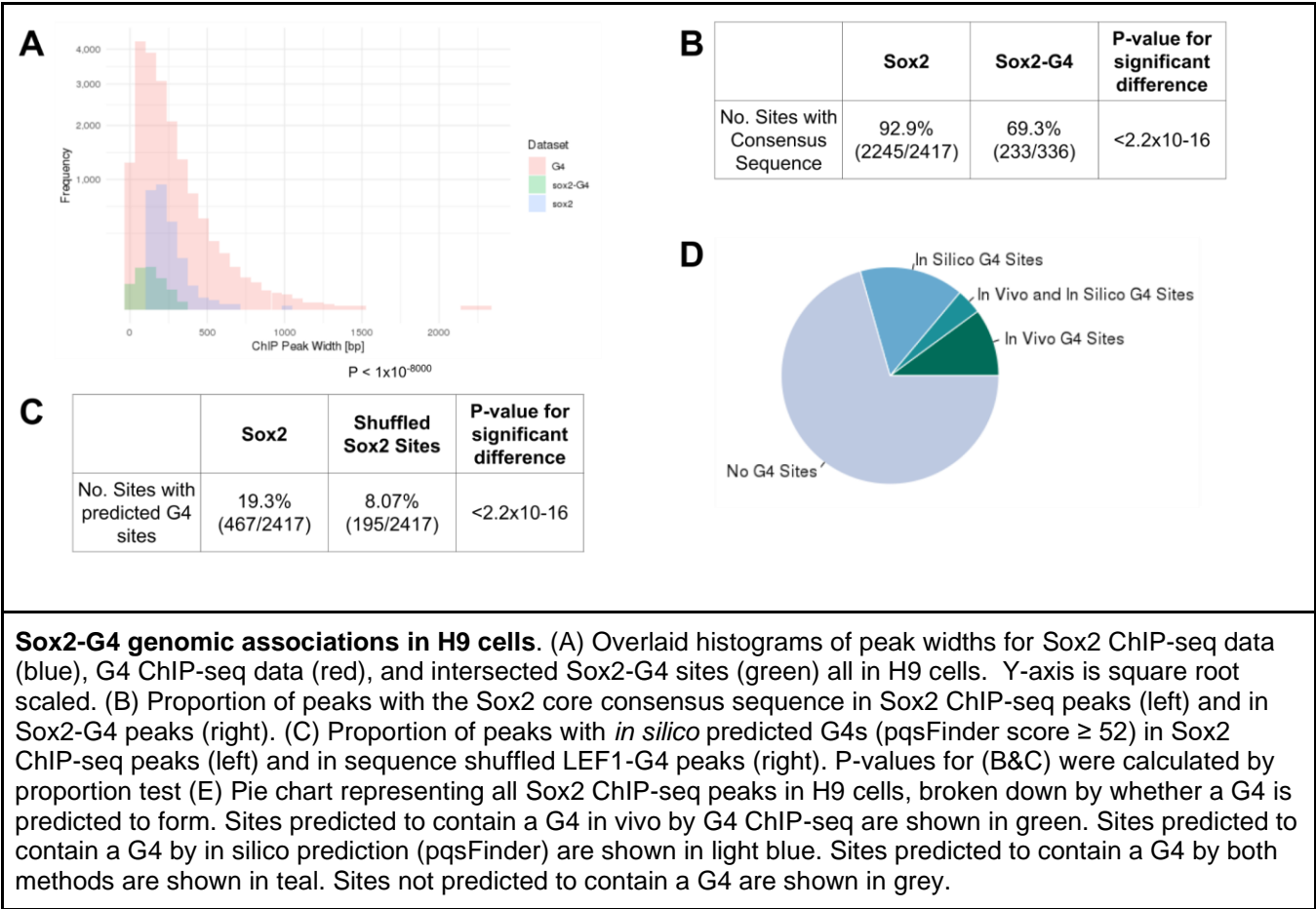


LEF1-G4 genomic associations in HEK293 cells. (A) Venn diagram displaying the number of overlaps between LEF1 ChIP-seq sites (yellow) and G4 ChIP-seq sites (red) in HEK293 cells. P-value displayed was calculated by Fisher's exact test. (B) Overlaid histograms of peak widths for LEF1 ChIP-seq data (blue), G4 ChIP-seq data (red), and intersected LEF1-G4 sites (green) all in HEK293 cells. Y-axis is square root scaled. (C) Proportion of peaks with the LEF1 core consensus sequence in LEF1 ChIP-seq peaks (left) and in LEF1-G4 peaks (right). (D) Proportion of peaks with *in silico* predicted G4s (pqsFinder score ≥ 52) in LEF1 ChIP-seq peaks (left) and in sequence shuffled LEF1-G4 peaks (right). P-values for (C&D) were calculated by proportion test (E) Pie chart representing all LEF1 ChIP-seq peaks in HEK293 cells, broken down by whether a G4 is predicted to form. Sites predicted to contain a G4 *in vivo* by G4 ChIP-seq are shown in green. Sites predicted to contain a G4 by *in silico* prediction (pqsFinder) are shown in light blue. Sites predicted to contain a G4 by both methods are shown in teal. Sites not predicted to contain a G4 are shown in grey.

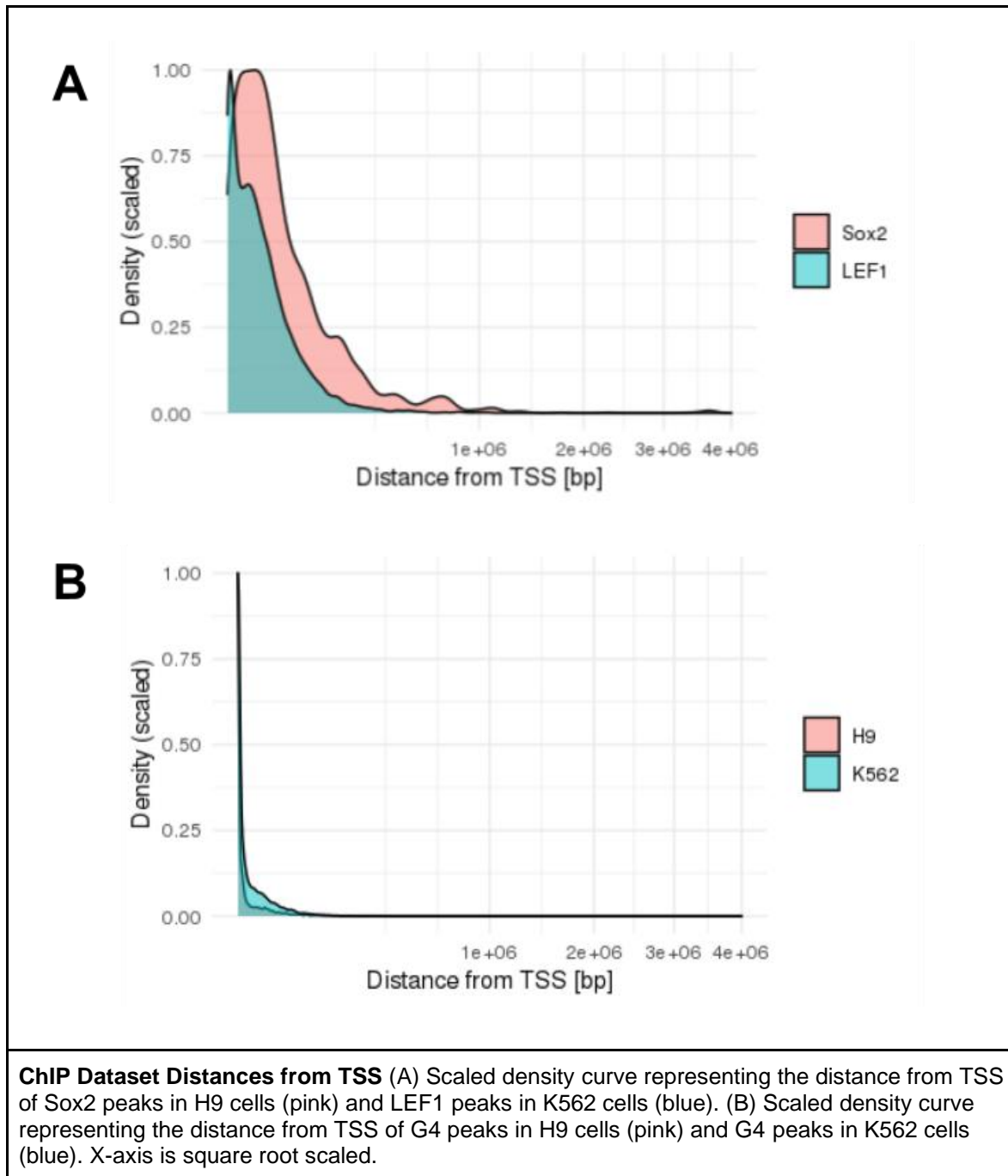
Appendix 3: LEF1-G4 associates with G4s predicted by G4 ChIP-seq and G4-seq



Appendix 4: Sox2-G4 associations in H9 cells



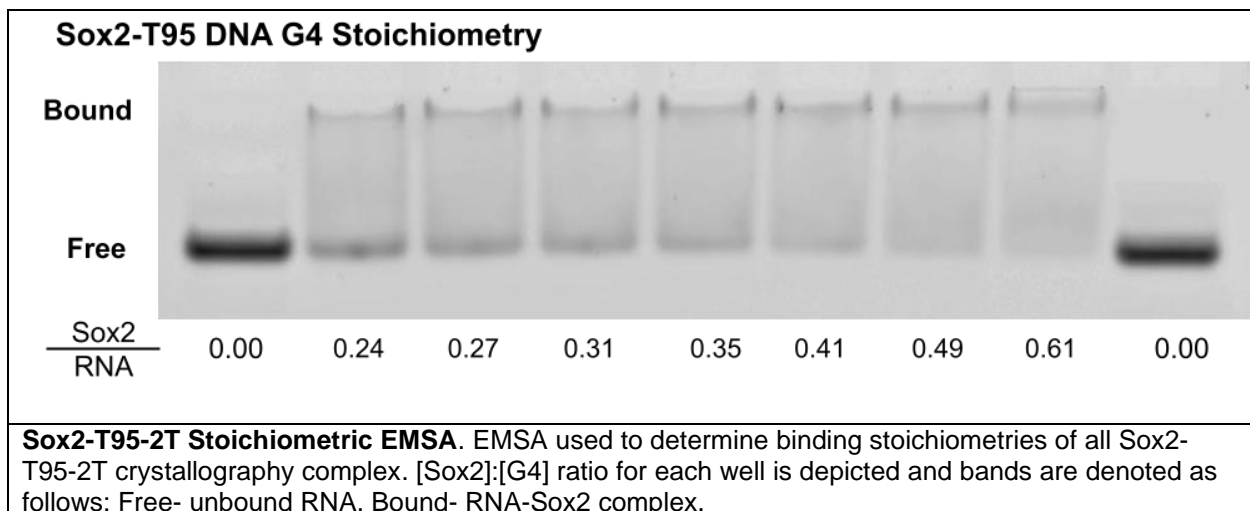
Appendix 5: ChIP Dataset Distances from TSS



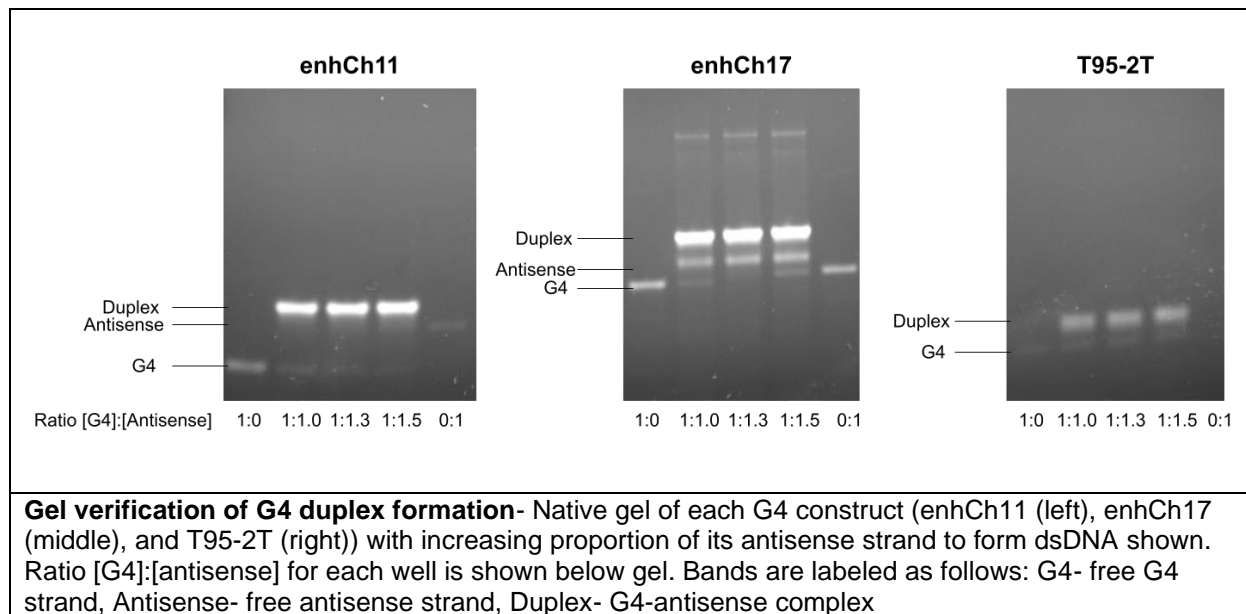
Appendix 6: *In Vitro* Survey Genomic G4s

Name	PQS Score	Sequence	Region	Core Consensus seq in ChIP site?
PromKDM2B	85	CGGGGCGGCCGGGGCCGGGCCCAGCGG GGC	>chr12:121904776-121904948	No
EnhCh17	71	TGGGGCTGGGGAGCCTCGGGAGAGGAAGG GTGGTGTGTTGGGA	>chr17:74964759-74964960	No
EnhCh11	70	AGGGAGCACAGGGTTGGGGGTGGGGT	>chr11:118910750-118910796	No

Appendix 7: Sox2-G4 Stoichiometric Binding Gel



Appendix 8: In Gel Verification of G4 Duplex Formation



Appendix 8: Nucleic Acid Constructs

Construct	Sequence
gBlock™ Template	GCGCGCGAATTCT TAATACGACTCACTATA *DNA template for RNA of interest*GCCGGCCATGGTCCCAGCCTCCTCGCTGGCGGCCGGTGGGCAACA TGCTTCGGCATGGCGAATGGGACCTCTAG <u>ACTGTGCATCGGGTCAGGA</u>
*Internal Bulge 0+1-1	GGAUUGUGCGAAAGCAAAUCCA
*Internal Bulge 0+1-2	GGAUUGUGACGAAAGUCAAAUCCA
*Internal Bulge 0+1-3	GGAUUGUGAGCGAAAGCUCAAUCCA
*Internal Bulge 0+1-4	GGAUUGUGAGCCGAAAGGCUCAAUCCA
*Internal Bulge 1+3	GGACUUGUGCGAAAGCACCCAGUCCA
*Internal Bulge 3+4	GGUACAUUCUACGAAAGUAGCCCCGUACCA
T95-2T	TTGGGTGGGTGGGTGGGT
Antisense T95-2T	ACCCACCCACCCACCCAA
Tel22	AGGGTTAGGGTTAGGGTTAGGGT
Antisense Tel22	ACCCTAACCCTAACCCTAACCCT
enhCh11	AGGGAGCACAGGGTTGGGGGTGGGGT
Antisense enhCh11	ACCCACCCCCAACCCCTGTGCTCCCT
enhCh17	TGGGGCTGGGAGCCTCGGGAGAGGAAGGGTGGTGTGGGGA
Antisense enhCh17	TCCCCAAACACCACCTTCTCTCCCGAGGCTCCCCAGCCCCA
promKDMB2	CGGGGCGGCCGGGGCCGGGCCCGAGCGGGGC
Antisense promKDMB2	GCCCCGCTCGGGCCCGGCCCGGCCGCCCGG
Sox2 consensus dsDNA	CGCGCCTTTGTTCCCGGGT
Sox2 non-consensus dsDNA	CGCGCGGCGCGGCCCGGGC
Hairpin RNA, Fully Paired (IL2 FREP)	GGUCUUAUCAUGCGGGCGAAAGUCUGUAUGAUGGGACC
Hairpin RNA, 0+1 (IL2 0+1)	GGUCUUAAGGUGCGGGCGAAAGUCUGUACCUCUGGGACC
Hairpin RNA, 0+1 (IL2 1+2)	GGUCUUAUGGUGCGGGCGAAAGUCUGUACCUCUGGGACC
Hairpin RNA, 0+1 (IL2 2+3)	GGUCUUAUCGUGCGGGCGAAAGUCUGUACCUCUGGGACC
Hairpin RNA, 0+1 (IL2 3+4)	GGUCUUAUCAUGCGGGCGAAAGUCUGUACCUCUGGGACC

Bold: T7 promoter

Underline: 3'-primer binding site

*Crystallography construct derived from gBlock™ template

Appendix 10: Protein Constructs

HMGB Domain	MBP-tagged Protein Sequence	Cleaved Protein Sequence
Sox2: Minimized HMGB Uniprot ID: P48431	MHHHHHHHHKIEEGKLVWINGDKGYNGLAEVGGKFEKDTGIKVTVEHPDKLEEKFPQVAAT GDGPDIIFWAHDRFGGYAQSGLLAEITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSLI YNKDLLPNPPKTWEEIPALDKELKAKGKSALMFNLQEPYFTWPLIAADGGYAFKYENGKYDI KDVGVNDAGAKAGLTFVLDLIKNNHMNADTDYSIAEAFNKGETAMTINGPWAWSNIDTSKV NYGVTVLPTFGQPSKPFVGVLSAGINAASPNKELAKEFLENYLLTDEGLEAVNKDKPLGAV ALKSYEEELAKDPRIATMENAQKGEIMPNIQMSAFWYAVRTAVINAASGRQTVDEALKDA QTNSSSVPGRGSIEGRA LEVLFQGP NSPDRVKRPMNAFMVWSRGQRRKMAQENPKMHN SEISKRLGAEWKLLSETEKRPFIDEAKRLRALHMKHEHPDYKYPRRKTKTLMK	GPNSPDRVKRPMNAFMVWSRGQRRKMAQENPKM HNSEISKRLGAEWKLLSETEKRPFIDEAKRLRALHMK EHPDYKYPRRKTKTLMK
Sox2: Extended HMGB Uniprot ID: P48431	MHHHHHHHHKIEEGKLVWINGDKGYNGLAEVGGKFEKDTGIKVTVEHPDKLEEKFPQVAAT GDGPDIIFWAHDRFGGYAQSGLLAEITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSLI YNKDLLPNPPKTWEEIPALDKELKAKGKSALMFNLQEPYFTWPLIAADGGYAFKYENGKYDI KDVGVNDAGAKAGLTFVLDLIKNNHMNADTDYSIAEAFNKGETAMTINGPWAWSNIDTSKV NYGVTVLPTFGQPSKPFVGVLSAGINAASPNKELAKEFLENYLLTDEGLEAVNKDKPLGAV ALKSYEEELAKDPRIATMENAQKGEIMPNIQMSAFWYAVRTAVINAASGRQTVDEALKDA QTNSSSVPGRGSIEGRA LEVLFQGP DRVKRPMNAFMVWSRGQRRKMAQENPKMHNSEIS KRLGAEWKLLSETEKRPFIDEAKRLRALHMKHEHPDYKYPRRKTKT	GPDRVKRPMNAFMVWSRGQRRKMAQENPKMHN EISKRLGAEWKLLSETEKRPFIDEAKRLRALHMKHEP DYKYPRRKTKT

* Cleavage sequence shown in bold

Data and Code Availability

Code written to conduct bioinformatic analysis described in this work and the data generated therein can be found at

https://github.com/abhe6819/HMGB_G4_Interactions.git

References

- (1) Latchman, D. S. Transcription Factors: An Overview. *Int. J. Biochem. Cell Biol.* **1997**, 29 (12), 1305–1312. [https://doi.org/10.1016/S1357-2725\(97\)00085-X](https://doi.org/10.1016/S1357-2725(97)00085-X).
- (2) Lambert, S. A.; Jolma, A.; Campitelli, L. F.; Das, P. K.; Yin, Y.; Albu, M.; Chen, X.; Taipale, J.; Hughes, T. R.; Weirauch, M. T. The Human Transcription Factors. *Cell* **2018**, 172 (4), 650–665. <https://doi.org/10.1016/j.cell.2018.01.029>.
- (3) Slattery, M.; Zhou, T.; Yang, L.; Dantas Machado, A. C.; Gordân, R.; Rohs, R. Absence of a Simple Code: How Transcription Factors Read the Genome. *Trends Biochem. Sci.* **2014**, 39 (9), 381–399. <https://doi.org/10.1016/j.tibs.2014.07.002>.
- (4) Schnepf, M.; von Reutern, M.; Ludwig, C.; Jung, C.; Gaul, U. Transcription Factor Binding Affinities and DNA Shape Readout. *iScience* **2020**, 23 (11), 101694. <https://doi.org/10.1016/j.isci.2020.101694>.
- (5) Yang, A.; Zhu, Z.; Kapranov, P.; McKeon, F.; Church, G. M.; Gingeras, T. R.; Struhl, K. Relationships between P63 Binding, DNA Sequence, Transcription

- Activity, and Biological Function in Human Cells. *Mol. Cell* **2006**, *24* (4), 593–602. <https://doi.org/10.1016/j.molcel.2006.10.018>.
- (6) Joseph, R.; Orlov, Y. L.; Huss, M.; Sun, W.; Li Kong, S.; Ukil, L.; Fu Pan, Y.; Li, G.; Lim, M.; Thomsen, J. S.; Ruan, Y.; Clarke, N. D.; Prabhakar, S.; Cheung, E.; Liu, E. T. Integrative Model of Genomic Factors for Determining Binding Site Selection by Estrogen Receptor- α . *Mol. Syst. Biol.* **2010**, *6* (1), 456. <https://doi.org/10.1038/msb.2010.109>.
 - (7) Chen, Y.; Bates, D. L.; Dey, R.; Chen, P.-H.; Machado, A. C. D.; Laird-Offringa, I. A.; Rohs, R.; Chen, L. DNA Binding by GATA Transcription Factor Suggests Mechanisms of DNA Looping and Long-Range Gene Regulation. *Cell Rep.* **2012**, *2* (5), 1197–1206. <https://doi.org/10.1016/j.celrep.2012.10.012>.
 - (8) Kitayner, M.; Rozenberg, H.; Rohs, R.; Suad, O.; Rabinovich, D.; Honig, B.; Shakked, Z. Diversity in DNA Recognition by P53 Revealed by Crystal Structures with Hoogsteen Base Pairs. *Nat. Struct. Mol. Biol.* **2010**, *17* (4), 423–429. <https://doi.org/10.1038/nsmb.1800>.
 - (9) Chen, J.; Zhang, Z.; Li, L.; Chen, B.-C.; Revyakin, A.; Hajj, B.; Legant, W.; Dahan, M.; Lionnet, T.; Betzig, E.; Tjian, R.; Liu, Z. Single-Molecule Dynamics of Enhanceosome Assembly in Embryonic Stem Cells. *Cell* **2014**, *156* (6), 1274–1285. <https://doi.org/10.1016/j.cell.2014.01.062>.
 - (10) Soufi, A.; Garcia, M. F.; Jaroszewicz, A.; Osman, N.; Pellegrini, M.; Zaret, K. S. Pioneer Transcription Factors Target Partial DNA Motifs on Nucleosomes to Initiate Reprogramming. *Cell* **2015**, *161* (3), 555–568. <https://doi.org/10.1016/j.cell.2015.03.017>.
 - (11) Luger, K. Dynamic Nucleosomes. *Chromosome Res.* **2006**, *14* (1), 5–16. <https://doi.org/10.1007/s10577-005-1026-1>.
 - (12) Zaret, K. S.; Carroll, J. S. Pioneer Transcription Factors: Establishing Competence for Gene Expression. *Genes Dev.* **2011**, *25* (21), 2227–2241. <https://doi.org/10.1101/gad.176826.111>.
 - (13) Dodonova, S. O.; Zhu, F.; Dienemann, C.; Taipale, J.; Cramer, P. Nucleosome-Bound SOX2 and SOX11 Structures Elucidate Pioneer Factor Function. *Nature* **2020**, *580* (7805), 669–672. <https://doi.org/10.1038/s41586-020-2195-y>.
 - (14) Echigoya, K.; Koyama, M.; Negishi, L.; Takizawa, Y.; Mizukami, Y.; Shimabayashi, H.; Kuroda, A.; Kurumizaka, H. Nucleosome Binding by the Pioneer Transcription Factor OCT4. *Sci. Rep.* **2020**, *10* (1), 11832. <https://doi.org/10.1038/s41598-020-68850-1>.
 - (15) Takahashi, K.; Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **2006**, *126* (4), 663–676. <https://doi.org/10.1016/j.cell.2006.07.024>.
 - (16) Cassidy, L. A. Having It Both Ways: Transcription Factors That Bind DNA and RNA. *Nucleic Acids Res.* **2002**, *30* (19), 4118–4126. <https://doi.org/10.1093/nar/gkf512>.
 - (17) Hudson, W. H.; Ortlund, E. A. The Structure, Function and Evolution of Proteins That Bind DNA and RNA. *Nat. Rev. Mol. Cell Biol.* **2014**, *15* (11), 749–760. <https://doi.org/10.1038/nrm3884>.
 - (18) Riley, K. J.-L.; Ramirez-Alvarado, M.; Maher, L. J. RNA-p53 Interactions in Vitro. *Biochemistry* **2007**, *46* (9), 2480–2487. <https://doi.org/10.1021/bi061480v>.

- (19) Clemens, K. R.; Wolf, V.; McBryant, S. J.; Zhang, P.; Liao, X.; Wright, P. E.; Gottesfeld, J. M. Molecular Basis for Specific Recognition of Both RNA and DNA by a Zinc Finger Protein. *Science* **1993**, *260* (5107), 530–533. <https://doi.org/10.1126/science.8475383>.
- (20) Kino, T.; Hurt, D. E.; Ichijo, T.; Nader, N.; Chrousos, G. P. Noncoding RNA Gas5 Is a Growth Arrest– and Starvation-Associated Repressor of the Glucocorticoid Receptor. *Sci. Signal.* **2010**, *3* (107). <https://doi.org/10.1126/scisignal.2000568>.
- (21) Parsonnet, N. V.; Lammer, N. C.; Holmes, Z. E.; Batey, R. T.; Wuttke, D. S. The Glucocorticoid Receptor DNA-Binding Domain Recognizes RNA Hairpin Structures with High Affinity. *Nucleic Acids Res.* **2019**, *47* (15), 8180–8192. <https://doi.org/10.1093/nar/gkz486>.
- (22) Huang, D.-B.; Vu, D.; Cassiday, L. A.; Zimmerman, J. M.; Maher, L. J.; Ghosh, G. Crystal Structure of NF-KB (P50)₂ Complexed to a High-Affinity RNA Aptamer. *Proc. Natl. Acad. Sci.* **2003**, *100* (16), 9268–9273. <https://doi.org/10.1073/pnas.1632011100>.
- (23) Fukunaga, J.; Nomura, Y.; Tanaka, Y.; Amano, R.; Tanaka, T.; Nakamura, Y.; Kawai, G.; Sakamoto, T.; Kozu, T. The Runt Domain of AML1 (RUNX1) Binds a Sequence-Conserved RNA Motif That Mimics a DNA Element. *RNA* **2013**, *19* (7), 927–936. <https://doi.org/10.1261/rna.037879.112>.
- (24) Ohe, K.; Lalli, E.; Sassone-Corsi, P. A Direct Role of SRY and SOX Proteins in Pre-mRNA Splicing. *Proc. Natl. Acad. Sci.* **2002**, *99* (3), 1146–1151. <https://doi.org/10.1073/pnas.022645899>.
- (25) Sigova, A. A.; Abraham, B. J.; Ji, X.; Molinie, B.; Hannett, N. M.; Guo, Y. E.; Jangi, M.; Giallourakis, C. C.; Sharp, P. A.; Young, R. A. Transcription Factor Trapping by RNA in Gene Regulatory Elements. *Science* **2015**, *350* (6263), 978–981. <https://doi.org/10.1126/science.aad3346>.
- (26) Williamson, J. R.; Raghuraman, M. K.; Cech, T. R. Monovalent Cation-Induced Structure of Telomeric DNA: The G-Quartet Model. *Cell* **1989**, *59* (5), 871–880. [https://doi.org/10.1016/0092-8674\(89\)90610-7](https://doi.org/10.1016/0092-8674(89)90610-7).
- (27) Huppert, J. L. Prevalence of Quadruplexes in the Human Genome. *Nucleic Acids Res.* **2005**, *33* (9), 2908–2916. <https://doi.org/10.1093/nar/gki609>.
- (28) Chariker, J. H.; Miller, D. M.; Rouchka, E. C. Computational Analysis of G-Quadruplex Forming Sequences across Chromosomes Reveals High Density Patterns Near the Terminal Ends. *PLOS ONE* **2016**, *11* (10), e0165101. <https://doi.org/10.1371/journal.pone.0165101>.
- (29) Bhattacharyya, D.; Mirihana Arachchilage, G.; Basu, S. Metal Cations in G-Quadruplex Folding and Stability. *Front. Chem.* **2016**, *4*. <https://doi.org/10.3389/fchem.2016.00038>.
- (30) Lee, M. P. H.; Parkinson, G. N.; Hazel, P.; Neidle, S. Observation of the Coexistence of Sodium and Calcium Ions in a DNA G-Quadruplex Ion Channel. *J. Am. Chem. Soc.* **2007**, *129* (33), 10106–10107. <https://doi.org/10.1021/ja0740869>.
- (31) Kouzine, F.; Liu, J.; Sanford, S.; Chung, H.-J.; Levens, D. The Dynamic Response of Upstream DNA to Transcription-Generated Torsional Stress. *Nat. Struct. Mol. Biol.* **2004**, *11* (11), 1092–1100. <https://doi.org/10.1038/nsmb848>.

- (32) Hänsel-Hertsch, R.; Beraldi, D.; Lensing, S. V.; Marsico, G.; Zyner, K.; Parry, A.; Di Antonio, M.; Pike, J.; Kimura, H.; Narita, M.; Tannahill, D.; Balasubramanian, S. G-Quadruplex Structures Mark Human Regulatory Chromatin. *Nat. Genet.* **2016**, *48* (10), 1267–1272. <https://doi.org/10.1038/ng.3662>.
- (33) Lago, S.; Nadai, M.; Cernilogar, F. M.; Kazerani, M.; Domínguez Moreno, H.; Schotta, G.; Richter, S. N. Promoter G-Quadruplexes and Transcription Factors Cooperate to Shape the Cell Type-Specific Transcriptome. *Nat. Commun.* **2021**, *12* (1), 3885. <https://doi.org/10.1038/s41467-021-24198-2>.
- (34) Balasubramanian, S.; Hurley, L. H.; Neidle, S. Targeting G-Quadruplexes in Gene Promoters: A Novel Anticancer Strategy? *Nat. Rev. Drug Discov.* **2011**, *10* (4), 261–275. <https://doi.org/10.1038/nrd3428>.
- (35) Mohaghegh, P. The Bloom's and Werner's Syndrome Proteins Are DNA Structure-Specific Helicases. *Nucleic Acids Res.* **2001**, *29* (13), 2843–2849. <https://doi.org/10.1093/nar/29.13.2843>.
- (36) Fry, M.; Loeb, L. A. Human Werner Syndrome DNA Helicase Unwinds Tetrahelical Structures of the Fragile X Syndrome Repeat Sequence d(CGG). *J. Biol. Chem.* **1999**, *274* (18), 12797–12802. <https://doi.org/10.1074/jbc.274.18.12797>.
- (37) Varshney, D.; Spiegel, J.; Zyner, K.; Tannahill, D.; Balasubramanian, S. The Regulation and Functions of DNA and RNA G-Quadruplexes. *Nat. Rev. Mol. Cell Biol.* **2020**, *21* (8), 459–474. <https://doi.org/10.1038/s41580-020-0236-x>.
- (38) Li, L.; Williams, P.; Ren, W.; Wang, M. Y.; Gao, Z.; Miao, W.; Huang, M.; Song, J.; Wang, Y. YY1 Interacts with Guanine Quadruplexes to Regulate DNA Looping and Gene Expression. *Nat. Chem. Biol.* **2021**, *17* (2), 161–168. <https://doi.org/10.1038/s41589-020-00695-1>.
- (39) Cogoi, S.; Zorzet, S.; Rapozzi, V.; Géci, I.; Pedersen, E. B.; Xodo, L. E. MAZ-Binding G4-Decoy with Locked Nucleic Acid and Twisted Intercalating Nucleic Acid Modifications Suppresses KRAS in Pancreatic Cancer Cells and Delays Tumor Growth in Mice. *Nucleic Acids Res.* **2013**, *41* (7), 4049–4064. <https://doi.org/10.1093/nar/gkt127>.
- (40) Raiber, E.-A.; Kranaster, R.; Lam, E.; Nikan, M.; Balasubramanian, S. A Non-Canonical DNA Structure Is a Binding Motif for the Transcription Factor SP1 in Vitro. *Nucleic Acids Res.* **2012**, *40* (4), 1499–1508. <https://doi.org/10.1093/nar/gkr882>.
- (41) Kumar, P.; Yadav, V. K.; Baral, A.; Kumar, P.; Saha, D.; Chowdhury, S. Zinc-Finger Transcription Factors Are Associated with Guanine Quadruplex Motifs in Human, Chimpanzee, Mouse and Rat Promoters Genome-Wide. *Nucleic Acids Res.* **2011**, *39* (18), 8005–8016. <https://doi.org/10.1093/nar/gkr536>.
- (42) Zhang, X.; Spiegel, J.; Martínez Cuesta, S.; Adhikari, S.; Balasubramanian, S. Chemical Profiling of DNA G-Quadruplex-Interacting Proteins in Live Cells. *Nat. Chem.* **2021**, *13* (7), 626–633. <https://doi.org/10.1038/s41557-021-00736-9>.
- (43) Spiegel, J.; Cuesta, S. M.; Adhikari, S.; Hänsel-Hertsch, R.; Tannahill, D.; Balasubramanian, S. G-Quadruplexes Are Transcription Factor Binding Hubs in Human Chromatin. *Genome Biol.* **2021**, *22* (1), 117. <https://doi.org/10.1186/s13059-021-02324-z>.

- (44) Hou, Y.; Li, F.; Zhang, R.; Li, S.; Liu, H.; Qin, Z. S.; Sun, X. Integrative Characterization of G-Quadruplexes in the Three-Dimensional Chromatin Structure. *Epigenetics* **2019**, *14* (9), 894–911. <https://doi.org/10.1080/15592294.2019.1621140>.
- (45) Robinson, J.; Raguseo, F.; Nuccio, S. P.; Liano, D.; Di Antonio, M. DNA G-Quadruplex Structures: More than Simple Roadblocks to Transcription? *Nucleic Acids Res.* **2021**, *49* (15), 8419–8431. <https://doi.org/10.1093/nar/gkab609>.
- (46) Štros, M.; Launholt, D.; Grasser, K. D. The HMG-Box: A Versatile Protein Domain Occurring in a Wide Variety of DNA-Binding Proteins. *Cell. Mol. Life Sci.* **2007**, *64* (19–20), 2590–2606. <https://doi.org/10.1007/s00018-007-7162-3>.
- (47) Malarkey, C. S.; Churchill, M. E. A. The High Mobility Group Box: The Ultimate Utility Player of a Cell. *Trends Biochem. Sci.* **2012**, *37* (12), 553–562. <https://doi.org/10.1016/j.tibs.2012.09.003>.
- (48) Singh, R. K.; Mukherjee, A. Molecular Mechanism of the Intercalation of the SOX-4 Protein into DNA Inducing Bends and Kinks. *J. Phys. Chem. B* **2021**, *125* (15), 3752–3762. <https://doi.org/10.1021/acs.jpcc.0c11496>.
- (49) Scaffidi, P.; Bianchi, M. E. Spatially Precise DNA Bending Is an Essential Activity of the Sox2 Transcription Factor. *J. Biol. Chem.* **2001**, *276* (50), 47296–47302. <https://doi.org/10.1074/jbc.M107619200>.
- (50) Lefebvre, V.; Huang, W.; Harley, V. R.; Goodfellow, P. N.; de Crombrughe, B. SOX9 Is a Potent Activator of the Chondrocyte-Specific Enhancer of the pro Alpha1(II) Collagen Gene. *Mol. Cell. Biol.* **1997**, *17* (4), 2336–2346. <https://doi.org/10.1128/MCB.17.4.2336>.
- (51) Connor, F.; Cary, P. D.; Read, C. M.; Preston, N. S.; Driscoll, P. C.; Denny, P.; Crane-Robinson, C.; Ashworth, A. DNA Binding and Bending Properties of the Postmeiotically Expressed Sry-Related Protein Sox-5. *Nucleic Acids Res.* **1994**, *22* (16), 3339–3346. <https://doi.org/10.1093/nar/22.16.3339>.
- (52) Murugesapillai, D.; McCauley, M. J.; Maher, L. J.; Williams, M. C. Single-Molecule Studies of High-Mobility Group B Architectural DNA Bending Proteins. *Biophys. Rev.* **2017**, *9* (1), 17–40. <https://doi.org/10.1007/s12551-016-0236-4>.
- (53) Love, J. J.; Li, X.; Case, D. A.; Giese, K.; Grosschedl, R.; Wright, P. E. Structural Basis for DNA Bending by the Architectural Transcription Factor LEF-1. *Nature* **1995**, *376* (6543), 791–795. <https://doi.org/10.1038/376791a0>.
- (54) Bowles, J.; Schepers, G.; Koopman, P. Phylogeny of the SOX Family of Developmental Transcription Factors Based on Sequence and Structural Indicators. *Dev. Biol.* **2000**, *227* (2), 239–255. <https://doi.org/10.1006/dbio.2000.9883>.
- (55) Mertin, S. The DNA-Binding Specificity of SOX9 and Other SOX Proteins. *Nucleic Acids Res.* **1999**, *27* (5), 1359–1364. <https://doi.org/10.1093/nar/27.5.1359>.
- (56) Stevanovic, M.; Drakulic, D.; Lazic, A.; Ninkovic, D. S.; Schwirtlich, M.; Mojsin, M. SOX Transcription Factors as Important Regulators of Neuronal and Glial Differentiation During Nervous System Development and Adult Neurogenesis. *Front. Mol. Neurosci.* **2021**, *14*, 654031. <https://doi.org/10.3389/fnmol.2021.654031>.

- (57) Sarkar, A.; Hochedlinger, K. The Sox Family of Transcription Factors: Versatile Regulators of Stem and Progenitor Cell Fate. *Cell Stem Cell* **2013**, *12* (1), 15–30. <https://doi.org/10.1016/j.stem.2012.12.007>.
- (58) Bergsland, M.; Ramsköld, D.; Zaouter, C.; Klum, S.; Sandberg, R.; Muhr, J. Sequentially Acting Sox Transcription Factors in Neural Lineage Development. *Genes Dev.* **2011**, *25* (23), 2453–2464. <https://doi.org/10.1101/gad.176008.111>.
- (59) Kondoh, H.; Kamachi, Y. SOX–Partner Code for Cell Specification: Regulatory Target Selection and Underlying Molecular Mechanisms. *Int. J. Biochem. Cell Biol.* **2010**, *42* (3), 391–399. <https://doi.org/10.1016/j.biocel.2009.09.003>.
- (60) Holmes, Z. E.; Hamilton, D. J.; Hwang, T.; Parsonnet, N. V.; Rinn, J. L.; Wuttke, D. S.; Batey, R. T. The Sox2 Transcription Factor Binds RNA. *Nat. Commun.* **2020**, *11* (1), 1805. <https://doi.org/10.1038/s41467-020-15571-8>.
- (61) Ng, S.-Y.; Bogu, G. K.; Soh, B. S.; Stanton, L. W. The Long Noncoding RNA RMST Interacts with SOX2 to Regulate Neurogenesis. *Mol. Cell* **2013**, *51* (3), 349–359. <https://doi.org/10.1016/j.molcel.2013.07.017>.
- (62) Cajigas, I.; Chakraborty, A.; Lynam, M.; Swyter, K. R.; Bastidas, M.; Collens, L.; Luo, H.; Ay, F.; Kohtz, J. D. Sox2- *Evf2* LncRNA-Mediated Mechanisms of Chromosome Topological Control in Developing Forebrain. *Development* **2021**, *148* (6), dev197202. <https://doi.org/10.1242/dev.197202>.
- (63) Jing, R.; Guo, X.; Yang, Y.; Chen, W.; Kang, J.; Zhu, S. Long Noncoding RNA Q Associates with Sox2 and Is Involved in the Maintenance of Pluripotency in Mouse Embryonic Stem Cells. *Stem Cells* **2020**, *38* (7), 834–848. <https://doi.org/10.1002/stem.3180>.
- (64) Chang, M. V.; Chang, J. L.; Gangopadhyay, A.; Shearer, A.; Cadigan, K. M. Activation of Wingless Targets Requires Bipartite Recognition of DNA by TCF. *Curr. Biol.* **2008**, *18* (23), 1877–1881. <https://doi.org/10.1016/j.cub.2008.10.047>.
- (65) Hrckulak, D.; Kolar, M.; Strnad, H.; Korinek, V. TCF/LEF Transcription Factors: An Update from the Internet Resources. *Cancers* **2016**, *8* (7), 70. <https://doi.org/10.3390/cancers8070070>.
- (66) Cadigan, K. M.; Waterman, M. L. TCF/LEFs and Wnt Signaling in the Nucleus. *Cold Spring Harb. Perspect. Biol.* **2012**, *4* (11), a007906–a007906. <https://doi.org/10.1101/cshperspect.a007906>.
- (67) Santiago, L.; Daniels, G.; Wang, D.; Deng, F.-M.; Lee, P. Wnt Signaling Pathway Protein LEF1 in Cancer, as a Biomarker for Prognosis and a Target for Treatment. *Am. J. Cancer Res.* **2017**, *7* (6), 1389–1406.
- (68) Reya, T.; Clevers, H. Wnt Signalling in Stem Cells and Cancer. *Nature* **2005**, *434* (7035), 843–850. <https://doi.org/10.1038/nature03319>.
- (69) Yan, K. S.; Janda, C. Y.; Chang, J.; Zheng, G. X. Y.; Larkin, K. A.; Luca, V. C.; Chia, L. A.; Mah, A. T.; Han, A.; Terry, J. M.; Ootani, A.; Roelf, K.; Lee, M.; Yuan, J.; Li, X.; Bolen, C. R.; Wilhelmy, J.; Davies, P. S.; Ueno, H.; von Furstenberg, R. J.; Belgrader, P.; Ziraldo, S. B.; Ordóñez, H.; Henning, S. J.; Wong, M. H.; Snyder, M. P.; Weissman, I. L.; Hsueh, A. J.; Mikkelsen, T. S.; Garcia, K. C.; Kuo, C. J. Non-Equivalence of Wnt and R-Spondin Ligands during Lgr5+ Intestinal Stem-Cell Self-Renewal. *Nature* **2017**, *545* (7653), 238–242. <https://doi.org/10.1038/nature22313>.

- (70) Merrill, B. J.; Gat, U.; DasGupta, R.; Fuchs, E. Tcf3 and Lef1 Regulate Lineage Differentiation of Multipotent Stem Cells in Skin. *Genes Dev.* **2001**, *15* (13), 1688–1705. <https://doi.org/10.1101/gad.891401>.
- (71) Hou, L.; Wei, Y.; Lin, Y.; Wang, X.; Lai, Y.; Yin, M.; Chen, Y.; Guo, X.; Wu, S.; Zhu, Y.; Yuan, J.; Tariq, M.; Li, N.; Sun, H.; Wang, H.; Zhang, X.; Chen, J.; Bao, X.; Jauch, R. Concurrent Binding to DNA and RNA Facilitates the Pluripotency Reprogramming Activity of Sox2. *Nucleic Acids Res.* **2020**, *48* (7), 3869–3887. <https://doi.org/10.1093/nar/gkaa067>.
- (72) Ferré-D'Amaré, A. R.; Zhou, K.; Doudna, J. A. A General Module for RNA Crystallization. *J. Mol. Biol.* **1998**, *279* (3), 621–631. <https://doi.org/10.1006/jmbi.1998.1789>.
- (73) Hoggan, D. B.; Chao, J. A.; Prasad, G. S.; Stout, C. D.; Williamson, J. R. Combinatorial Crystallization of an RNA–Protein Complex. *Acta Crystallogr. D Biol. Crystallogr.* **2003**, *59* (3), 466–473. <https://doi.org/10.1107/S0907444902023399>.
- (74) Schultz, S. C.; Shields, G. C.; Steitz, T. A. Crystallization of Escherichia Coli Catabolite Gene Activator Protein with Its DNA Binding Site. *J. Mol. Biol.* **1990**, *213* (1), 159–166. [https://doi.org/10.1016/S0022-2836\(05\)80128-7](https://doi.org/10.1016/S0022-2836(05)80128-7).
- (75) Nowakowski, J.; Shim, P. J.; Joyce, G. F.; Stout, C. D. Crystallization of the 10-23 DNA Enzyme Using a Combinatorial Screen of Paired Oligonucleotides. *Acta Crystallogr. D Biol. Crystallogr.* **1999**, *55* (11), 1885–1892. <https://doi.org/10.1107/S0907444999010550>.
- (76) Berger, I.; Kang, C.; Sinha, N.; Wolters, M.; Rich, A. A Highly Efficient 24-Condition Matrix for the Crystallization of Nucleic Acid Fragments. *Acta Crystallogr. D Biol. Crystallogr.* **1996**, *52* (3), 465–468. <https://doi.org/10.1107/S0907444995013564>.
- (77) Remenyi, A. Crystal Structure of a POU/HMG/DNA Ternary Complex Suggests Differential Assembly of Oct4 and Sox2 on Two Enhancers. *Genes Dev.* **2003**, *17* (16), 2048–2059. <https://doi.org/10.1101/gad.269303>.
- (78) The UniProt Consortium; Bateman, A.; Martin, M.-J.; Orchard, S.; Magrane, M.; Agivetova, R.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E. H.; Britto, R.; Bursteinas, B.; Bye-A-Jee, H.; Coetzee, R.; Cukura, A.; Da Silva, A.; Denny, P.; Dogan, T.; Ebenezer, T.; Fan, J.; Castro, L. G.; Garmiri, P.; Georghiou, G.; Gonzales, L.; Hatton-Ellis, E.; Hussein, A.; Ignatchenko, A.; Insana, G.; Ishtiaq, R.; Jokinen, P.; Joshi, V.; Jyothi, D.; Lock, A.; Lopez, R.; Luciani, A.; Luo, J.; Lussi, Y.; MacDougall, A.; Madeira, F.; Mahmoudy, M.; Menchi, M.; Mishra, A.; Moulang, K.; Nightingale, A.; Oliveira, C. S.; Pundir, S.; Qi, G.; Raj, S.; Rice, D.; Lopez, M. R.; Saidi, R.; Sampson, J.; Sawford, T.; Speretta, E.; Turner, E.; Tyagi, N.; Vasudev, P.; Volynkin, V.; Warner, K.; Watkins, X.; Zaru, R.; Zellner, H.; Bridge, A.; Poux, S.; Redaschi, N.; Aimo, L.; Argoud-Puy, G.; Auchincloss, A.; Axelsen, K.; Bansal, P.; Baratin, D.; Blatter, M.-C.; Bolleman, J.; Boutet, E.; Breuza, L.; Casals-Casas, C.; de Castro, E.; Echioukh, K. C.; Coudert, E.; Cuche, B.; Doche, M.; Dornevil, D.; Estreicher, A.; Famiglietti, M. L.; Feuermann, M.; Gasteiger, E.; Gehant, S.; Gerritsen, V.; Gos, A.; Gruaz-Gumowski, N.; Hinz, U.; Hulo, C.; Hyka-Nouspikel, N.; Jungo, F.; Keller, G.; Kerhornou, A.; Lara, V.; Le Mercier, P.; Lieberherr, D.; Lombardot, T.; Martin, X.; Masson, P.; Morgat, A.;

- Neto, T. B.; Paesano, S.; Pedruzzi, I.; Pilbout, S.; Pourcel, L.; Pozzato, M.; Pruess, M.; Rivoire, C.; Sigrist, C.; Sonesson, K.; Stutz, A.; Sundaram, S.; Tognolli, M.; Verbregue, L.; Wu, C. H.; Arighi, C. N.; Arminski, L.; Chen, C.; Chen, Y.; Garavelli, J. S.; Huang, H.; Laiho, K.; McGarvey, P.; Natale, D. A.; Ross, K.; Vinayaka, C. R.; Wang, Q.; Wang, Y.; Yeh, L.-S.; Zhang, J.; Ruch, P.; Teodoro, D. UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49* (D1), D480–D489. <https://doi.org/10.1093/nar/gkaa1100>.
- (79) Jauch, R.; Ng, C. K. L.; Narasimhan, K.; Kolatkar, P. R. The Crystal Structure of the Sox4 HMG Domain–DNA Complex Suggests a Mechanism for Positional Interdependence in DNA Recognition. *Biochem. J.* **2012**, *443* (1), 39–47. <https://doi.org/10.1042/BJ20111768>.
- (80) Klaus, M.; Prokoph, N.; Girbig, M.; Wang, X.; Huang, Y.-H.; Srivastava, Y.; Hou, L.; Narasimhan, K.; Kolatkar, P. R.; Francois, M.; Jauch, R. Structure and Decoy-Mediated Inhibition of the SOX18/ *Prox1* -DNA Interaction. *Nucleic Acids Res.* **2016**, *44* (8), 3922–3935. <https://doi.org/10.1093/nar/gkw130>.
- (81) Palasingam, P.; Jauch, R.; Ng, C. K. L.; Kolatkar, P. R. The Structure of Sox17 Bound to DNA Reveals a Conserved Bending Topology but Selective Protein Interaction Platforms. *J. Mol. Biol.* **2009**, *388* (3), 619–630. <https://doi.org/10.1016/j.jmb.2009.03.055>.
- (82) Masuzawa, T.; Sato, S.; Niwa, T.; Taguchi, H.; Nakamura, H.; Oyoshi, T. G-Quadruplex-Proximity Protein Labeling Based on Peroxidase Activity. *Chem. Commun.* **2020**, *56* (78), 11641–11644. <https://doi.org/10.1039/D0CC02571B>.
- (83) Tikhonova, P.; Pavlova, I.; Isaakova, E.; Tsvetkov, V.; Bogomazova, A.; Vedekhina, T.; Luzhin, A. V.; Sultanov, R.; Severov, V.; Klimina, K.; Kantidze, O. L.; Pozmogova, G.; Lagarkova, M.; Varizhuk, A. DNA G-Quadruplexes Contribute to CTCF Recruitment. *Int. J. Mol. Sci.* **2021**, *22* (13), 7090. <https://doi.org/10.3390/ijms22137090>.
- (84) Amato, J.; Cerofolini, L.; Brancaccio, D.; Giuntini, S.; Iaccarino, N.; Zizza, P.; Iachettini, S.; Biroccio, A.; Novellino, E.; Rosato, A.; Fragai, M.; Luchinat, C.; Randazzo, A.; Pagano, B. Insights into Telomeric G-Quadruplex DNA Recognition by HMGB1 Protein. *Nucleic Acids Res.* **2019**, *47* (18), 9950–9966. <https://doi.org/10.1093/nar/gkz727>.
- (85) Amato, J.; Madanayake, T. W.; Iaccarino, N.; Novellino, E.; Randazzo, A.; Hurley, L. H.; Pagano, B. HMGB1 Binds to the *KRAS* Promoter G-Quadruplex: A New Player in Oncogene Transcriptional Regulation? *Chem. Commun.* **2018**, *54* (68), 9442–9445. <https://doi.org/10.1039/C8CC03614D>.
- (86) Rubio-Cosials, A.; Battistini, F.; Gansen, A.; Cuppari, A.; Bernadó, P.; Orozco, M.; Langowski, J.; Tóth, K.; Solà, M. Protein Flexibility and Synergy of HMG Domains Underlie U-Turn Bending of DNA by TFAM in Solution. *Biophys. J.* **2018**, *114* (10), 2386–2396. <https://doi.org/10.1016/j.bpj.2017.11.3743>.
- (87) Lonnais, S.; Tarrés-Solé, A.; Rubio-Cosials, A.; Cuppari, A.; Brito, R.; Jaumot, J.; Gargallo, R.; Vilaseca, M.; Silva, C.; Granzhan, A.; Teulade-Fichou, M.-P.; Eritja, R.; Solà, M. The Human Mitochondrial Transcription Factor A Is a Versatile G-Quadruplex Binding Protein. *Sci. Rep.* **2017**, *7* (1), 43992. <https://doi.org/10.1038/srep43992>.

- (88) Hon, J.; Martínek, T.; Zendulka, J.; Lexa, M. Pqsfinder: An Exhaustive and Imperfection-Tolerant Search Tool for Potential Quadruplex-Forming Sequences in R. *Bioinformatics* **2017**, *33* (21), 3373–3379. <https://doi.org/10.1093/bioinformatics/btx413>.
- (89) Puig Lombardi, E.; Londoño-Vallejo, A. A Guide to Computational Methods for G-Quadruplex Prediction. *Nucleic Acids Res.* **2020**, *48* (1), 1–15. <https://doi.org/10.1093/nar/gkz1097>.
- (90) Robinson, J. T.; Thorvaldsdóttir, H.; Winckler, W.; Guttman, M.; Lander, E. S.; Getz, G.; Mesirov, J. P. Integrative Genomics Viewer. *Nat. Biotechnol.* **2011**, *29* (1), 24–26. <https://doi.org/10.1038/nbt.1754>.
- (91) Chen, H.; Long, H.; Cui, X.; Zhou, J.; Xu, M.; Yuan, G. Exploring the Formation and Recognition of an Important G-Quadruplex in a HIF1 α Promoter and Its Transcriptional Inhibition by a Benzo[*c*]Phenanthridine Derivative. *J. Am. Chem. Soc.* **2014**, *136* (6), 2583–2591. <https://doi.org/10.1021/ja412128w>.
- (92) Kaiser, C. E.; Van Ert, N. A.; Agrawal, P.; Chawla, R.; Yang, D.; Hurley, L. H. Insight into the Complexity of the I-Motif and G-Quadruplex DNA Structures Formed in the *KRAS* Promoter and Subsequent Drug-Induced Gene Repression. *J. Am. Chem. Soc.* **2017**, *139* (25), 8522–8536. <https://doi.org/10.1021/jacs.7b02046>.
- (93) Hänsel-Hertsch, R.; Spiegel, J.; Marsico, G.; Tannahill, D.; Balasubramanian, S. Genome-Wide Mapping of Endogenous G-Quadruplex DNA Structures by Chromatin Immunoprecipitation and High-Throughput Sequencing. *Nat. Protoc.* **2018**, *13* (3), 551–564. <https://doi.org/10.1038/nprot.2017.150>.
- (94) Biffi, G.; Tannahill, D.; McCafferty, J.; Balasubramanian, S. Quantitative Visualization of DNA G-Quadruplex Structures in Human Cells. *Nat. Chem.* **2013**, *5* (3), 182–186. <https://doi.org/10.1038/nchem.1548>.
- (95) Chambers, V. S.; Marsico, G.; Boutell, J. M.; Di Antonio, M.; Smith, G. P.; Balasubramanian, S. High-Throughput Sequencing of DNA G-Quadruplex Structures in the Human Genome. *Nat. Biotechnol.* **2015**, *33* (8), 877–881. <https://doi.org/10.1038/nbt.3295>.
- (96) Spivakov, M.; Fisher, A. G. Epigenetic Signatures of Stem-Cell Identity. *Nat. Rev. Genet.* **2007**, *8* (4), 263–271. <https://doi.org/10.1038/nrg2046>.
- (97) Liu, X.; Huang, J.; Chen, T.; Wang, Y.; Xin, S.; Li, J.; Pei, G.; Kang, J. Yamanaka Factors Critically Regulate the Developmental Signaling Network in Mouse Embryonic Stem Cells. *Cell Res.* **2008**, *18* (12), 1177–1189. <https://doi.org/10.1038/cr.2008.309>.
- (98) Abdelalim, E. M.; Emara, M. M.; Kolatkar, P. R. The SOX Transcription Factors as Key Players in Pluripotent Stem Cells. *Stem Cells Dev.* **2014**, *23* (22), 2687–2699. <https://doi.org/10.1089/scd.2014.0297>.
- (99) Masui, S.; Nakatake, Y.; Toyooka, Y.; Shimosato, D.; Yagi, R.; Takahashi, K.; Okochi, H.; Okuda, A.; Matoba, R.; Sharov, A. A.; Ko, M. S. H.; Niwa, H. Pluripotency Governed by Sox2 via Regulation of Oct3/4 Expression in Mouse Embryonic Stem Cells. *Nat. Cell Biol.* **2007**, *9* (6), 625–635. <https://doi.org/10.1038/ncb1589>.
- (100) Rizzino, A. Sox2 and Oct-3/4: A Versatile Pair of Master Regulators That Orchestrate the Self-renewal and Pluripotency of Embryonic Stem Cells. *Wiley*

- Interdiscip. Rev. Syst. Biol. Med.* **2009**, 1 (2), 228–236.
<https://doi.org/10.1002/wsbm.12>.
- (101) Liu, Z.; Legant, W. R.; Chen, B.-C.; Li, L.; Grimm, J. B.; Lavis, L. D.; Betzig, E.; Tjian, R. 3D Imaging of Sox2 Enhancer Clusters in Embryonic Stem Cells. *eLife* **2014**, 3, e04236. <https://doi.org/10.7554/eLife.04236>.
 - (102) Wei, C.-L.; Nicolis, S. K.; Zhu, Y.; Pagin, M. Sox2-Dependent 3D Chromatin Interactomes in Transcription, Neural Stem Cell Proliferation and Neurodevelopmental Diseases. *J. Exp. Neurosci.* **2019**, 13, 117906951986822. <https://doi.org/10.1177/1179069519868224>.
 - (103) Bertolini, J. A.; Favaro, R.; Zhu, Y.; Pagin, M.; Ngan, C. Y.; Wong, C. H.; Tjong, H.; Vermunt, M. W.; Martynoga, B.; Barone, C.; Mariani, J.; Cardozo, M. J.; Tabanera, N.; Zambelli, F.; Mercurio, S.; Ottolenghi, S.; Robson, P.; Creighton, M. P.; Bovolenta, P.; Pavesi, G.; Guillemot, F.; Nicolis, S. K.; Wei, C.-L. Mapping the Global Chromatin Connectivity Network for Sox2 Function in Neural Stem Cell Maintenance. *Cell Stem Cell* **2019**, 24 (3), 462-476.e6. <https://doi.org/10.1016/j.stem.2019.02.004>.
 - (104) Gao, T.; Qian, J. EnhancerAtlas 2.0: An Updated Resource with Enhancer Annotation in 586 Tissue/Cell Types across Nine Species. *Nucleic Acids Res.* **2019**, gkz980. <https://doi.org/10.1093/nar/gkz980>.
 - (105) Rocher, V.; Genais, M.; Nassereldine, E.; Mourad, R. DeepG4: A Deep Learning Approach to Predict Cell-Type Specific Active G-Quadruplex Regions. *PLOS Comput. Biol.* **2021**, 17 (8), e1009308. <https://doi.org/10.1371/journal.pcbi.1009308>.
 - (106) He, J.; Shen, L.; Wan, M.; Taranova, O.; Wu, H.; Zhang, Y. Kdm2b Maintains Murine Embryonic Stem Cell Status by Recruiting PRC1 Complex to CpG Islands of Developmental Genes. *Nat. Cell Biol.* **2013**, 15 (4), 373–384. <https://doi.org/10.1038/ncb2702>.
 - (107) Mohanty, J.; Barooah, N.; Dhamodharan, V.; Harikrishna, S.; Pradeepkumar, P. I.; Bhasikuttan, A. C. Thioflavin T as an Efficient Inducer and Selective Fluorescent Sensor for the Human Telomeric G-Quadruplex DNA. *J. Am. Chem. Soc.* **2013**, 135 (1), 367–376. <https://doi.org/10.1021/ja309588h>.
 - (108) Xu, S.; Li, Q.; Xiang, J.; Yang, Q.; Sun, H.; Guan, A.; Wang, L.; Liu, Y.; Yu, L.; Shi, Y.; Chen, H.; Tang, Y. Thioflavin T as an Efficient Fluorescence Sensor for Selective Recognition of RNA G-Quadruplexes. *Sci. Rep.* **2016**, 6 (1), 24793. <https://doi.org/10.1038/srep24793>.
 - (109) Zhang, X. F.; Xu, H. M.; Han, L.; Li, N. B.; Luo, H. Q. A Thioflavin T-Induced G-Quadruplex Fluorescent Biosensor for Target DNA Detection. *Anal. Sci.* **2018**, 34 (2), 149–153. <https://doi.org/10.2116/analsci.34.149>.
 - (110) Do, N. Q.; Phan, A. T. Monomer-Dimer Equilibrium for the 5'-5' Stacking of Propeller-Type Parallel-Stranded G-Quadruplexes: NMR Structural Study. *Chem. - Eur. J.* **2012**, 18 (46), 14752–14759. <https://doi.org/10.1002/chem.201103295>.
 - (111) Ambrus, A.; Chen, D.; Dai, J.; Bialis, T.; Jones, R. A.; Yang, D. Human Telomeric Sequence Forms a Hybrid-Type Intramolecular G-Quadruplex Structure with Mixed Parallel/Antiparallel Strands in Potassium Solution. *Nucleic Acids Res.* **2006**, 34 (9), 2723–2735. <https://doi.org/10.1093/nar/gkl348>.

- (112) Uhlén, M.; Fagerberg, L.; Hallström, B. M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, Å.; Kampf, C.; Sjöstedt, E.; Asplund, A.; Olsson, I.; Edlund, K.; Lundberg, E.; Navani, S.; Szgyarto, C. A.-K.; Odeberg, J.; Djureinovic, D.; Takanen, J. O.; Hober, S.; Alm, T.; Edqvist, P.-H.; Berling, H.; Tegel, H.; Mulder, J.; Rockberg, J.; Nilsson, P.; Schwenk, J. M.; Hamsten, M.; von Feilitzen, K.; Forsberg, M.; Persson, L.; Johansson, F.; Zwahlen, M.; von Heijne, G.; Nielsen, J.; Pontén, F. Tissue-Based Map of the Human Proteome. *Science* **2015**, *347* (6220), 1260419. <https://doi.org/10.1126/science.1260419>.
- (113) Zyner, K. G.; Simeone, A.; Flynn, S. M.; Doyle, C.; Marsico, G.; Adhikari, S.; Portella, G.; Tannahill, D.; Balasubramanian, S. G-Quadruplex DNA Structures in Human Stem Cells and Differentiation. *Nat. Commun.* **2022**, *13* (1), 142. <https://doi.org/10.1038/s41467-021-27719-1>.
- (114) Quan, J.; Tian, J. Circular Polymerase Extension Cloning for High-Throughput Cloning of Complex and Combinatorial DNA Libraries. *Nat. Protoc.* **2011**, *6* (2), 242–251. <https://doi.org/10.1038/nprot.2010.181>.
- (115) Edwards, A. L.; Garst, A. D.; Batey, R. T. Determining Structures of RNA Aptamers and Riboswitches by X-Ray Crystallography. In *Nucleic Acid and Peptide Aptamers*; Mayer, G., Ed.; Methods in Molecular Biology; Humana Press: Totowa, NJ, 2009; Vol. 535, pp 135–163. https://doi.org/10.1007/978-1-59745-557-2_9.
- (116) Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinforma.* **2014**, *47* (1). <https://doi.org/10.1002/0471250953.bi1112s47>.
- (117) Yu, G.; Wang, L.-G.; He, Q.-Y. ChIPseeker: An R/Bioconductor Package for ChIP Peak Annotation, Comparison and Visualization. *Bioinformatics* **2015**, *31* (14), 2382–2383. <https://doi.org/10.1093/bioinformatics/btv145>.
- (118) Hsu, F.; Kent, W. J.; Clawson, H.; Kuhn, R. M.; Diekhans, M.; Haussler, D. The UCSC Known Genes. *Bioinformatics* **2006**, *22* (9), 1036–1046. <https://doi.org/10.1093/bioinformatics/btl048>.
- (119) Yu, G.; He, Q.-Y. ReactomePA: An R/Bioconductor Package for Reactome Pathway Analysis and Visualization. *Mol. Biosyst.* **2016**, *12* (2), 477–479. <https://doi.org/10.1039/C5MB00663E>.