The FOXP2-TBR1 Interaction and its Role in the Regulation of DNA Binding

by

Ashleigh Anne Blane

(940523)

Thesis

Submitted in fulfilment of the requirements for the degree

Philosophiae Doctor

in

Molecular and Cell Biology

in the Faculty of Science, University of the Witwatersrand, Johannesburg, South Africa

Supervisor: Dr Sylvia Fanucchi

January 2021

DECLARATION

I, Ashleigh Anne Blane (940523), am a student registered for the degree of Doctor of Philosophy (PhD) at the University of the Witwatersrand.

I hereby declare the following:

- I am aware that plagiarism is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where explicitly indicated otherwise and acknowledged.
- I have not submitted this work before for any other degree or examination at this or any other University.
- I have followed the required conventions in referencing the work of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is evidence that this is not my own unaided work or that I have failed to acknowledge the sources of the ideas in my writing.

A.A Blane 08 January 2021

RESEARCH OUTPUTS

Conference Oral and poster presentations

9th Cross Faculty Postgraduate Symposium:

- Poster Presentation
- Structural characterisation of the TBR1 T–Box domain
- Ashleigh Blane and Sylvia Fanucchi
- University of the Witwatersrand, Johannesburg, South Africa
- 29 – 30 October 2018

Molecular Biosciences Research Thrust Research Day 2018:

- Oral Presentation
- Expression, purification, and structural characterisation of the TBR1 T–Box domain
- Ashleigh Blane and Sylvia Fanucchi
- University of the Witwatersrand, Johannesburg, South Africa
- 29 November 2018

10th Cross Faculty Postgraduate Symposium:

- Poster Presentation (1st place in the Faculty of Science)
- A study on the dimerisation and DNA binding of the TBR1 T–Box domain
- Ashleigh Blane and Sylvia Fanucchi
- University of the Witwatersrand, Johannesburg, South Africa
- 3 – 4 September 2019

Molecular Biosciences Research Thrust Research Day 2019:

- Poster Presentation (3rd place, PhD/Post Doc division)
- The Role of the TBR1 – FOXP2 Interaction on TBR1 DNA – Binding
- Ashleigh Blane and Sylvia Fanucchi
- University of the Witwatersrand, Johannesburg, South Africa
- 28 November 2019

11th Cross Faculty Postgraduate Symposium:

- Grad Flash (5th place)
- The FOXP2–TBR1 interaction and its role in regulation of DNA binding
- Ashleigh Blane and Sylvia Fanucchi
- University of the Witwatersrand, Johannesburg, South Africa
- 12–16 October 2020

Publications

Previous publications

Blane, A. & Fanucchi, S. (2015) Effect of pH on the Structure and DNA Binding of the FOXP2 Forkhead Domain. *Biochemistry*, **54**(25), 4001–4007.

Webb, H., Steeb, O., Blane, A., Rotherman, L., Aron, S., Machanick, P., Dirr, H. & Fanucchi, S. (2017) The FOXP2 forkhead domain binds to a variety of DNA sequences with different rates and affinities. *The Journal of Biochemistry*, **126**(1), 45-54.

Blane, A., Dirr, H. & Fanucchi, S. (2018) A Phosphomimetic Study Implicates Ser557 in Regulation of FOXP2 DNA Binding. *The Protein Journal*, **37**(4),311–323.

Thulo, M., Rabie, MA., Pahad, N., Donald, HL., Blane, AA., Permual, CM., Penedo, JC and Fanucchi, S. (2021) The influence of various regions of the FOXP2 sequence on its structure and DNA-binding function. *Bioscience Reports*, **41**(1),1-13.

ABSTRACT

Forkhead Box P2 (FOXP2) and T–Brain Related Protein 1 (TBR1) are transcription factors associated with neurological development and disorders such as autism spectrum disorder. FOXP2 and TBR1 each contain highly conserved DNA binding domains, the forkhead domain (FHD) and the T–Box domain (T–Box), respectively. TBR1 and FOXP2 appear to interact via the T–Box domain of TBR1 and either the FHD or N–terminal region of FOXP2. Mutations in both proteins have been implicated in autism and some of these result in the disruption of the interaction between the two proteins. The aim of this study was to determine whether the TBR1 T–Box domain and the FOXP2 FHD interact and the subsequent effect of the interaction on DNA binding. Both proteins were successfully expressed in *E. coli* cells and purified using liquid chromatography. The tertiary and quarterly structure of each protein was evaluated using tryptophan fluorescence and size exclusion chromatography. The proteins were confirmed to be functional by assessing their DNA binding functions using electrophoretic mobility shift assays. As the DNA binding properties of the TBR1 T–Box have never been characterised, DNA binding to T–box was further characterised using isothermal titration calorimetry (ITC) and single molecule Förster resonance energy transfer (smFRET) using DNA containing a single site (SSL) and two palindromic sites (PAL). Finally, fluorescence anisotropy (FA) was used to assess the effect of the T–Box–FHD interaction on each protein's DNA binding function. Both T–Box and FHD were found to be monomeric in solution and had a folded structure with DNA binding capabilities. T–Box bound both SSL and Pal DNA by interacting with both the major and minor grooves, like other T–Box domains have exhibited in crystal structures. T–Box bound the SSL DNA with a 10X greater affinity than Pal DNA, likely because of interactions that occur in regions flanking the binding site. Furthermore, this work found evidence to suggest that although T–Box can occupy both sites on Pal DNA, T–Box binds these sites independently rather than dimerising in solution and then binding as a dimer. Finally, FA studies showed that the FHD–T–Box interaction disrupts the interaction of the FHD with DNA whereas T–Box can interact with DNA and FHD simultaneously. This work provides insight into a complex regulatory pathway that can cause neurological disorders such as ASD when mis–regulation occurs.

To my most favourite humans;

Mom and Dad

Craig, Lol and Kirst

Mr Dan

"Mischief Managed"

– *Harry Potter and the Prisoner of Azkaban* by J.K. Rowling

ACKNOWLEDGEMENTS

A huge thank you to my supervisor Dr Sylvia Fanucchi, you gave me the freedom to pursue my project and gave me the help and guidance I needed to succeed. Thank you for the hours of chats, laughs and French toast.

I wish to show my gratitude to Prof Heini Dirr, Prof Yasien Sayed and Dr Ikechukwu Achilonu. I am extremely grateful for your insight and advice over the years.

I would like to express my very great appreciation to Dr Juan Carlos Penedo for his invaluable technical and intellectual contribution to smFRET.

Thank you to all my colleagues in the PSFRU and FOXy group, present and past, for all the support, discussions, and assistance. A special thanks to Allison Williams, Cardon Perumal, Ilan Kirkel, Jake Zondagh, Joni Symon, Kimberley Wiid and Monare Thulo.

A very special thank you to my best lab friends Dean Sherry, Gary Robertson, Gavin Morris, Heather Donald, Mia Færch, Naadira Pahad, Roland Worth and Zaahida Sheik Ismail. I would never had made it to the end without you guys. You all made the bad days good and the good days better.

I am enormously grateful to the National research foundation and the Wits postgraduate merit award for funding.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

AAVP5	Adeno–Associated Virus Type 2 P5 promoter
AIRS	Ambiguous Interaction Restraints
AmpR	Ampicillin Resistance
ART	Acid Rich Tail
ASD	Autism Spectrum Disorder
AUTS2	Autism Susceptibility Gene 2
A_{xxx}	Absorbance at xxx nm
BRET	Bioluminescence Resonance Energy Transfer
C	Concentration
CASK	Calcium/Calmodulin–Dependent Serine Protein Kinase
CINAP	CASK–Interacting Nucleosome Assembly Protein
CNTNAP2	Contactin–Associated Protein–Like 2 Protein
CTBP2	C–Terminal–Binding Protein 2
DBD	DNA Binding Domain
DNA	Deoxyribonucleic Acid
DPE	Downstream Promoter Element
DTT	1,4–Dithiothreitol
EDTA	Ethylenediaminetetraacetic Acid
EMSA	Electrophoretic Mobility Shift Assay
EOMES	Eomesodermin
Ɛ	Extinction Coefficient
F*	Fraction Number
FA	Fluorescence Anisotropy
FHD	Forkhead Domain
FOG1	Friend of GATA Protein 1
FOX	Forkhead Box
FOXA3	Forkhead Box A3
FOXP	Forkhead Box P
FOXP1	Forkhead Box P1
FOXP2	Forkhead Box P2
FOXP3	Forkhead Box P3
FOXP4	Forkhead Box P4
FRET	Förster Resonance Energy Transfer
FT	Flow Through
G Factor	Grating Factor
GATA1	GATA–Binding Factor 1
GRIN2B	Glutamate Ionotropic Receptor NMDA Type Subunit 2B
H	Hours
HADDOCK	High Ambiguity Driven Docking Approach
His–tag	Hexahistidine Tag
HTH	Helix–Turn–Helix
ID	Intellectual Disability
IMAC	Immobilised Metal Affinity Chromatography
INR	Initiator Element

IPTG	Isopropyl β–D–1–Thiogalactopyranoside
ITC	Isothermal Titration Calorimetry
K_a	Association Constant
K_d	Dissociation Constant
K_{OBS}	Observed Rate
k_{OFF}	Off Rate
k_{ON}	On Rate
LZ	Leucine Zipper
Mw	Molecular Weight
Nel	Nelson DNA
Nel C	Nelson CNTNAP2 Promoter
Nel S	Nelson short
NLS	Nuclear Localisation Sequence
NMDA	N–Methyl–D–Aspartate
NMDAR	N–Methyl–D–Aspartate Receptor
NR2F1	COUP transcription factor 1
NR2F2	COUP transcription factor 2
NTA	Nitrilotriacetic Acid
P	Pellet
PAGE	Polyacrylamide Gel Electrophoresis
Pal	T–Box palindrome DNA
Pal L	T–Box palindrome long
PCR	Polymerase Chain Reaction
PIC	Preinitiation Complex
PMSF	Phenylmethylsulfonyl Fluoride
PQ	Polyglutamine Rich Region
PTM	Post Translational Modification
RELN	Reelin
RNA	Ribonucleic Acid
ROX	Rhodamine X
S	Supernatant
SDS	Sodium Dodecyl Sulfate
SEC	Size Exclusion Chromatography
smFRET	Single Molecule Förster Resonance Energy Transfer
SOC	Super Optimal Broth with Catabolite Repression
SPR	Surface Plasmon Resonance
SSL	T–Box Single Site Long DNA
SSS	T–Box Single Site Short
TBE	Tris Borate EDTA
T–BET	T–Box Protein Expressed in T Cells
TBP	TATA Binding Protein
TBR1	T–Brain Protein 1
TBR2	T–Brain Protein 2
TBX1	T–Box Protein 1
TBX21	T–Box Protein 21
TBX3	T–Box Protein 3
TBX5	T–Box Protein 5

TCEP	Tris(2–Carboxyethyl)Phosphine
TEMED	N,N,N′,N′–Tetramethyl Ethylenediamine
TIR	Total Internal Reflection
TM	Melting Temperature
UV	Ultra–Violet
Ve	Elution Volume
Vo	Void Volume
W*	Wash fraction number
W–HTH	Winged Helix–Turn–Helix
WT	Wild Type
Xbra	T protein for *Xenopus*
YY1	Yin Yang 1
ZF	Zinc Finger
l	Pathlength
n	Number of Amino Acid Residues

The IUPAC abbreviations for the chemical elements and the IUPAC–IUBMB abbreviations for amino acids and DNA nucleotide bases were used.

INTRODUCTION

1.1. TRANSCRIPTION

The central dogma of molecular biology states that in cells, DNA is used to produce RNA, which is subsequently used to produce protein (Crick 1970). Transcription is an important step in this process, where RNA polymerases use DNA templates to produce RNA (**Figure 1**).

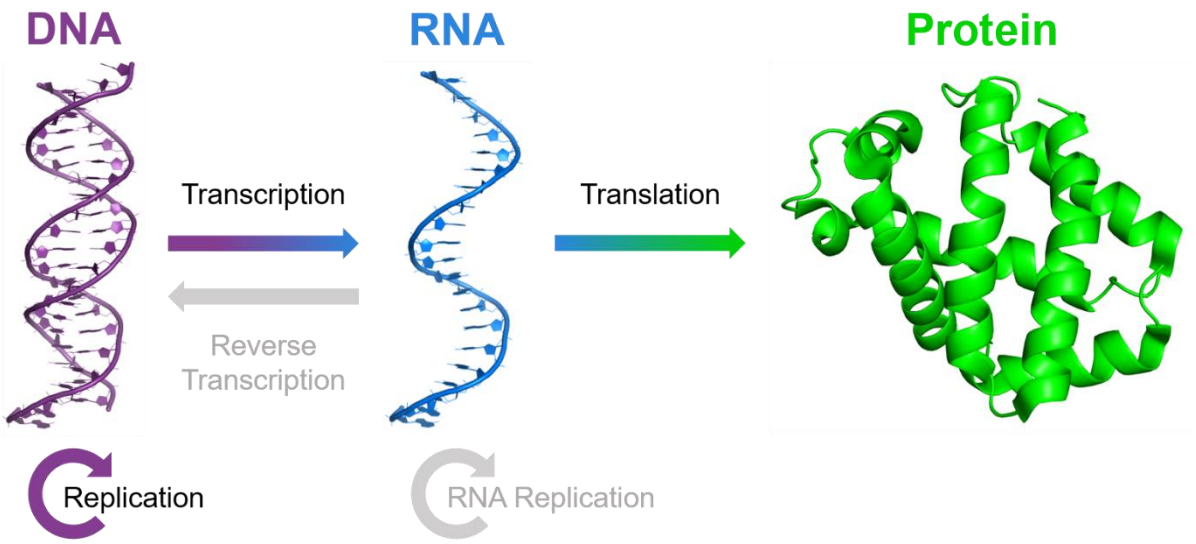


Figure 1: The central dogma of molecular biology.

In most cells, a DNA template is used in the production of RNA in a process called transcription. RNA is then used as a template to produce protein by a process known as translation. In special circumstances RNA may be replicated or used to create DNA (reverse transcription)(Crick 1970). PDB ID: 1h97 (Trematode haemoglobin from *Paramphistomum epiclitum*)(Pesce *et al.* 2001). Image rendered with PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.

Transcription factors are responsible for the basal and differential regulation of transcription. Some genes, known as housekeeping genes or constitutive genes, are transcribed continually at a constant level. They are responsible for maintenance of basic cellular function and undergo basal transcription regulated by general/basal transcription factors. The general transcription factors are classified as the transcription factors that form the most basic complex at a promoter and initiate transcription (Latchman 2005; Pedersen *et al.* 1999). Some genes, however, are transcribed differentially (facultative genes). This differential transcription is regulated by specific transcription factors binding to enhancer and silencer regions on DNA. Differential expression of genes allows multiple cell and tissue types to form from the same genetic material. Highly regulated differential expression of multiple genes also allows cells to react instantly to even slight changes in environment (Huffman & Brennan 2002).

1.2. Transcription of Housekeeping genes

In eukaryotes, for RNA polymerase II to transcribe DNA to form messenger RNA it must form a complex with general transcription factors at the promoter site on the DNA. This complex is known as the basal transcription complex or pre–initiation complex (PIC) (**Figure 2**). This basal transcription complex formation is all that is required for the expression of housekeeping genes (Conaway & Conaway 1993; Grünberg & Hahn 2013; Roeder 1991; Sainsbury *et al.* 2015; Van Dyke *et al.* 1988)

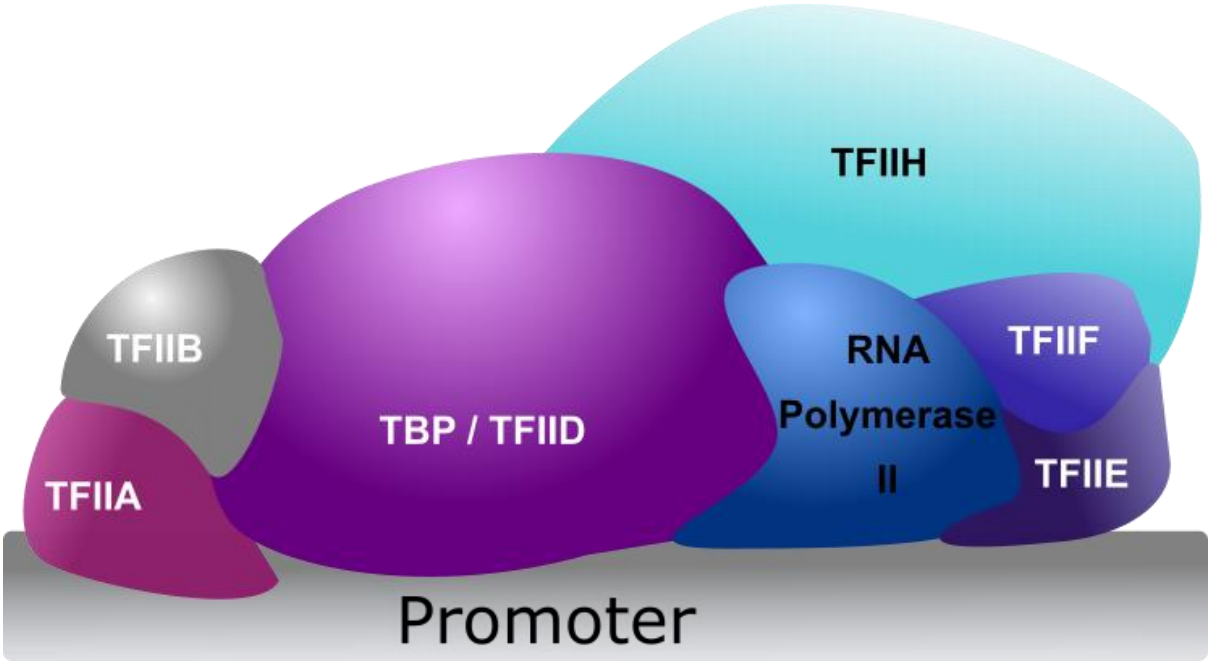


Figure 2: The basal transcription complex or preinitiation complex (PIC) that forms on the promoter and initiates transcription.

The TATA binding protein (TBP) otherwise known as TFIID binds the promoter first. It is followed by RNA polymerase II and the other general transcription factors (TFIIA, TFIIB, TFIIF, TFIIE and TFIIH) which are recruited in this order. (Roeder 1991). Image created using Inkscape 0.92.4

1.3. Regulation of Transcription

The regulation of differentially expressed genes is controlled by specific transcription factors. These transcription factors tend to have multiple domains which generally facilitate DNA binding and protein–protein interactions. The DNA binding domains are involved in specifically binding regulatory sites on the DNA to regulate expression of specific genes. Transcription factors either act as activators or repressors. These actions generally occur through protein–protein interactions with the PIC (Latchman 1997).

Regulation of the transcription of genes is very important as dysregulation can result in disease. Differential regulation of genes allows cells to differentiate and to react to stimuli. This is generally achieved through signal transduction pathways, which through

multiple proteins, relay information from the cell's surface to the nucleus where transcription factors are the end targets (Tootle & Rebay 2005). Transcription is regulated in numerous ways and is in fact a complex combination of events including interactions amongst DNA binding sites, transcription activators and transcription repressors (Latchman 2004).

1.3.1. Initiation of Transcription

Transcription is initiated when the PIC forms on the promoter region of DNA (**Figure 2**). Common promoters in eukaryotes are the TATA box (found 25 to 30 bp upstream of the transcription start site), initiator sequence (which spans the transcription start site) and the downstream promoter element (which is located 25 bp downstream of the transcription start site) (**Figure 3**). First the TBP (TATA box binding protein) binds directly to promoter or proximal promoter elements after which several other components of the PIC, including polymerase II, assemble at the promoter and the PIC spans from position –30 to +30 (Sainsbury *et al.* 2015).

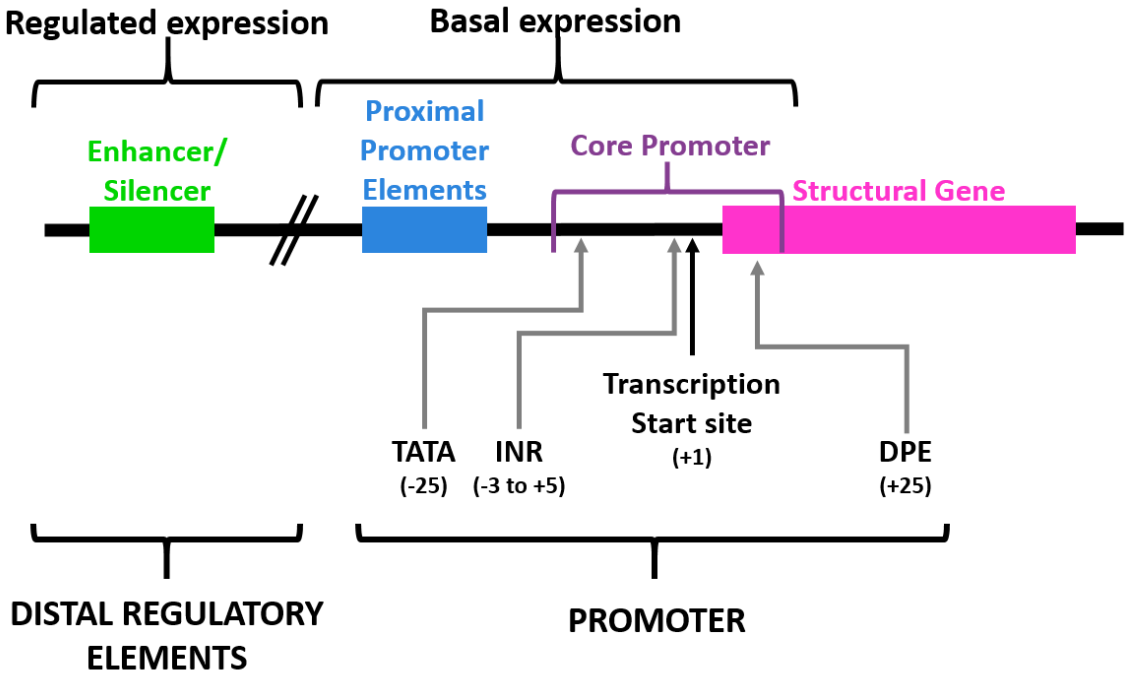


Figure 3: Structure of an RNA polymerase II promoter.

The structural gene (pink) is transcribed when the transcription initiation complex forms at the core promoter (purple). Common eukaryotic promoters (TATA– TATA box, INR– Initiator element and DPE– Downstream promoter element) from part of the core promoter. The promoter is considered to contain the core promoter and proximal promoter elements (blue). Enhancer/ silencer sites (green) are bound by transcription factors which either activate or repress transcription of specific genes. These are referred to as distal regulatory elements.

Eukaryotes contain regulatory regions of DNA known as enhancers or silencers to which specific transcription factors bind in order to regulate transcription. Transcription factors which act as activators bind enhancer regions on the DNA, whereas transcription factors which act as repressors bind the silencer regions on the DNA. Neither transcription activators nor repressors typically directly bind the promoter element, but they may interact with the PIC which forms at the promoter. The enhancer and silencer sites in the DNA can be found kilobases before and after the transcription start sites (**Figure 4**). These regulatory regions of DNA contain binding sites for multiple transcription factors and thus the control of transcription is also dependent on the types of transcription factors present in the cell at any given time (Pedersen *et al.* 1999).

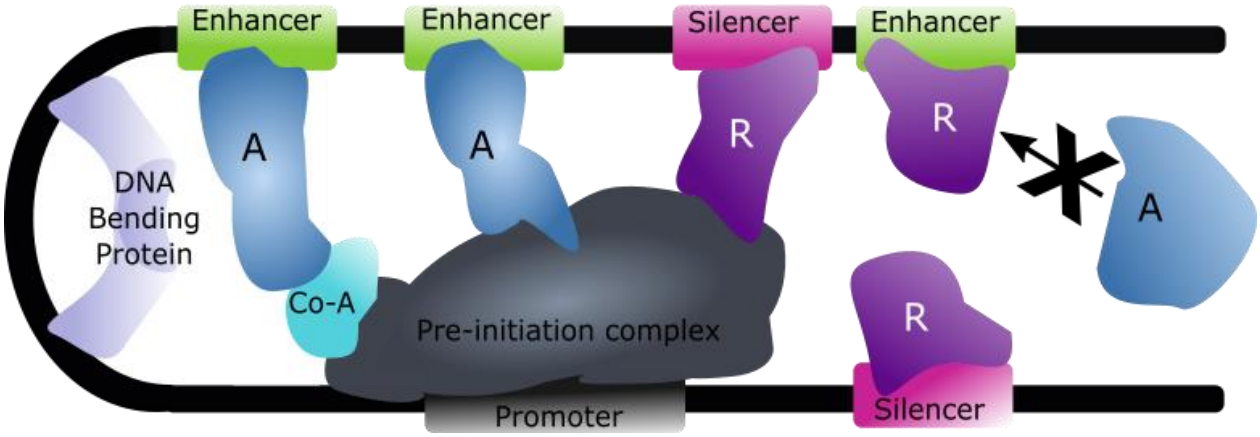


Figure 4: Transcription factors bound to the preinitiation complex and enhancer/ silencers.

The preinitiation complex (PIC, dark grey) is bound to the promoter. The DNA is looped by a DNA bending protein (lilac). Activators (A, blue) bound to enhancers either interact with the PIC directly or indirectly through co–activators (cyan). Repressors (R, purple) bind silencers and either prevent activators from binding enhancer sites or prevent them from binding the PIC. Image created using Inkscape 0.92.4

An important factor in transcription regulation is the structure of chromatin as this affects the accessibility of DNA to transcription factors and RNA polymerase (Olins & Olins, 2003). A nucleosome is a strand of DNA that is wrapped twice around a complex of eight histone proteins. Nucleosomes represent the basic structure of chromatin which is further divided into the more uncompact euchromatin form and the more compact heterochromatin form. In heterochromatin, nucleosomes are packed together by histone H1 which results in a tightly packed chromatin fibre. Typically, genes located in euchromatin are more readily transcribed as they are accessible to transcription factors. In addition to the density of chromatin packing, the location of different enhancer and silence sites can vary relative to the promoter regions. This is a result of how the chromosome is packed, as chromosomal looping can occur to bring parts of the

sequence that would otherwise be far away from each other, close together(Cao *et al.* 2015).

1.3.2. Transcription Activators and Repressors

Transcription activation is defined as the increase in the rate of gene expression whereas transcription repression is defined as the decrease in the rate of gene expression. The way in which transcription factors act as either repressors or activators varies. Transcription factors function as transcription activators by interacting with either the PIC or other co–activators. They activate transcription by either increasing the rate of the basal transcription complex formation or enhancing its level of activity (Roberts 2000). Many transcription factors act as transcriptional repressors. The first mechanism of transcriptional repression is competitive binding to the transcription activator's DNA binding site (**Figure *5***). Transcription repression is sometimes achieved by binding to the activation domain on a transcription activator or by interaction (either directly or indirectly) with members of the PIC. Transcription repressors can also bind to their own specific regulatory DNA sequences (Cowell 1994; Latchman 1997).

Some transcription factors can even act as both activators and repressors. (Sharrocks 2001). For example, the nuclear transcription factor Y subunit gamma can both activate and repress the Human von Willebrand factor gene depending on which transcription co–factors are recruited and also depending on the location of its different DNA binding sites (Peng & Jahroudi 2002). Yin Yang 1 (YY1) is another transcription factor that has been proven to function as both a transcriptional activator and repressor (Weill *et al.* 2003). It is hypothesised to function as a repressor by employing multiple mechanisms which include displacing transcription activators from binding sites which overlap activator binding sites, interacting with co–repressors and transcription activators and bending DNA. It functions as an activator when it interacts with co–activators which block the repressor domains and expose the activation domain (Shrivastava & Calame 1994; Yang *et al.* 1997). The mechanisms of transcription activation and/or repression for a transcription factor can be different for each gene they regulate and can differ in each cell type that expresses the factor.

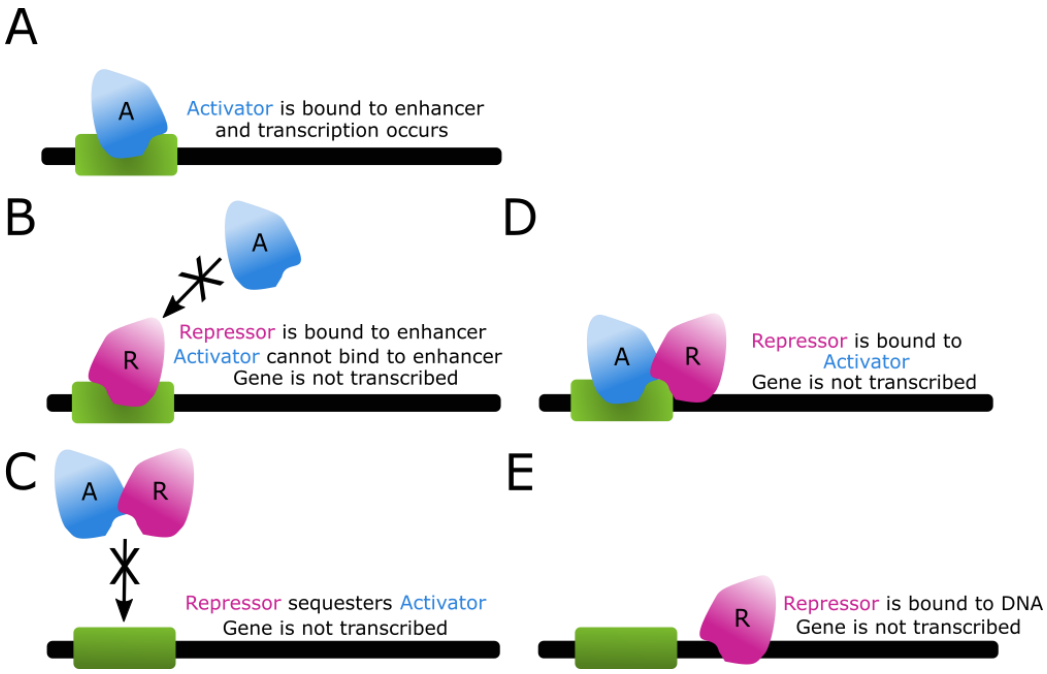


Figure 5: Mechanisms of repression of gene expression by transcription factors.

A: Transcription of the gene is active when the activator (A) is bound to the enhancer region of the DNA. **B:** Transcription is repressed when the repressor (R) competitively binds the activator's (A) DNA binding site. **C:** The repressor (R) binds the activator (A) preventing the activator from binding the DNA, therefore, repressing gene expression. **D:** The repressor (R) directly binds the activator (A) or the preinitiation complex thereby inhibiting activation of gene expression. **E:** The repressor (R) binds to a specific inhibitory DNA sequence (silencer) preventing gene expression. Image created using Inkscape 0.92.4

1.3.3. Regulation of Transcription Factors

Since transcription factors are responsible for the tight regulation of gene expression they are, themselves highly regulated. Regulation of transcription factors occurs by a variety of mechanisms of which synthesis, nuclear localisation, binding of ligands, interactions with other proteins or post translational modifications are examples (Latchman 1997).

The first mechanism by which a transcription factor is regulated is the expression of the gene and subsequent synthesis of the transcription factor. Transcription factor gene expression, as with any protein, is regulated. POU domain class 5 transcription factor 1 otherwise known as Octamer–Binding Transcription Factor 4 is an example of a transcription factor that regulates the expression of its own gene. Octamer–Binding Transcription Factor 4 acts as a repressor for the expression of itself and thus forms a negative feedback loop (Pan *et al.* 2006).

Transcription factors function in the nucleus where transcription occurs and thus control over their subcellular location will result in regulation of their transcriptional effect. Many transcription factors and proteins contain small stretches of sequence, known as

nuclear localisation signals (NLS), which facilitate their transport through the nuclear envelope via interactions with NLS–binding–protein. Nuclear import via some NLS is enhanced by phosphorylation, binding of ligands and the oligomerisation state of the transcription factor (Whiteside & Goodbourn 1993) and these effects thus indirectly result in the regulation of transcription via transcription factors containing NLS.

Post translational modification (PTM) of transcription factors is one of the ways in which transcription factors with highly similar binding sequences can regulate very distinct genes. Transcription factors have been shown to undergo many types of PTMs. These include phosphorylation, acetylation, ubiquintation, glycosylation and sumoylation (Benayoun & Veitia 2009). PTMs of transcription factors, either individually or in combination, have been shown to regulate transcription factors by influencing protein–protein interactions, DNA interactions, cellular localisation, activity and stability (Gill 2004; Tootle & Rebay 2005; Verger *et al.* 2003). PTMs have been shown to occur on multiple sites on a protein (multisite modification). This is important as multiple pathways can converge on the same target transcription factor via different modifications. These different modification patterns present on the target transcription factor at any given time can cause different structural and functional changes (Benayoun & Veitia 2009; Tootle & Rebay 2005).

The formation of large multiprotein complexes is fundamental to transcription. Many transcription factors are required to form these complexes in order to bind specific DNA sites, recruit additional transcription factors or co–factors as well as members of the PIC. For example, the interaction between Friend of GATA protein 1 (FOG1) and GATA–binding factor 1 (GATA1) can either activate or repress transcription of GATA1 regulated genes depending on the binding sites present (Fox *et al.* 1998). FOG1 acts as a repressor when it interacts with co–repressor C–terminal–binding protein 2 (Fox *et al.* 1999). The formation of homo and heterodimers of transcription factors themselves can be a powerful regulatory mechanism. A monomer may bind different DNA sites to the dimeric form of a protein and different combinations of heterodimers can, therefore, further alter the DNA binding specificity. Dimerisation/oligomerisation being a concentration dependent event can also alter regulation of transcription depending on the amount of each transcription factor expressed at any given time. The concentrations of various monomers, any PTMs and the binding affinities for other monomers will influence which oligomers form and subsequently how transcription is regulated (Amoutzias *et al.* 2008).

1.4. Protein–DNA Interactions

Protein–DNA interactions are involved in numerous processes including transcription, DNA packing, DNA repair, cell differentiation and translation. These interactions normally occur via the binding of the protein to a short specific site on the DNA. When these interactions occur, it is not uncommon for both the protein and DNA to undergo conformational changes (Brennan *et al.* 1990; Masse *et al.* 2002; Thompson & Landy 1988). In the context of transcription regulation, the formation of large complexes consisting of a variety of proteins is required to rapidly form at specific DNA sites located in the genome. This is quite a feat considering not only the large size of the human genome but also that many of these sites are inaccessible due to packing of the DNA. So how do transcription factors bind their target sequences with specificity so quickly?

1.4.1. DNA Scanning

Transcription factors along with other DNA binding proteins have the ability to bind their specific target sequences even though their target sequences are likely to be at a low concentration and surrounded by other non–target sequences. Furthermore, they tend to do this at a rate that is faster than what can be accounted for by diffusional collisions (Berg *et al.* 1981). There are proposed mechanisms as to how proteins achieve this and they all generally involve two steps where first they bind non–specifically to DNA and then they undergo translocation to their target sequence where they bind with specificity and greater affinity (Gerland *et al.* 2002; Von Hippel & Berg 1989).

It has been proposed that proteins use a number of mechanisms to scan the DNA, going from non–specific binding of DNA to specific binding of their target sequence. Firstly, the protein might bind the DNA non–specifically and then simply slide along the DNA until the target sequence is found (**Figure 6A**) (Winter *et al.* 1981). Secondly, a protein, after binding non–specifically to DNA, might hop along the DNA (**Figure 6B**). Usually the protein would re–associate a little further along the same DNA segment, however, it could also hop to an entirely new DNA molecule (Berg 1978; Gowers & Halford 2003). A third mechanism is termed intersegmental transfer where a protein will transfer to a new or looped DNA strand by binding both strands simultaneously (**Figure 6C**) (Bellomy & Record 1990).

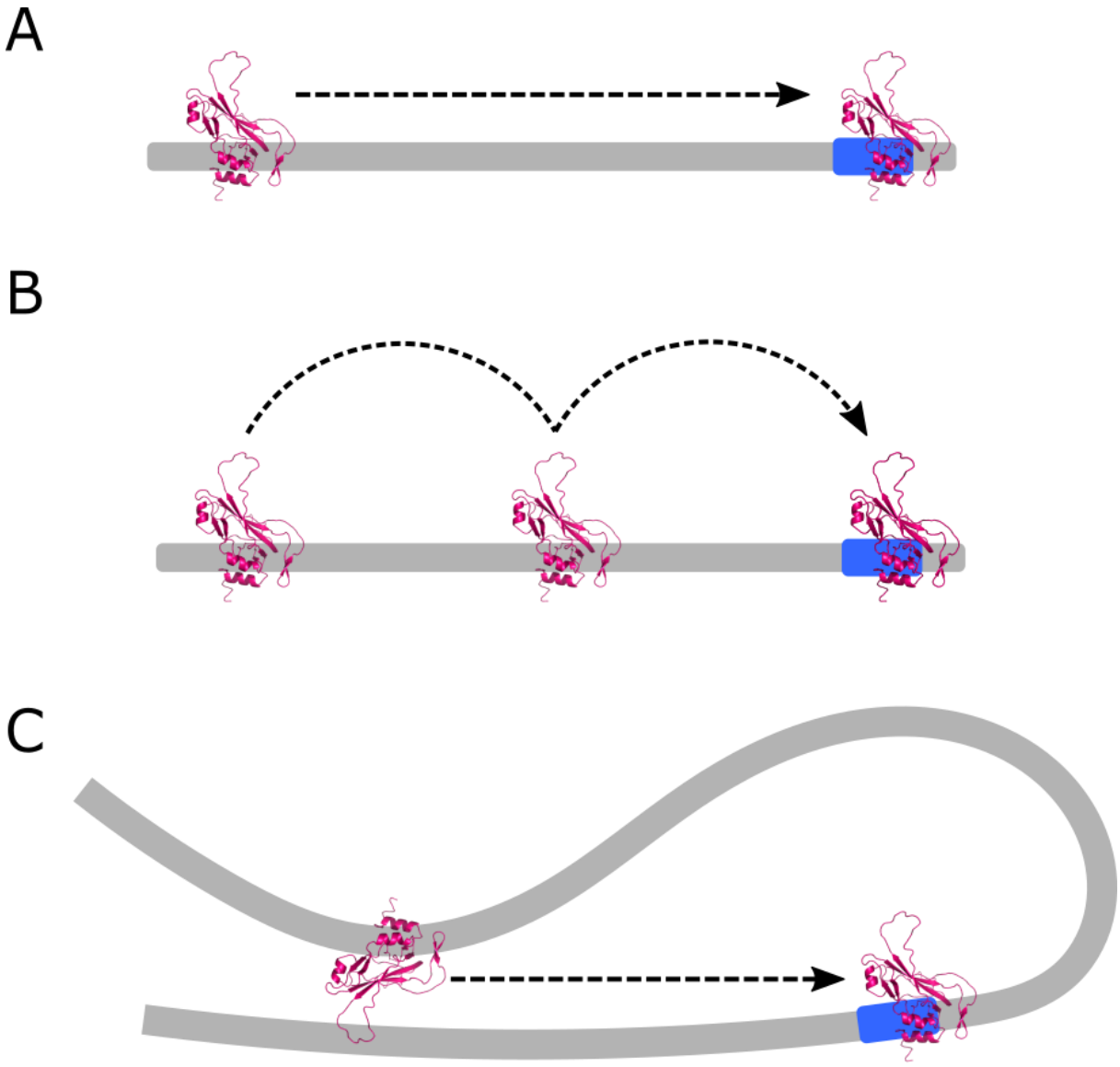


Figure 6: Mechanisms of DNA scanning

Once the protein has bound DNA non–specifically it must scan the DNA and bind its target sequence specifically. This is achieved using 3 mechanisms. **A:** The protein simply slides along the DNA until it finds its target sequence. **B:** The protein hops along the DNA, occasionally hopping to a new DNA molecule, until it finds its target sequence. **C:** The protein may transfer itself from one DNA molecule/ regions to another by binding both DNA segments simultaneously. It is using a combination of these mechanisms along with diffusion that allows a protein to find its target sequence rapidly. The protein structure used was 1XBR(Muller & Herrmann 1997). Image rendered with PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC. Image created using Inkscape 0.92.4.

1.4.2. Specific DNA Binding

There are many structurally diverse transcription factors that all share the ability to bind DNA. The interactions that occur between the amino acid side chains and the bases of the DNA are known as direct or base readout (**Figure 7**). Hydrogen bonds can occur between the protein and the DNA major or minor groove. The specificity between the protein and DNA conferred by hydrogen bonds is not just dependent on the number of bonds that occur, but also the geometry. For example, when two hydrogen bonds occur between different donor and acceptor atoms (bidentate hydrogen bonds), this confers greater specificity than when two hydrogen bonds share the same donor (bifurcated hydrogen bond)(Coulocheri *et al.* 2007). In addition to direct hydrogen bonds, water mediated hydrogen bonds also play a role in base readout as seen with the Tryptophan repressor (Harrison & Aggarwal 1990; Luisi *et al.* 1991). Along with hydrogen bond formation, hydrophobic contacts are the other main driving force of direct readout. Although hydrophobic interactions occur in both the major and minor groove, they tend to be more prevalent in the minor groove. The extensive minor groove contacts often result in the widening of the minor groove. This facilitates extensive hydrophobic contact formation with non–polar side chains and the DNA bases (Bewley *et al.* 1998; Kim *et al.* 1993a, 1993b).

Initially it was thought that base readout was the only factor in binding specificity (Seeman *et al.* 1976). However, the specific interactions between bases and amino acids are not sufficient to fully explain the specificity of DNA binding by transcription factors (Otwinowski *et al.* 1988). In addition to the role that base/direct readout plays, the complimentary shape of the protein and DNA are also important. This is known as indirect or shape readout (**Figure 7**). An important aspect for shape readout is the conformational change that occur to both the protein and DNA during DNA binding that results in a stabilised complex (Rohs *et al.* 2010). In many cases there are conformational deviations from the typical B–DNA shape upon DNA binding. These changes in conformation allow the stabilisation of the complex by promoting the formation of additional Van der Waals contacts, salt bridges and charge–charge interactions between the phosphate backbone and basic amino acid residue side chains (Luscombe *et al.* 2001; Rohs *et al.* 2010). The DNA conformational changes that commonly occur include DNA bending, base flipping, and the formation of kinks (Coulocheri *et al.* 2007; Koudelka *et al.* 2006; Rohs *et al.* 2010).

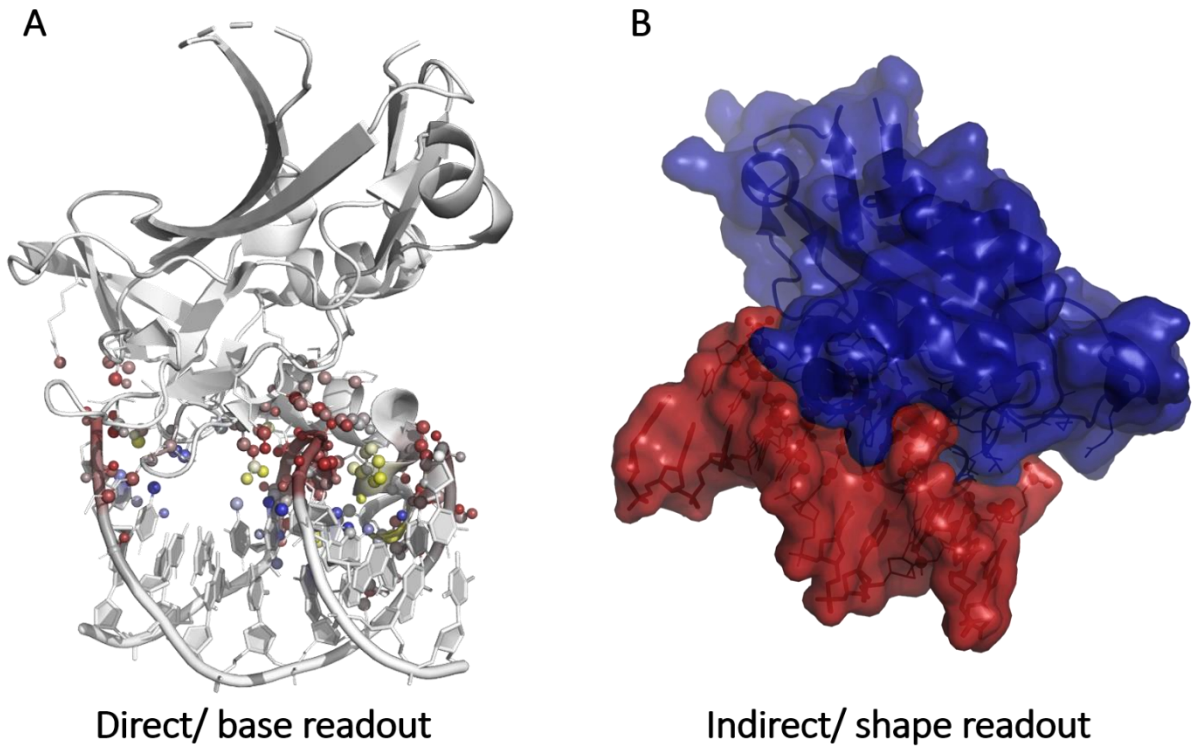


Figure 7: Base and shape readout

A: Base readout is displayed by highlighting atoms that are involved in interactions between the protein and DNA bases (blue), backbone (red) and both bases and backbone (yellow). **B:** Shape readout is displayed by showing the DNA (red) and protein (blue) surfaces. The protein fits perfectly into the DNA minor and Major grooves. The protein and DNA structure used was 2X6V(Stirnimann *et al.* 2010). Image rendered with PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC. The interactions were displayed using PDIviz 1.2.1. (Ribeiro *et al.* 2015).

1.5. Protein–Protein Interactions

Similar to the protein folding problem, understanding the basis for and predicting how two proteins interact is difficult simply because of how diverse protein–protein interactions can be (Wodak & Janin 2004). Protein–protein interactions can be classified by the types of interaction, how structured the interacting proteins are and the strength of the interaction.

The nature of the bonds between the proteins can either be covalent or non–covalent. Covalent protein–protein interactions are rare and generally are disulphide bonds except for the attachment of ubiquitin and small ubiquitin–like modifier. Non–covalent interactions are far more common and are mediated by hydrogen bonds, ionic interactions, Van der Waals interactions and hydrophobic interactions. Non–covalent interactions are generally weaker and therefore, transient. These interactions allow dissociation of complexes which is critical for many regulatory pathways (Westermarck *et al.* 2013).

Protein–protein interactions can be classified in two groups namely domain–domain interactions and domain–peptide interactions. Domain–domain interactions encompass protein interactions between two fully and independently folded domains (**Figure 8**). Classically, domain–domain interactions are complimentary in shape and charge. The interfaces are generally well packed and contain regions where solvent is excluded in the centre surrounded by solvent–mediated interactions (Wodak & Janin 2004). Domain–peptide interactions consist of interactions between one fully folded domain and an unstructured usually small partner. Sometimes the unstructured domain folds to some degree when the interaction forms (**Figure 8**). Transcription factors are generally modular multi–domain proteins with regions capable of binding many proteins and DNA. Many are intrinsically disordered and fold to some degree in the presence of their binding partners (Wodak & Janin 2004). This means that a large modular multi–domain transcription factor might form interactions with proteins via both fully ordered domains and with disordered regions.

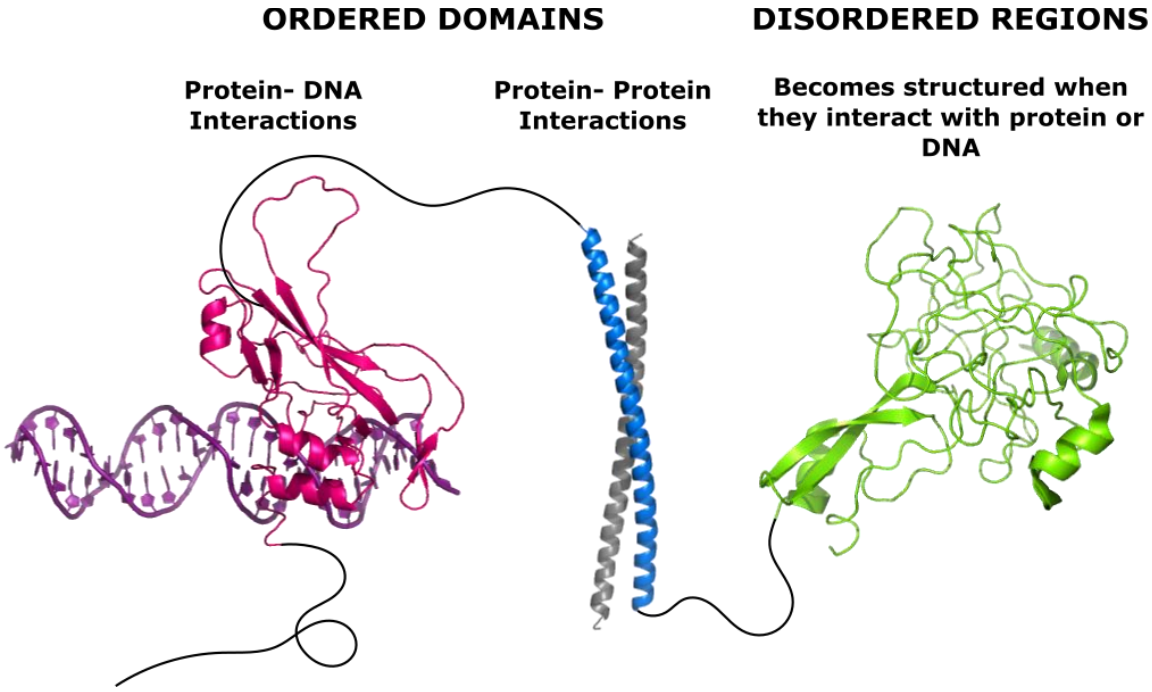


Figure 8: A theoretical example of a multidomain transcription factor containing ordered and disordered domains to facilitate DNA binding and protein– protein interactions.

Transcription factors often contain multiple domains or regions. Some domains are ordered/structured and often interact with DNA (pink) or proteins (blue). These protein–protein interactions can be homo– or heterodimerisation events as displayed above. Some regions are disordered and may fold when an interaction occurs (green). Many transcription factors are intrinsically disordered proteins which allows them to have diverse binding partners. The protein and DNA structures were image rendered with PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC. The structures used were T domain from *Xenopus Laevis* bound to DNA (1XBR) (Muller & Herrmann 1997), MDV1 coiled coil domain (1XU6)(Chattopadhyay *et al.*

2005) and the disordered activation domain in Trypsinogen (2TPI)(Walter *et al.* 1982). Image created using Inkscape 0.92.4.

Many transcription factors regulate multiple genes, governed by complex combinations of interactions. These interactions are often a result of multi–complex protein–protein interactions which are cooperative and form at promoters (Naiche *et al.* 2005). Different combinations of transcription factor interactions either activate or repress expression.

Expression of genes is a highly regulated process. It is critical that the correct genes be expressed to the appropriate levels at any given time. This changes constantly due to various signals as well as in different cell and tissue types. Mis–regulation of gene expression can lead to disease which is why it is important to not only understand the regulatory pathways that transcription factors regulate but also the interactions and binding events that occur along each step (Lee & Young 2013).

1.6. Classification of Transcription Factors

It is estimated that there are at least 2000 transcription factor encoding genes in the human genome (Vaquerizas *et al.* 2009; Venter *et al.* 2001). These transcription factors are classified based on their DNA binding domains (DBD)

According to the TFclass classification system, mammalian transcription factors are grouped into 9 superclasses based on the general topology of their main DBD and the way they interact with DNA (Wingender *et al.* 2015). A class is defined by the way the DBD folds and interacts with DNA through interfaces and oligomerisation events. The classes are divided into families and subfamilies which are based on sequence similarity and function (Wingender 2013).

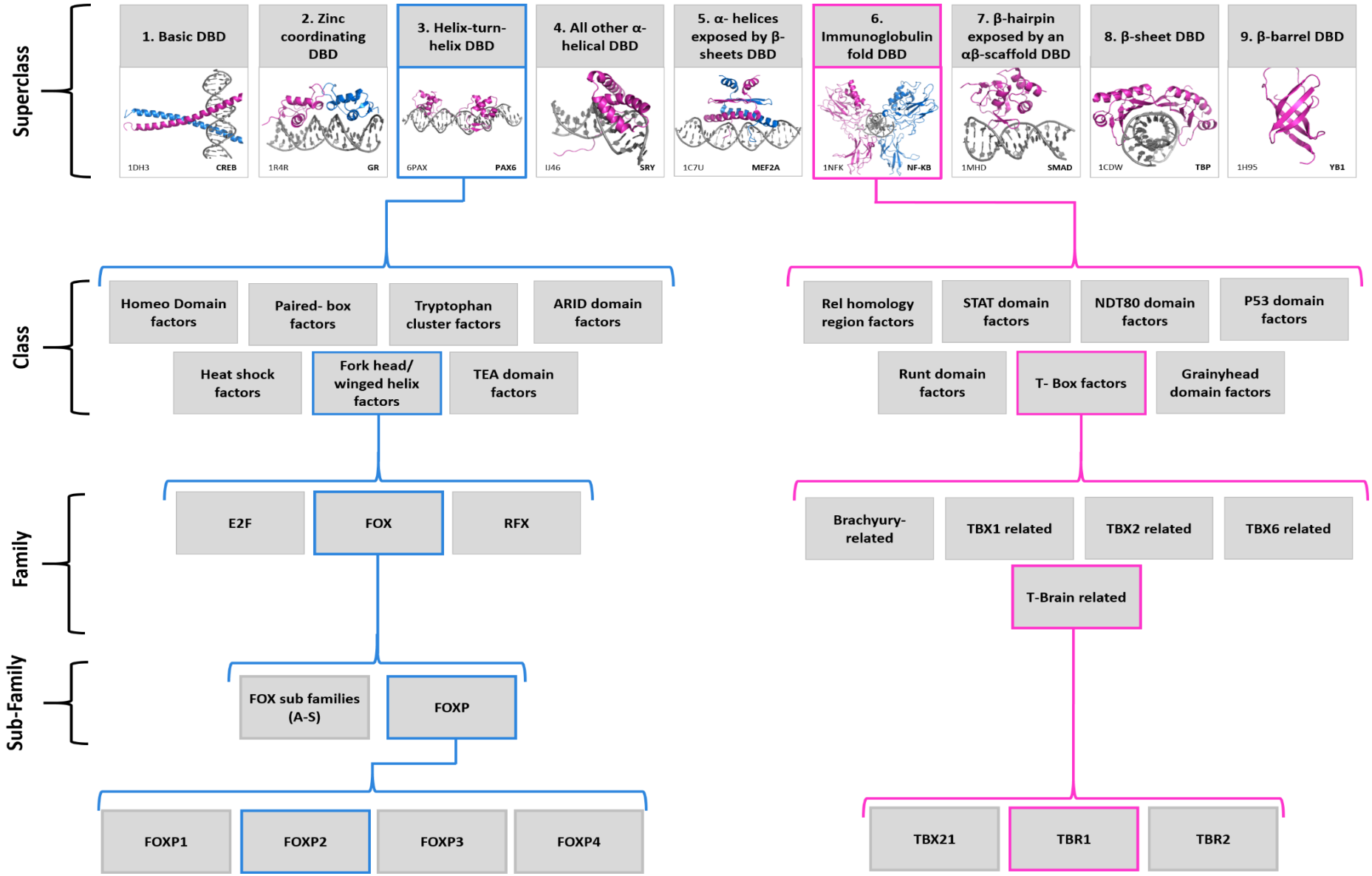

Superclass
1. Basic DBD
1DH3 CREB
2. Zinc coordinating DBD
1R4R GR
3. Helix-turn-helix DBD
6PAX PAX6
4. All other α-helical DBD
IJ46 SRY
5. α- helices exposed by β-sheets DBD
1C7U MEF2A
6. Immunoglobulin fold DBD
1NFK NF-KB
7. β-hairpin exposed by an αβ-scaffold DBD
1MHD SMAD
8. β-sheet DBD
1CDW TBP
9. β-barrel DBD
1H95 YB1
Class
Homeo Domain factors
Paired- box factors
Tryptophan cluster factors
ARID domain factors
Heat shock factors
Fork head/ winged helix factors
TEA domain factors
Rel homology region factors
STAT domain factors
NDT80 domain factors
P53 domain factors
Runt domain factors
T- Box factors
Grainyhead domain factors
Family
E2F
FOX
RFX
Brachyury-related
TBX1 related
TBX2 related
TBX6 related
T-Brain related
Sub-Family
FOX sub families (A-S)
FOXP
FOXP1
FOXP2
FOXP3
FOXP4
TBX21
TBR1
TBR2

Figure 9: A structural classification of mammalian transcription factors based on their DNA binding domains displaying TBR1 and FOXP2.

According to TFclass classification scheme there are 9 superclasses of transcription factors. A superclass defines the general topology of the DNA binding domain. A class defines the structural blueprint of the DNA binding domain. A family is defined by function and sequence similarity. A subfamily is a group based on further sequence similarity. The genus is the transcription factor gene. The TBR1 (T–Brain protein 1) classification is indicated in pink whereas the FOXP2 (Forkhead Box P2) classification is indicated in blue. A structure of a member of each superclass is shown along with the PDB ID and abbreviated name (Wingender *et al.* 2015). The structure images are rendered with PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC

The helix–turn–helix (HTH) domains superfamily contains more than a quarter of human transcription factors (Wingender 2013). These transcription factors contain a recognition helix that forms specific interactions with the DNA. The recognition helix is surrounded by two or three additional helices which in many cases lie almost perpendicular to one another (Burley 1997).

The classical HTH motif has three alpha helices arranged in a triangle (**Figure 10**). The turn between the second and third helices is well conserved and sharp. The interaction with DNA occurs via an interaction between helix 3 and the major groove of DNA (Warren 2002). The winged helix–turn–helix (W–HTH) motif, the subclass FOXP proteins fall under, is a variation of the HTH. It differs by the presence of a β–strand hairpin wing at the C–terminal end. The W–HTH interacts with DNA by insertion of helix 3 into the DNA major groove like the classical HTH and it often makes additional DNA contacts, facilitated by the wing, with the minor groove (Aravind *et al.* 2005).

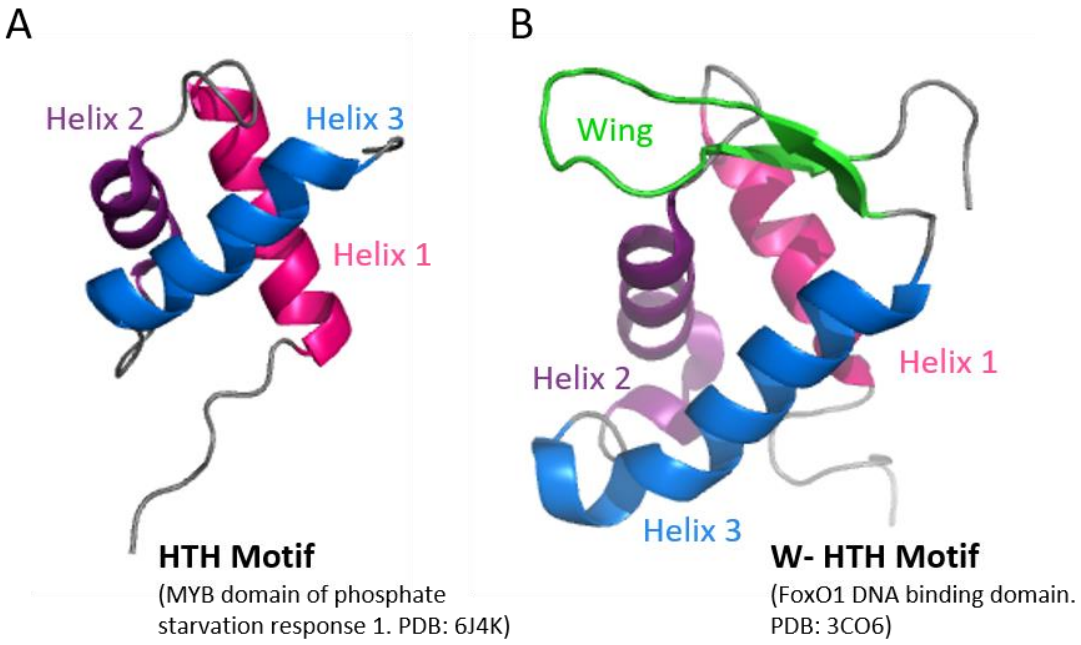


Figure 10 A: A helix– turn– helix motif B: A winged helix–turn–helix motif.

A: The MYB domain of phosphate starvation response 1 is an example of a classical helix– turn– helix motif (PDB ID: 6J4K)(Jiang *et al.* 2019). Helices 1, 2 and 3 are coloured in pink, purple and blue. **B**: The FoxO1 DNA binding domain is a forkhead domain which is a good example of a winged helix–turn–helix motif (PDB ID: 3CO6)(Brent *et al.* 2008). Helices 1, 2 and 3 are coloured in pink, purple and blue and the wing is coloured in green. Images rendered with PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.

Many DBDs rely on a β–sheet motif in order to function. Transcription factors in the β–sheet DBD superfamily like the TATA Box binding protein (TBP) interact with the DNA minor groove via β–strands (Nikolov *et al.* 1996; Wingender 2013). The β–sheet structure can further form distinct morphologies, one of which is the β–barrel. A β–barrel is simply a twisted β–sheet, often antiparallel, which folds to form either a closed or rounded structure (**Figure 11**). Members of the β–barrel DBD superclass interact with single stranded DNA via side chains facing outward from the β–barrel and via loops protruding from the β–strands making up the barrel (Kloks *et al.* 2002). Many proteins do not fit the strict definition of a β–barrel and are in fact combinations between distinct motifs, such as proteins with the immunoglobulin like fold motif (Alberts *et al.* 2002). The immunoglobulin–like fold superclass consists of proteins characterised by some form of β–structured core, often a β–barrel and a β–sheet lid, with exposed coils and helices which form the interactions with the DNA minor and major grooves (Wingender 2013).

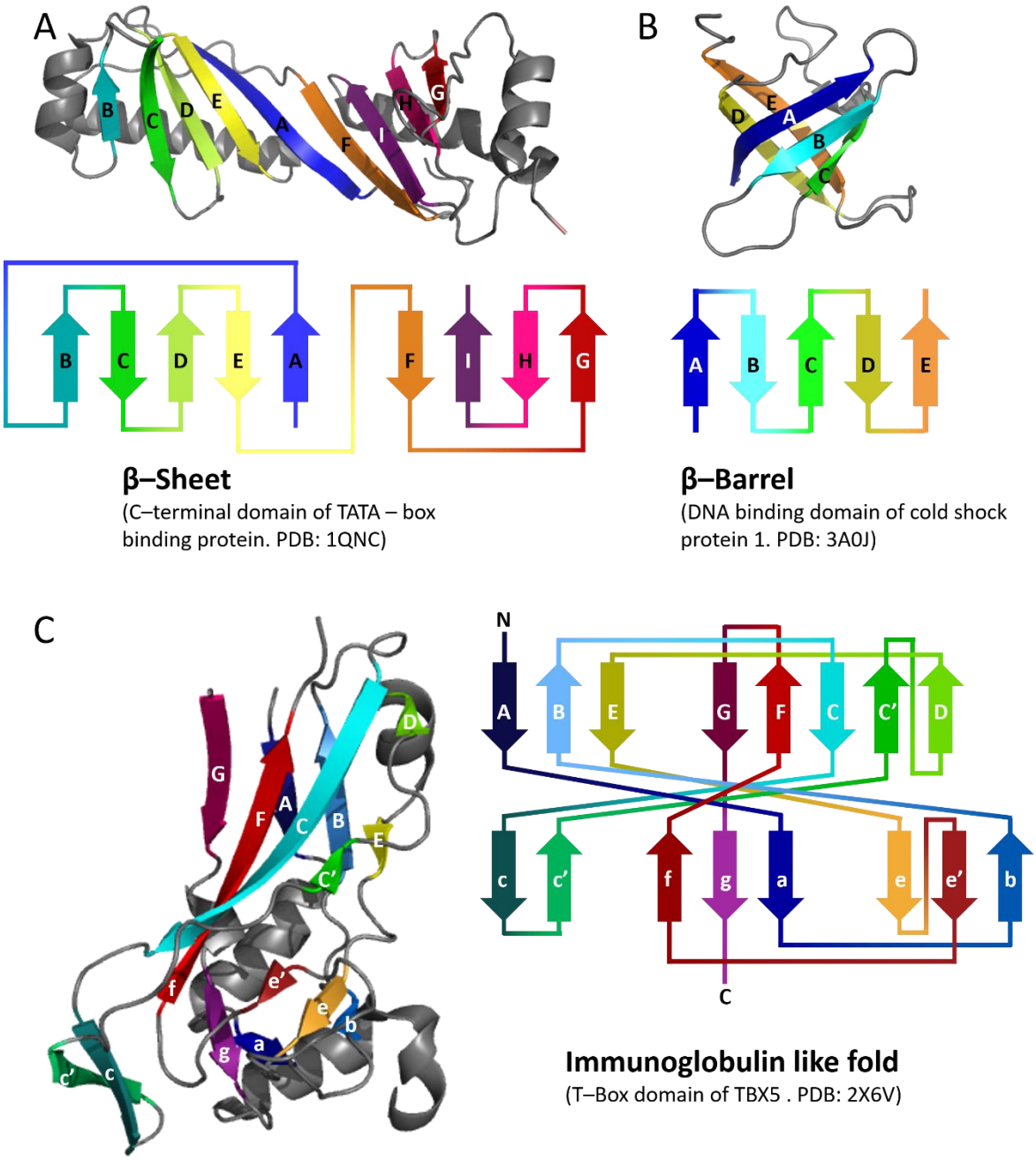


Figure 11: β–Sheet, β–Barrel and immunoglobulin like fold protein structures and topology diagrams.

A: A typical β–sheet motif (C–terminal domain of TATA–box binding protein, PDB ID 1QNC)(Patikoglou *et al.* 1999) and a schematic showing the direction and order of the β–strands that form the β–sheet motif. **B:** A typical β–barrel motif (DNA binding domain of cold shock protein 1, PDB ID 3A0J) and a schematic showing the direction and order of the β–strands in the β–barrel. **C**: A typical immunoglobulin like fold (T–Box domain of TBX5, PDB ID 2X6V) (Stirnimann *et al.* 2010). A β–barrel motif is formed at the top of the structure (capital letter named strands) and a lid like structure forms at the bottom (lower case named strands). Image rendered with PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.

1.7. FOX Transcription Factors

The forkhead box (FOX) family of transcription factors is one of three families in the winged helix class of mammalian transcription factors. FOX proteins all contain a highly conserved DNA binding domain known as the forkhead domain (FHD). Despite the conservation of the FHD, FOX transcription factors are functionally diverse. Some have been implicated in numerous processes including embryonic development in multiple tissues (Shu *et al.* 2007), immune regulation (Koh *et al.* 2009), metabolism (Friedman & Kaestner 2006) and cancer (Sharma *et al.* 2005).

Figure 12, shows the first solved crystal structure of a FOX FHD: that of the FOXA3 FHD bound to DNA (Clark *et al.* 1993). This structure is considered to be a generic FOX FHD model. A FOX FHD is a W–HTH motif which consists of 90 to 100 residues. (Clark *et al.* 1993; Friedman & Kaestner 2006) The wing regions are the least conserved amongst the FOX FHDs. The main interactions between FOX FHDs and DNA occur via insertion of helix three into the major groove of DNA. Helix 3 is highly conserved in all FOX proteins and thus it is likely that other regions of the FHD confer DNA binding specificity (Clark *et al.* 1993; Friedman & Kaestner 2006)

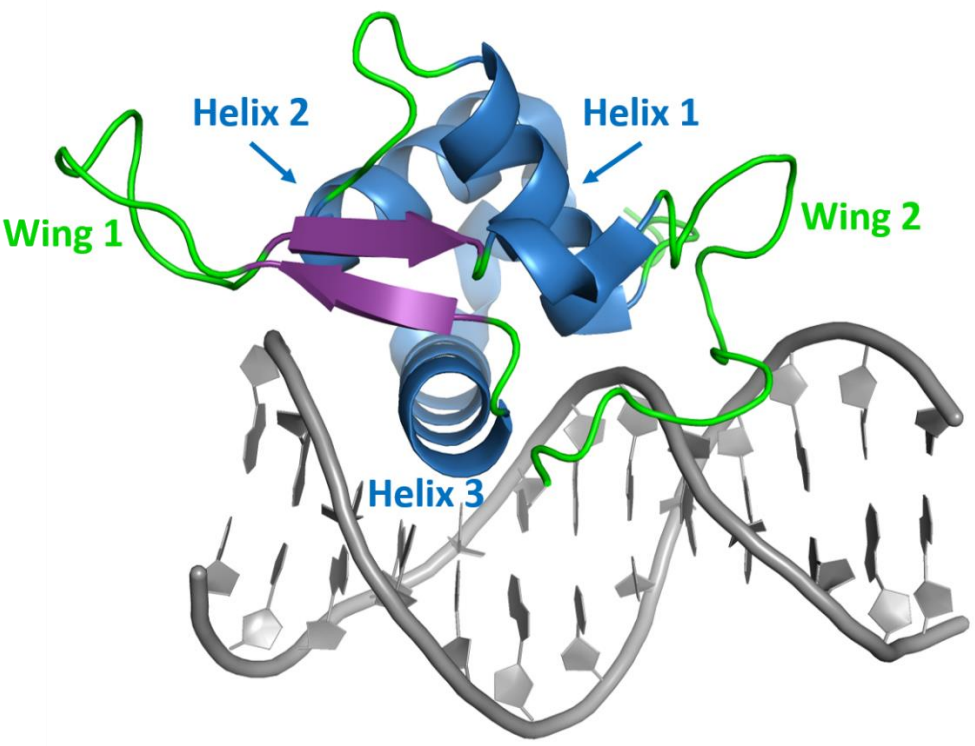


Figure 12: FOXA3 forkhead domain bound to DNA.

Helix 3 is wedged into the major groove of the DNA. Helices are blue, the coils are green, and the β–sheet is purple. Wing 1 is located between the β–strands and wing 2 is located on the C–terminal (PDB ID: 1VTN)(Clark *et al.* 1993). Image of rendered with the PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.

1.7.1. FOXP SUBFAMILY

The FOX family of transcription factors consists of 19 subfamilies, labelled A–S, which differ based on their structural characteristics. There are four members within the FOXP subfamily which are distinguished from other FOX subfamilies as they contain a glutamine–rich region, a zinc finger and a leucine zipper (**Figure 13**) (Wang *et al.* 2003). Their FHD is also located near the C–terminus rather than the N–terminus of the full length protein (Shu *et al.* 2001).

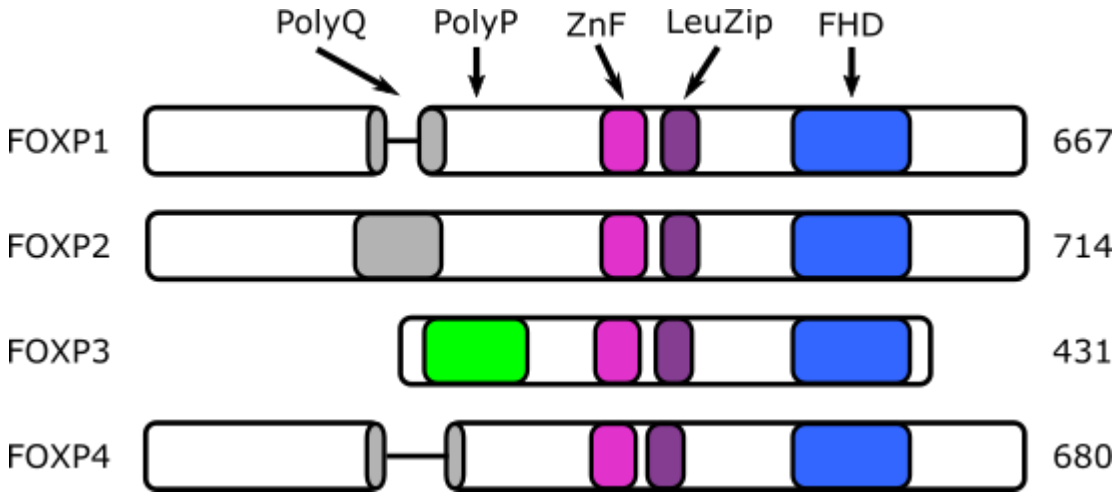


Figure 13: The domain architecture of the four members of the FOXP subfamily.

The FHD (blue) is found near the C–terminal end of each member. FOXP1, FOXP2 and FOXP4 all contain N–terminal polyglutamine rich regions (grey) whereas FOXP3 contains a proline rich region (green). They each also contain a zinc finger (pink) and a leucine zipper (purple).

The multiple domains in FOXPs are typical of the modular structure of transcription factors. Glutamine rich regions in transcription factors generally have transcriptional repression/ activation functions which can be mediated by protein–protein interactions (Shu *et al.* 2001; Wang *et al.* 2003). The leucine zipper and zinc finger motifs have been shown to function in the formation of homo and heterodimers of the full length FOXP proteins (Wang *et al.* 2003). Furthermore, studies were conducted on dimerisation of FOXP1, FOXP2 and FOXP4 found using a reporter assay. This study demonstrated that their ability to repress transcription diminished when the leucine zipper was removed preventing dimerisation of the full length proteins (Li *et al.* 2004).

1.7.2. FOXP FUNCTION

The four members of the FOXP subfamily display diverse functions despite the highly conserved FHD, leucine zipper and zinc finger. FOXP1 has been shown to be involved in tumour suppression (Banham *et al.* 2001), lung development (Shu *et al.* 2007), regulation of B cell development (Hu *et al.* 2006) and brain development (Ferland *et al.* 2003). FOXP3, one of the more studied members, has been implicated in regulation of

the immune system (Koh *et al.* 2009) and dysregulation of FOXP3 results in immune dysregulation and polyendocrinopathy, enteropathy X–linked inheritance(Bennett *et al.* 2001). FOXP4, the least studied member, has been associated with prostate and kidney cancer (Long *et al.* 2012; Teufel *et al.* 2003).

FOXP2 is expressed in intestinal, neural and cardiovascular tissues during embryonic development (Shu *et al.* 2001). The importance of FOXP2 was first recognised when the R553H mutation in its forkhead domain was discovered. The mutation was discovered in a family, known as the KE family, suffering from a speech disorder manifesting as the inability to sequence mouth movements resulting in difficulty with pronunciation (verbal dyspraxia). In addition to verbal dyspraxia, members of the KE family, despite displaying close to normal nonverbal IQ, also display defects of written language and receptive linguistic skills (Fisher & Scharff 2009). In addition to this, FOXP2 has been implicated in autism (Gong *et al.* 2004; Mukame *et al.* 2011), schizophrenia (Sanjuán *et al.* 2006; Španiel *et al.* 2011; Tolosa *et al.* 2010) and dyslexia (Wilcke *et al.* 2012). Furthermore, down regulation of FOXP2 has been associated with a variety of cancers (reviewed by Herrero & Gitton 2018) including hepatocellular carcinoma (Yan *et al.* 2015), gastric cancer (Jia *et al.* 2016) and breast cancer (Chen *et al.* 2018; Cuiffo *et al.* 2014).

1.7.3. FOXP STRUCTURE

FOXP proteins, like other FOX proteins, bind DNA by inserting helix 3 into the DNA major groove. However, unlike other FOX proteins, FOXPs have the ability to form a domain swapped dimer via the FHD in addition to dimerisation of the full length proteins (mediated by the leucine zippers)(Häußermann *et al.* 2019; Li *et al.* 2004; Wang *et al.* 2003). Studies in mammalian cells have shown that different combinations of FOXP1, FOXP2 and FOXP4 homo and heterodimers are capable of differentially regulating multiple FOXP2 gene targets (Mendoza & Scharff 2017).

When domain swapping occurs in FOXP2, FOXP3 and FOXP1 (nothing is known about FOXP4's ability to form domain swapped dimers), helix 3, strand 2 and strand 3 are swapped by two FOXP2 monomers. Domain swapping of the FHD is made possible due to the extension of the turn connecting helix 2 and helix 3 (Bandukwala *et al.* 2011; Chu *et al.* 2011; Stroud *et al.* 2006). There are multiple residues that interact along the hydrophobic dimer interface of the FOXP2 domain swapped dimer which include Tyr509, Tyr531, Tyr540, Trp533, Trp548, Phe507, Phe534, Phe538 and Phe541 from both monomers (Stroud *et al.* 2006). In **Figure *14*** the domain swapped dimer and monomer of the FOXP2 FHD are displayed (PDB ID: 2A07). The main difference between the dimer and the monomer is that in the dimer, helix 2 is extended to form the hinge region.

The FOXP subfamily have the only FOX FHDs known to be capable of forming a domain swapped dimer. Alanine 539 in FOXP2 and the equivalent alanine in FOXP3 and FOXP1

allows the flexibility in the hinge loop region required to domain swap. This is not possible in other FHDs that contain a proline in that position (Stroud *et al.* 2006). The mutation of Ala539 to a proline in FOXP2 eliminates domain swapping entirely and results in a completely monomeric FHD (Stroud *et al.* 2006). The wild type FOXP2 FHD, however, can bind DNA with a greater affinity and has more conformational flexibility than the A539P mutant (Morris & Fanucchi 2016), implicating the significance of domain swapping in the interaction of this protein with DNA.

The FOXP3 FHD exists almost solely in dimeric form (Bandukwala *et al.* 2011). The FOXP2 FHD, on the other hand, appears to exist as a mixture of monomeric and dimeric species (Stroud *et al.* 2006). In both cases the dimer interface is hydrophobic but the FOXP3 domain swapped dimer is more stable than the FOXP2 domain swapped dimer (Bandukwala *et al.* 2011; Stroud *et al.* 2006). Perumal *et al*, 2015 showed that the FOXP2 FHD domain swapped dimer can be stabilised by mutation of tyrosine 540 (in the dimer interface) to a phenylalanine as in FOXP3. This result suggests that the stability of the dimer interface is dependent on its hydrophobicity.

It has been postulated that the FOXP2 domain swapped dimer might simultaneously bind two separate DNA strands due to the proximity of the two DNA binding surfaces of the domain swapped dimer. It has been postulated that the FOXP2 domain swapped dimer may loop DNA or even mediate interchromosomal contacts in this way (Stroud *et al.* 2006).

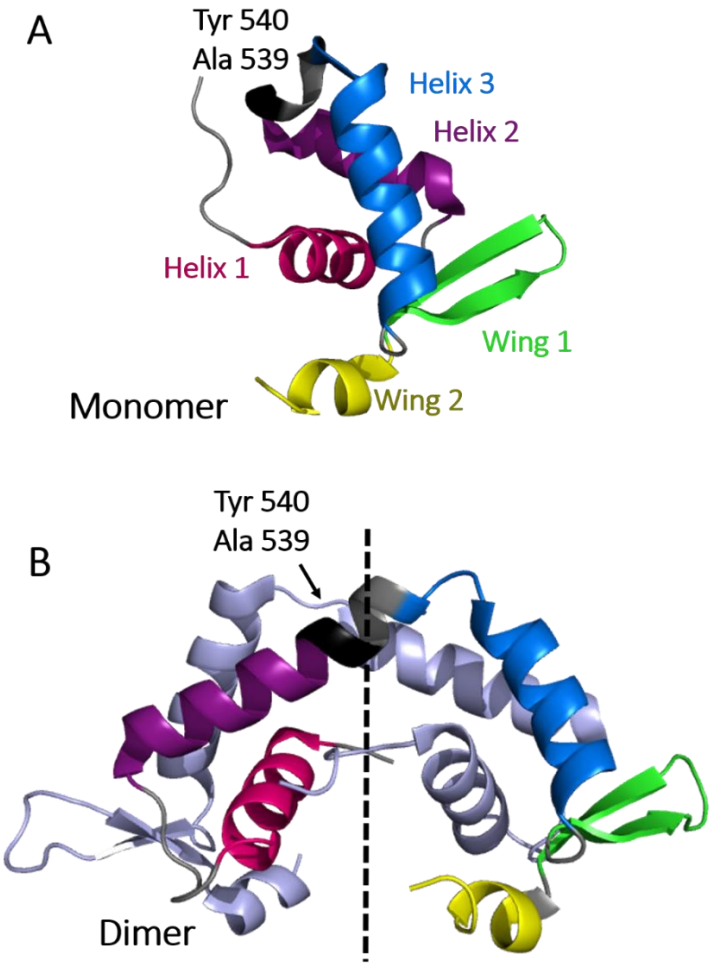


Figure 14: The FOXP2 FHD domain monomer (A) and domain swapped dimer (B).

A: The monomeric FHD is coloured according to the W–HTH regions (Helix 1, Helix 2, Helix 3, Wing 1, Wing 2). Alanine 539 and Tyrosine 540 are coloured in Black. **B:** One monomer of the domain swapped dimer is displayed in light blue while is the other is coloured according to the W–HTH regions (Helix 1, Helix 2, Helix 3, Wing 1, Wing 2) observed in the monomer. The dimer is formed by the interchange of helix 3 (blue), wing 1 (green) and wing 2 (yellow). Alanine 539 and Tyrosine 540, amino acids significant in dimerisation, are coloured in Black (PDB ID: 2A07)(Stroud *et al.* 2006). Image rendered with PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC

1.8. FOXP2

1.8.1. FOXP2 DNA BINDING

The FOXP2 FHD has been shown to bind DNA in both the dimeric and monomeric form via helix 3(Stroud *et al.* 2006). The main DNA interactions formed by the monomeric FHD, as determined from the crystal structure (PDB ID: 2A07) occur through Asn550, Arg553 and His554. Asn550 and His554 from direct hydrogen bonds with the DNA while Arg553 forms a water mediated hydrogen bond with the DNA bases. Hydrogen bonds to the DNA phosphate backbone are formed by Ser557, Thr547, Arg504, Trp573, Tyr509,

Arg583 and Arg584. In addition to these hydrogen bonds, various residues also form extensive Van der Waals interactions with the DNA. Certain aromatic and hydrophobic residues from helix 1 and helix 3 interact with the DNA, stabilising the DNA protein interaction (Stroud *et al.* 2006).

As the forkhead domain of FOX proteins is so highly conserved, it is predicted that FOX proteins have similar DNA sequences to which they will bind. The general FOX consensus sequence is 5' [A/G][C/T][A/C]AA[C/T]A 3' based on the conserved nature of residues that make DNA contacts in helix 3 (Carlsson & Mahlapuu 2002). Regions in the FHD such as wing 1 and wing 2 as well as helix 3 confer DNA binding specificity (Clark *et al.* 1993). FOX proteins including FOXA3 utilise the wing 1 and wing 2 regions to form extensive contacts with the DNA minor groove and backbone (Clark *et al.* 1993; Stroud *et al.* 2006). As the FOXP2 forkhead domain has shorter wing 1 and wing 2 regions than most other FHDs, there are limited DNA contacts. It has been postulated that this could result in a decrease in binding affinity of FOXP2 compared to some other FOX proteins (Stroud *et al.* 2006). Furthermore, the flexibility which facilitates domain swapping, also allows for the formation of more electrostatic contacts with the DNA (Morris *et al.* 2018). According to the crystal structure, FOXP2 utilises many Van der Waals forces and fewer hydrogen bonds to make contacts in the major groove compared to other FOX proteins, which could allow the FOXP2 FHD more flexibility in terms of DNA binding sequences (Stroud *et al.* 2006). A study determined that the FOXP2 FHD has the ability to bind multiple DNA sequences with various rates and affinities (Webb *et al.* 2017). The two most commonly referred to sequences are the Wang sequence (5' A[C/T]AAATA 3') (Wang *et al.* 2003) and the Nelson sequence (5' TGTTTAC 3') (Nelson *et al.* 2013). When the FOXP2 FHD binds the Nelson sequence with high affinity it becomes more structured and rigid than when it binds the Wang sequence with lower affinity. This suggests that lower affinity binding may be a mechanism of DNA scanning in a partially unfolded state (Webb *et al.* 2017). It is also possible that lower affinity DNA binding may be enhanced by protein–protein interactions involving the full length FOXP2 protein.

1.8.2. FOXP2–Protein Interactions

As previously stated, the formation of protein–protein interactions between different transcription factors is important because formation of large multi macromolecular complexes is required for transcriptional regulation. FOXP2, along with FOXP1 and FOXP4, has the ability to form homo and hetero–oligomers which are mediated via their leucine zippers (Li *et al.* 2004). These three proteins are expressed specifically in various organs and cell types and in some cases their expression patterns overlap. For instance, FOXP1, FOXP2 and FOXP4 are all expressed in the brain, lung and gut (Shu *et al.* 2001). A study by Sin *et al.* (2014) demonstrated that the various combinations of FOXP1, FOXP2 and FOXP4 homo and hetero–oligomers differentially regulated 10 genes (CER1, SFRP4, WISP2, PICKLE1, NCOR2, SNWI, NEUROD2, PAX3, EFNB3 and SLIT1) known to be regulated by FOXP2 alone. Furthermore, (Mendoza & Scharff 2017), show that the

FOXP1/FOXP2 dimer binds the Contactin–associated protein–like 2 (CNTNAP2) gene whereas the FOXP2/FOXP4 dimer does not.

FOXP2 also forms protein–protein interactions with non–FOXP proteins (**Table 1**). For instance, FOXP2 interacts with CTBP1 (C–terminal–binding protein 1), a protein known to act as a co–repressor for FOG1, FOG2 and class II histone deacetylases (Li *et al.* 2004). FOXP2 has also been shown to interact with Transcription factor SOX–5 and Zinc finger MYM–type protein 2 via a region within amino acids 93 to 258 (Estruch *et al.* 2018). Stabilin–1 and Stabilin–2 have been found to interact with not only FOXP2 but also various other FOX proteins suggesting they interact with the FHD which is common to all (Estruch *et al.* 2018; Li *et al.* 2015). COUP transcription factor 1 (NR2F1) and COUP transcription factor 2 (NR2F2), were found to interact with the leucine zipper/zinc finger region. However, their interactions did not appear to be dependent on oligomerisation (Estruch *et al.* 2018). Interestingly, YY1 was found to interact with multiple regions of FOXP2 with apparent differences in affinity (Estruch *et al.* 2018).

Table 1: Interacting partners of FOXP2

	Protein	Interaction type	Experiment	Reference
CCNC	Cyclin–C	Physical association	Two hybrid prey pooling approach, Two hybrid array, Validated two hybrid	(Rolland *et al.* 2014)
CTBP1	C–terminal–binding protein 1	Physical association	Two hybrid prey pooling approach, Two hybrid array, Validated two hybrid	(Corominas *et al.* 2014; Deriziotis *et al.* 2014a)(Estruch *et al.* 2018)
CTBP2	C–terminal–binding protein 2	physical association, co–localisation	Bioluminescence resonance energy transfer	(Deriziotis *et al.* 2014a)
DSG4	Desmoglein–4	Association	Anti–bait coimmunoprecipitation	(Huttlin *et al.* 2017)
EFHD1	EF–hand domain–containing protein D1	Association	Anti–bait coimmunoprecipitation	(Huttlin *et al.* 2017)
FOXP1	Forkhead Box Protein P1	physical association, co–localisation	Bioluminescence resonance energy transfer, Confocal microscopy, Tandem affinity purification, Coimmunoprecipitation	(Deriziotis *et al.* 2014a; Konopka *et al.* 2009; Li *et al.* 2015; Sollis *et al.* 2016)
FOXP2	Forkhead Box Protein P2	Direct interaction, physical association	Molecular Sieving, Bioluminescence resonance energy transfer, Light scattering, Tandem affinity purification	(Deriziotis *et al.* 2014a; Li *et al.* 2015; Stroud *et al.* 2006)
FOXP3	Forkhead Box Protein P3	Physical association	Tandem affinity purification	(Li *et al.* 2015)
FOXP4	Forkhead Box Protein P4	Physical association	Tandem affinity purification, Coimmunoprecipitation	(Konopka *et al.* 2009; Li *et al.* 2015)
INSR	Insulin receptor	Association	Anti–bait coimmunoprecipitation	(Boutchueng–Djidjou *et al.* 2018)
LRRC15	Leucine–rich repeat–containing protein 15	Association	Anti–bait coimmunoprecipitation	(Huttlin *et al.* 2017)
MAPK3	Mitogen–activated protein kinase 3	Physical association	Two hybrid array	(Corominas *et al.* 2014)
MKP6	Mitogen–activated protein kinase phosphatase 6	association	Anti–tag coimmunoprecipitation	(Huttlin *et al.* 2017)
MTO	Methanethiol oxidase	Association	Anti–bait coimmunoprecipitation	(Huttlin *et al.* 2017)
NEK4	Serine/threonine–protein kinase Nek4	association	Anti–bait coimmunoprecipitation	(Basei *et al.* 2015)
NR2F1	COUP transcription factor 1	Physical association	Bioluminescence resonance energy transfer	(Estruch *et al.* 2018)
NR2F2	COUP transcription factor 2	Physical association	Bioluminescence resonance energy transfer	(Estruch *et al.* 2018)
PIN1	Peptidyl–prolyl cis–trans isomerase NIMA–interacting 1	Physical association	Two hybrid prey pooling approach, Two hybrid array, Validated two hybrid	(Rolland *et al.* 2014)

PKP3	Plakophilin–3	Association	Anti–bait coimmunoprecipitation	(Huttlin *et al.* 2017)
RPIA	Ribose–5–phosphate isomerase	Physical association	Two hybrid prey pooling approach, Two hybrid array, Validated two hybrid	(Corominas *et al.* 2014; Rolland *et al.* 2014)
S100A2	Protein S100–A2	association	Anti–bait coimmunoprecipitation	(Huttlin *et al.* 2017)
SDCB1	Syntenin–1	Physical association	Two hybrid prey pooling approach, Two hybrid array, Validated two hybrid	(Rolland *et al.* 2014)
SOX5	Transcription factor SOX–5	Physical association	Bioluminescence resonance energy transfer	(Estruch *et al.* 2018)
SP4	Transcription factor Sp4	Physical association	Two hybrid prey pooling approach, Two hybrid array, Validated two hybrid	(Rolland *et al.* 2014)
STAB1	Stabilin–1	Physical association	Bioluminescence resonance energy transfer	(Estruch *et al.* 2018)
STAB2	Stabilin–2	Physical association	Bioluminescence resonance energy transfer	(Estruch *et al.* 2018)
TBR1	T–Brain Related Protein 1	physical association, co–localisation	Bioluminescence resonance energy transfer, Confocal microscopy	(Deriziotis *et al.* 2014b)
TLE5	TLE family member 5	Physical association	Two hybrid prey pooling approach, Two hybrid array, Validated two hybrid	(Rolland *et al.* 2014)
TNIK	TRAF2 and NCK–interacting protein kinase	Physical association	Anti–bait coimmunoprecipitation	(Li *et al.* 2017)
TSACC	TSSK6–activating co–chaperone protein	Physical association	Two hybrid prey pooling approach, Two hybrid array, Validated two hybrid	(Rolland *et al.* 2014)
VSIG8	V–set and immunoglobulin domain–containing protein 8	Association	Anti–bait coimmunoprecipitation	(Huttlin *et al.* 2017)
YY1	Yin Yang 1	Physical association	Bioluminescence resonance energy transfer	(Estruch *et al.* 2018)
ZMYM2	Zinc finger MYM–type protein 2	Physical association	Bioluminescence resonance energy transfer	(Estruch *et al.* 2018)

Data from the table was curated by the EMBL–EBI IMEx database (Orchard *et al.* 2012) and various papers. A direct interaction is defined as two molecules in direct contact with each other. Physical association is defined as molecules that are in the same complex, they do not necessarily interact directly. Co–localisation is defined as molecules in the same subcellular location, the methodology used is not high resolution enough to confirm an interaction.

1.9. T–Box Transcription Factors

The T–Box class of proteins consists of approximately 50 members, 18 in mammals, all comprising of a structurally conserved fold known as the T–Box domain (Wilson & Conlon 2002). There are 5 families of T–Box proteins in the T–Box factors order: Brachyury related, TBX1 related, TBX2 related, TBX6 related and T–brain related (**Figure 9 and Figure 15**) (Wingender *et al.* 2015). Many T–Box proteins are conserved among vertebrates, however, there are many organisms which contain distinct T–Box proteins. They tend to display very little conservation outside of the T–Box region and are functionally diverse (Kavka & Green 1997; Smith 1997). T–Box transcription factors are expressed in multiple organisms and tissues and appear to be involved in multiple developmental processes ranging from gastrulation to organogenesis (Kavka & Green 1997; Smith 1997).

T–Box proteins are generally between 50– 70 kDa and comprise of two domains. The first domain is the T–Box domain which can form interactions with both DNA and proteins (Herrmann 1995). The second domain is a transactivation or repression domain. The order and location of the domains in T–Box transcription factors differs although it is conserved between each member of the family and its orthologs. All proteins in the T–Box family thus far, are capable of binding a consensus DNA sequence (5' TCACACCT '3) despite fairly large differences in their amino acid sequences (Wilson & Conlon 2002).

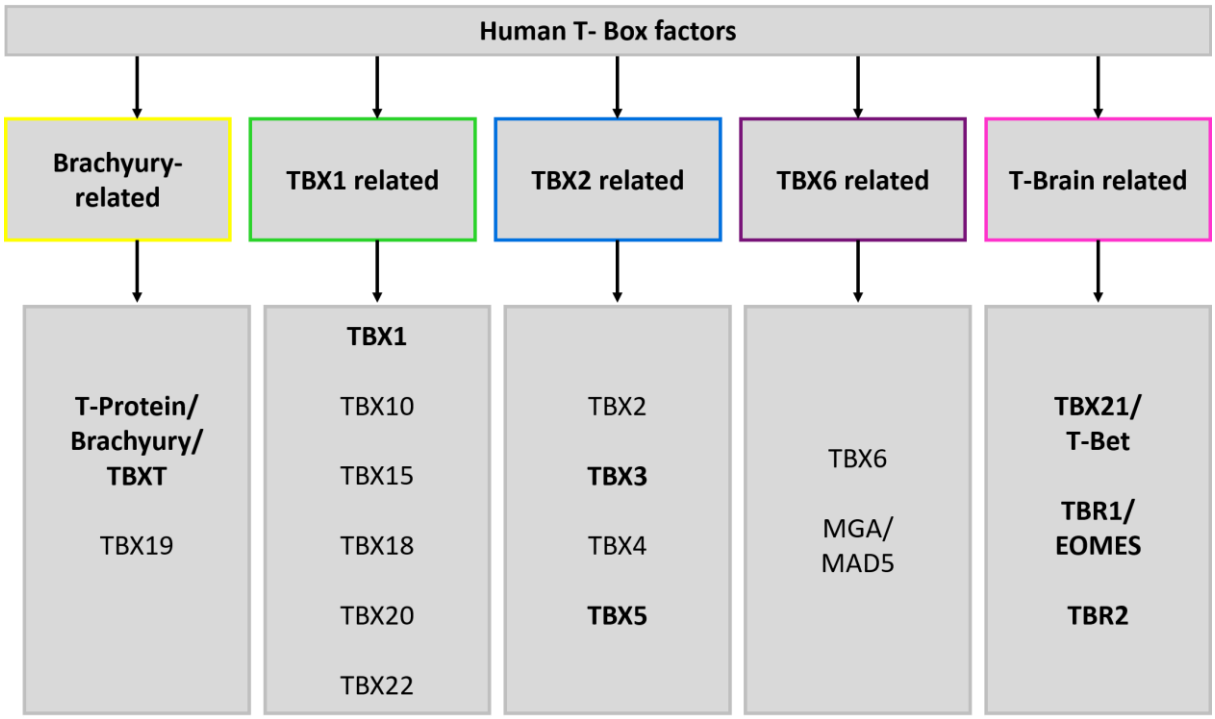


Figure 15: All T–Box proteins found in humans grouped into their families.

The T–Box proteins found in humans grouped according to the families using the TFclass classification scheme (Wingender *et al.* 2015). T–Box proteins mentioned within this document are highlighted in bold.

1.9.1. T–Brain Related Family

The T–Brain related/ TBR family consists of three proteins TBR1 (otherwise known as T–Brain 1, TBR–1 and TES–56), TBR2 (otherwise known as EOMES) and TBX21 (otherwise known as T–BET). These proteins are grouped together in the family due to the fact that they are expressed in various regions of the brain and all have a degree of sequence similarity between them (Bulfone *et al.* 1995). The members of the TBR1 family are involved in development of the brain, eye, immune system, mesoderm, and placenta.

TBR2 (EOMES) functions in the development of the central nervous system and the mesoderm in addition to differentiation of $CD8_+$ T–cells (Buggert *et al.* 2014; Kimura *et al.* 1999). TBX21 (T–BET) is an important transcription factor in regulation of the immune system. It is involved in the development of T helper 1 cells from naïve T–cells (Oestreich & Weinmann 2012). TBX21 expression in the brain appears to be exclusively located in the mitral and tufted cells (types of neurons) of the olfactory bulb where it acts as a transcriptional activator (Mitsui *et al.* 2011). TBR1 is involved in brain development, and more specifically it plays in role in differentiation and migration of neurons within specific regions of the cerebral cortex (Bulfone *et al.* 1995).

Each member of the T–Brain related family contains a T–Box domain in the centre followed by a C–terminal region known as the T–Box transcription factor associated domain. This domain tends to lie on the C–terminal end of multiple T–Box proteins and

its role and structure as of yet is unclear. The T–Box domain is highly conserved among the 3 members of the TBR1 family and shows sequence similarity to a lesser degree with the Brachyury gene protein (T–protein, the first discovered T–Box protein) and TBX5 (A member of the TBX2 family) (**Figure *16***).

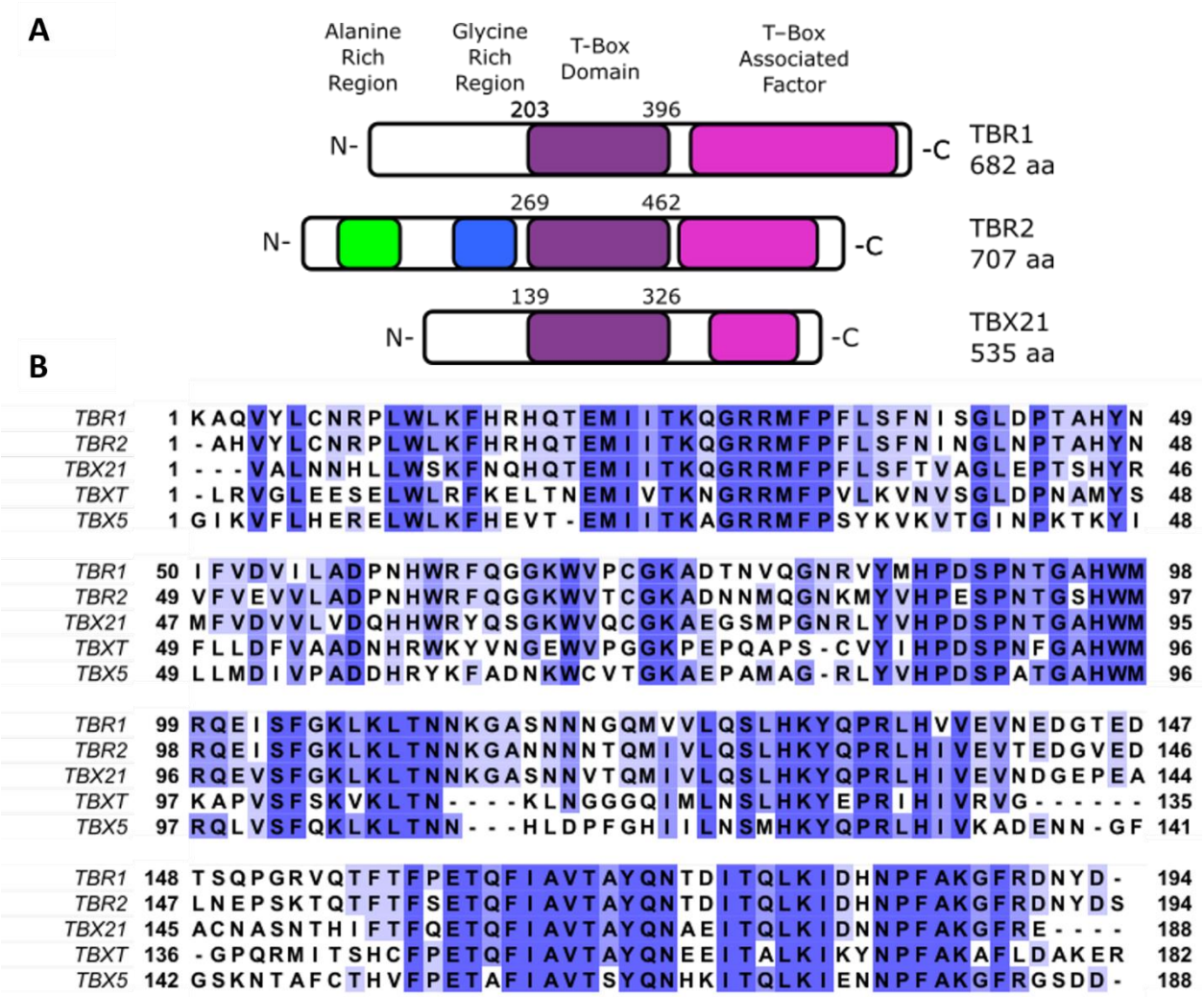


Figure 16. The domain architecture of the three members of the TBR1 family (A) and multiple sequence alignment of the portions of TBR1, TBR2, TBX21, Brachyury (TBXT) and TBX5 which contain the T–Box domain (B).

A: An alanine rich region (green) and glycine rich region (blue) are found on the N–terminal end of TBR2 only. The T–Box domain (purple) is found in the centre followed by a T–Box associated factor (pink) of each T–Box order member. **B**: Multiple sequence alignment of the portions of TBR1, TBR2, TBX21, Brachyury and TBX5 which contain the T–Box domain. The darker the blue the amino acids are highlighted the more conserved the residue is. The TBR2, TBX21, TBXT and TBX5 T–Box domains were aligned to the TBR1 T–Box domain sequence using ClustalW (Larkin *et al.* 2007) and visualised using Jalview version 2 (Waterhouse *et al.* 2009).

1.9.2. T–Box Domain Structure

The T–Box domain has an immunoglobulin–like fold. This is a fold which consists of a 6–7 stranded β–barrel (Strands A–G) consisting of two β–sheets (A,B,E and G,F,C,C',D) arranged in a Greek–key topology (Bork *et al.* 1994). Additionally, the T–Box domain has three more β–sheets (c,c' , f,g,a and e,e',b) which form a lid like structure on the β–barrel (**Figure 11** and **Figure 17**).

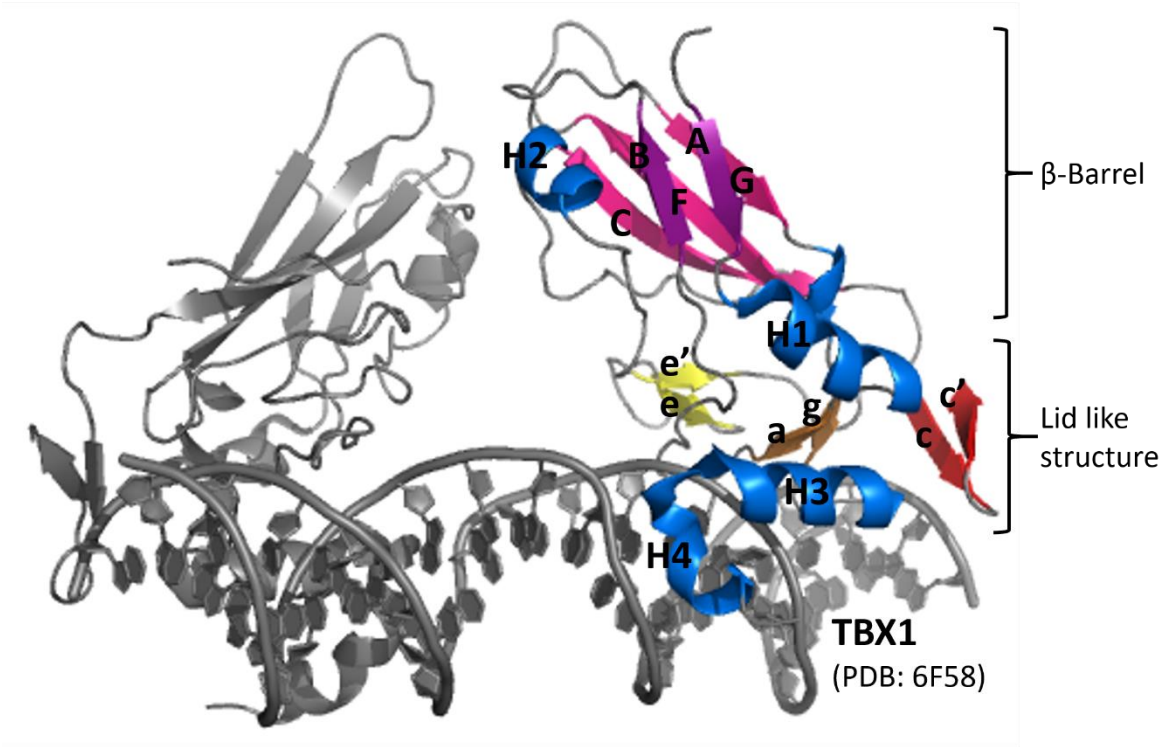


Figure 17: Structure of the human T–Box domain of Brachyury (TBX1) highlighting the secondary structural elements.

A: The β–barrel core is formed by two separate β–sheets made up of strands G, F, and C (pink) and A and B (purple). The lid is made up of three smaller β–sheets made up of strands c' and c (red), e' and e (yellow) and a and g (orange). The four helices (blue) and coils (grey) make up most of the exterior of the protein. (PDB ID: 6F58). Image rendered with PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.

The T–Box domain of the T–protein from *Xenopus laevis* (XBra) (PDB ID: 1XBR) is considered a generic structure of the T–Box domain and was solved in 1997 to a resolution of 2.5 Å (Muller & Herrmann 1997). To date crystal structures of Xbra (PDB ID: 1XBR), TBX1 (PDB ID:4A04), TBX3 (PDB ID: 1H6F), TBX5 (PDB ID: 2X6U, 2X6V, 4S0H, 5BQD and 5FLV), TBX21 (PDB ID: 5T1J) and TBXT (human brachyury) have been solved.

In (**Figure 17**) the two β–sheets that make up the β–barrel are indicated along with the lid structure. The T–Box domain contains four helices, one of which (helix 2) replaces

what is strand D in most immunoglobulin–like folds. Helix 1 is pressed up against and lies almost parallel with helix 3. Helix 3 and helix 4 lie perpendicular to one another and interact with the DNA minor groove (Muller & Herrmann 1997). Despite variations in the T–Box domain, members all seem to bind the same consensus DNA sequence. This is understandable as the portion of the protein involved in the interaction is highly conserved (Stirnimann *et al.* 2010). Some T–Box proteins were found to preferentially bind sequences containing multiple core binding sites, often in a palindrome orientation (Wilson & Conlon 2002).

1.9.3. T–Box DNA Binding and Dimerisation

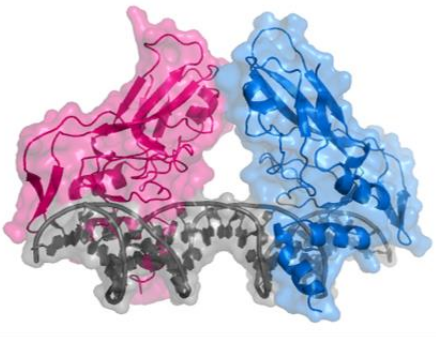
The crystal structures of Xbra (T–protein from *Xenopus laevis*), TBXT, TBX1 and TBX3 all show two T–Box proteins bound to a palindromic DNA sequence. Whether or not these two proteins actually dimerise when they encounter the DNA is a matter of some speculation. The interface between these monomers is very small (**Figure 18**) and so each monomer likely binds independently to the palindromic DNA (Coll *et al.* 2002; El Omari *et al.* 2012). In addition to this, the Xbra T–Box domain (Muller & Herrmann 1997) was shown to be monomeric in solution.

Crystal structures exist for the TBX5 T–Box domain both in the presence and absence of DNA. The DNA–free structures exist for both monomeric (2X6U) and dimeric (5BQD) TBX5 T–BOX domain. The DNA–bound structure shows that one of the monomers of the dimer binds to a DNA sequence containing only a single binding site (**Figure 18**). Interestingly, like the Xbra T–Box, the TBX5 T–Box domain has been reported to be monomeric in solution which suggests that the dimer might be a result of crystal packing or that the protein may require additional domains in order to stabilise the dimer when in solution (Stirnimann *et al.* 2010).

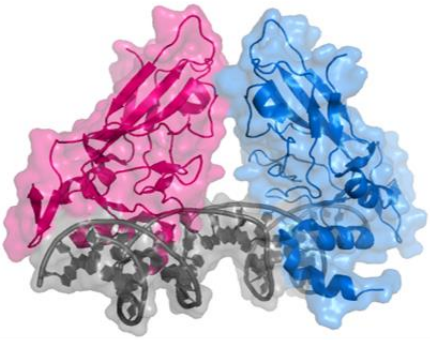
The TBX21 (T–Bet) T–Box domain, the only TBR1 family member to have a solved crystal structure, was crystallised in the presence of palindromic DNA. It bound as a dimer to two separate DNA strands with a much more extensive dimer interfaces observed in the Xbra, TBXT, TBX1 and TBX3 crystal structures (**Figure 18**). The TBX21 dimer interface is made up by two loops (loop 1, residues 248–256 and loop 2, residues 276–293). According to sequence alignments, Loop one appears to be conserved in the TBR1 family but not in other T–Box families and loop two is not conserved at all. The dimer interface is stabilised by salt bridges and electrostatic interactions. TBX21 has been shown to exist as a mixture of monomer and dimer in solution unlike other T–Box domains (Liu *et al.* 2016) which appear to be monomeric in solution.

In the Xbra, TBXT, TBX1 and TBX3 crystal structures, the T–Box domains have been observed as dimers bound to palindrome DNA with very small dimer interfaces (**Figure 18**). However, as of yet only single DNA binding sites have been identified *in vivo*. This calls to question the relevance of these observed T–Box dimers, with exception of TBX21

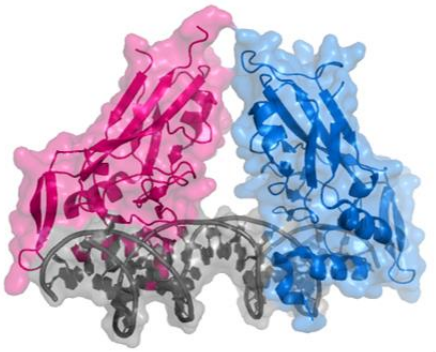
and TBX5 as both form dimers with extensive dimer interfaces (Stirnimann *et al.* 2010; Wilson & Conlon 2002).



Brcahyury T-protein from *Xenopus laevis* (Xbra)
(PDB ID: 1XBR)

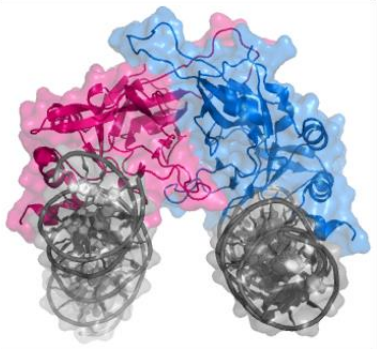


Human Brachyury T-protein (TBXT)
(PDB ID: 6F58)



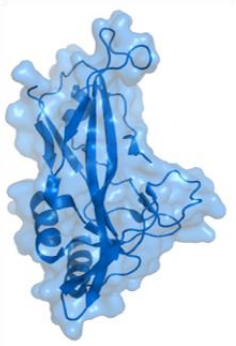
T-box domain of human TBX1
(PDB ID: 4A04)

T-box domain of human TBX3
(PDB ID: 1H6F)

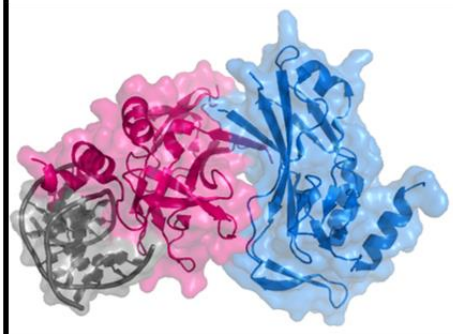


T-box domain of human TBX21 (T-Bet)
(PDB ID: 5T1J)

Dimeric T-box domain of human TBX5
(PDB ID: 5BQD)



Monomeric T-box domain of human TBX5
(PDB ID: 2X6U)



Dimeric T-box domain of human TBX5 bound to DNA
(PDB ID: 2X6V)

Figure 18: The crystal structures of various T–Box domains.

The T–Box domains of various T–Box–containing proteins are displayed. Each monomer is displayed in either pink or blue and the DNA is coloured in grey. XBra, TBXT, TBX1 and TBX3 are seen as dimers bound to palindromic DNA. The dimer interfaces are very small. TBX21 is seen as a dimer binding to separate strands of DNA. The Dimer interface is large. TBX5 is seen as a dimer and a monomer in the absence of DNA. In the DNA–bound form, only a single monomer of the dimer is seen interacting with DNA. The dimer interface is large and there appears to be no large protein conformational change upon DNA binding. The PDB codes are displayed below each structure (El Omari *et al.* 2012; Liu *et al.* 2016; Muller & Herrmann 1997; Pradhan *et al.* 2016; Stirnimann *et al.* 2010). Image rendered with PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.

The crystal structure of the T–Box domain of the Brachyury T–protein bound to DNA (Xbra) (PDB ID: 1XBR) shows that the T–Box domain binds DNA via interactions with both the major and minor groove (**Figure 19**). These interactions are mediated by helix 3, helix 4, as well as the loops and strands of the two small β–sheets that form the lid of the β–barrel. Helices 3 and 4 interact via the minor groove. More specifically, helix 3 lies across the DNA backbone making contacts with phosphate groups and forming a hydrophobic interaction between an isoleucine residue and a guanine. Helix 4 is inserted into the minor groove mediated by hydrophobic interactions with the bases and sidechains. The loops extending from the smaller β–sheets form various interactions with the DNA backbone. Extensive hydrophobic interactions are typical of minor groove binding proteins; what differs in T–Box proteins, however, is the lack of DNA bending present (Muller & Herrmann 1997). TBX5, the only T–Box domain structure solved in the presence and absence of DNA, shows no major differences between the DNA bound and unbound forms (Stirnimann *et al.* 2010).

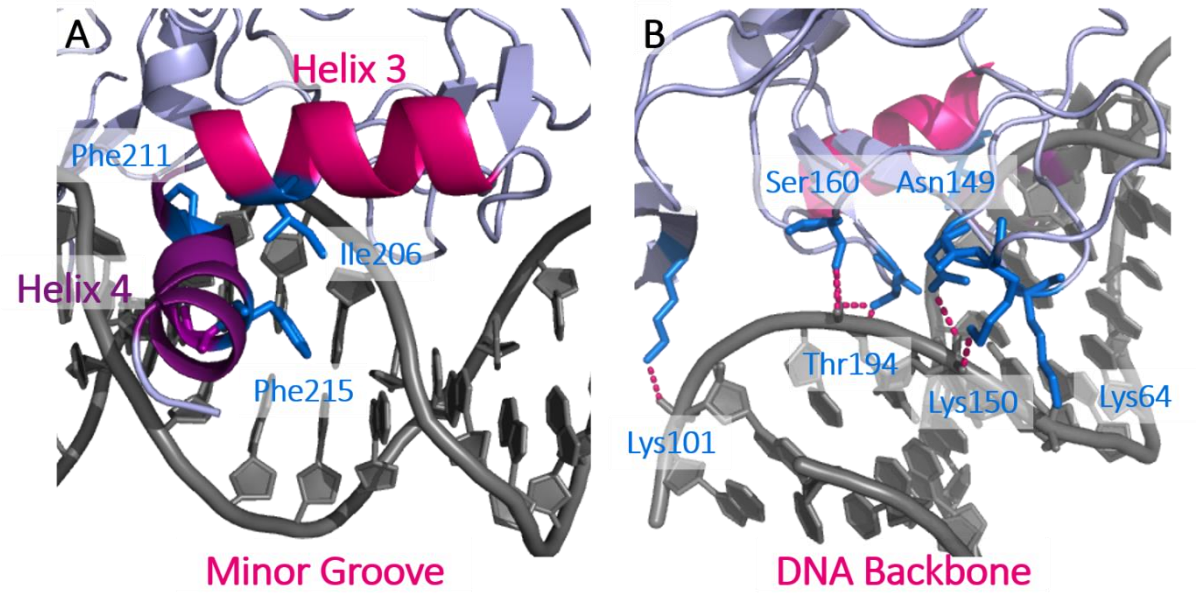


Figure 19: Hydrophobic interactions in the minor groove and hydrogen bonds with the backbone of the DNA and Xbra.

Interactions that occur between the T–Box domain of Brachyury T–protein and DNA as per the 1XBR crystal structure (Muller & Herrmann 1997). **A:** Helix 4 (purple) is inserted into the minor groove, whereas helix 3 (pink) lies across the DNA backbone. Three residues which are inserted into the minor groove and that partake in a hydrophobic interaction are coloured in blue. **B**: Residues (blue) protruding from loops in the lid portion of the structure form hydrogen bonds (pink) with the DNA backbone (grey). Image rendered with PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.

1.9.4. T–Box Protein–Protein Interactions

T–Box transcription factors form important protein–protein interactions. In many cases they have activation and repression domains that lie on either side of the T–Box domain. The T–Box domain itself has been shown to be involved in various protein–protein interactions. These protein–protein interactions alter the ability of these proteins to activate or repress transcription (Hiroi *et al.* 2001; Pradhan *et al.* 2016; Vance *et al.* 2010), nuclear import (Hsueh *et al.* 2000) and even epigenetic modification (Miller *et al.* 2008).

1.10. TBR1

T–Brain protein 1 (TBR1) is a T–Box protein in the T–Brain related family. As previously stated, it contains a T–Box domain, capable of binding DNA and forming protein–protein interactions. In addition to this it contains a C–terminal T–Box associated domain which functions as a transcriptional activation or repressor domain.

The structure of TBR1 has not yet been solved and its DNA binding mechanism and any other macromolecular interactions that it may form have not been characterised to

date. However, there are many studies available that have investigated its expression patterns and identified interactions, especially in association with autism spectrum disorders (ASD). Multiple gene targets of TBR1 are known, three of which are: Reelin (RELN) , glutamate receptor subunit epsilon 2 (GRIN2B) and autism susceptibility gene 2 protein (AUTS2) (Deriziotis *et al.* 2014b). The RELN (a glycoprotein involved in neuronal migration) gene was shown to be upregulated in the presence of TBR1 and further enhanced by the interaction between TBR1 and its binding partner calcium/calmodulin–dependent protein kinase (CASK) (Hsueh *et al.* 2000). TBR1 also upregulates the GRIN2B gene which encodes for the glutamate receptor ionotropic, NMDA 2B (GluN2B). GluN2B forms part of the N–Methyl–D–Aspartate Receptor (NMDAR). TBR1 itself is then upregulated by NMDAR and CASK showing that TBR1 forms part of a positive feedback loop (Chuang *et al.* 2014). AUTS2, a protein of unknown function has been implicated in autism and mental retardation (Huang & Hsueh 2015). It has been shown that the TBR1 and AUTS2 genes are co–expressed in various tissues at different stages of development (Bedogni *et al.* 2011). **Table 2** shows the various proteins that are known to interact with TBR1 to date.

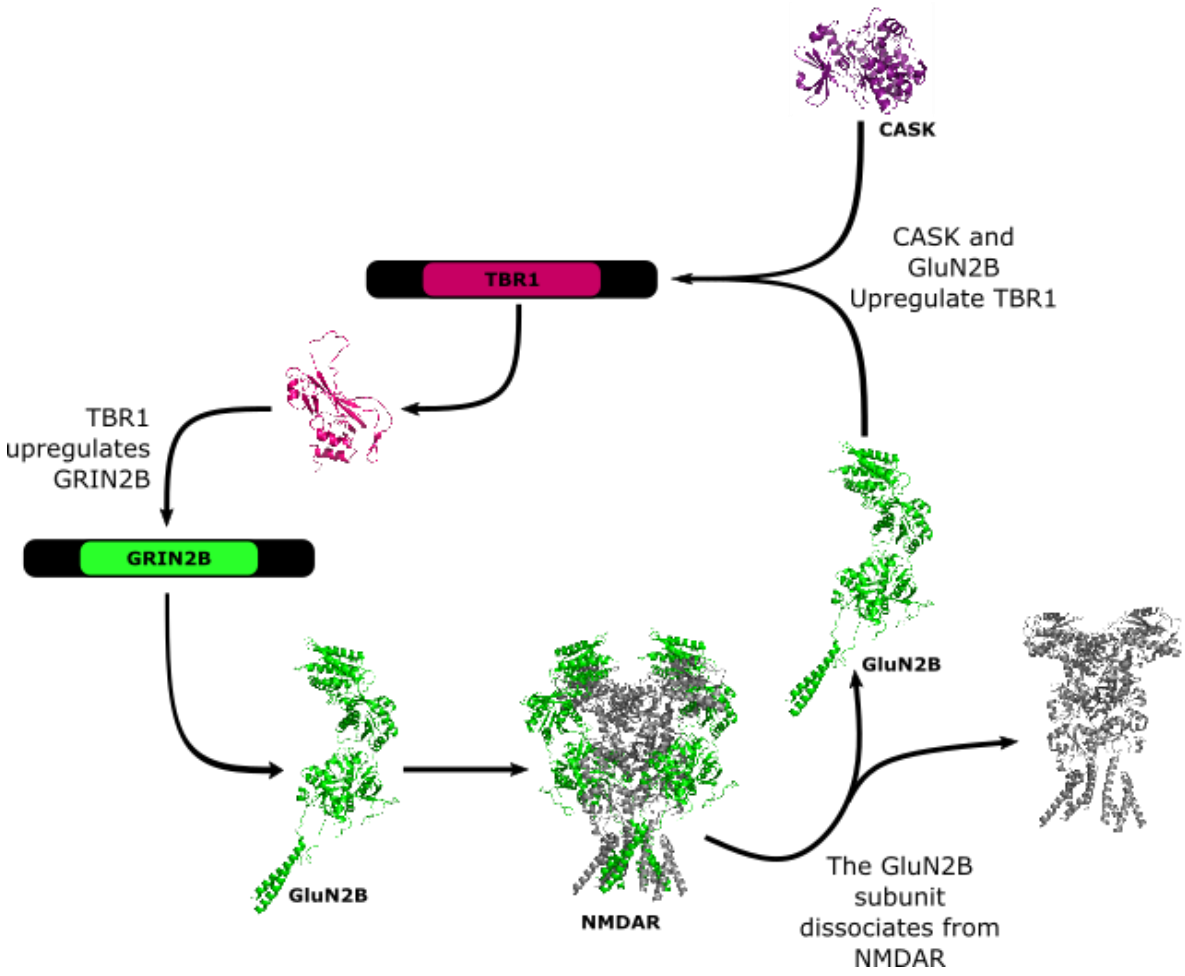


Figure 20: TBR1, GRIN2B and CASK positive feedback loop.

TBR1 (pink) upregulates expression of GRIN2B (green). The GluN2B (green) subunit of NMDAR (grey) dissociates and in conjunction with CASK (purple) upregulates expression of TBR1 (pink), forming a positive feedback loop. The protein structures used to represent TBR1, GluN2B/NMDAR and CASK were the T–Box domain of Xbra (1XBR) (Muller & Herrmann 1997), GluN1/GluN2B NMDA receptor (4TLL)(Lee *et al.* 2014) and CASK–CaMK (6LMN). Image rendered with PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC. Image created using Inkscape 0.92.4.

Table 2: Interacting partners of TBR1

	Protein	Interaction type	Experiment	Reference
AP1	Transcription factor AP–1	association	Anti–bait coimmunoprecipitation	(Li *et al.* 2015)
BCL11A	B–cell lymphoma/leukemia 11A	physical association colocalization	bioluminescence resonance energy transfer	(den Hoed *et al.* 2018)
CASK	Calcium/calmodulin–dependent protein kinase	physical association colocalization	yeast two–hybrid screen, coimmunoprecipitation	(Hsueh *et al*. 2000)
ESR1	Ethylene–responsive transcription factor	association	tandem affinity purification	(Stellato *et al.* 2015)
EZRI	Ezrin	association	pull down	(Liu *et al.* 2018)
FOXP1	Forkhead Box Protein 1	physical association colocalization	bioluminescence resonance energy transfer, confocal microscopy	(Deriziotis *et al.* 2014b)
FOXP2	Forkhead Box Protein 2	physical association colocalization	bioluminescence resonance energy transfer, confocal microscopy	(Deriziotis *et al.* 2014b)
FOXP4	Forkhead Box Protein 4	physical association colocalization	bioluminescence resonance energy transfer, confocal microscopy	(Deriziotis *et al.* 2014b)
HscB	Iron–sulfur cluster co–chaperone protein	Association	Anti– bait coimmunoprecipitation	(Maio *et al.* 2017)
MTA70	N6–adenosine–methyltransferase catalytic subunit	Association	Anti– bait coimmunoprecipitation	(Yue *et al.* 2018)
RMND5B	E3 Ubiquitin–Protein Transferase	Association	tandem affinity purification	(Boldt *et al.* 2016)
TBR1	T Brain Related Protein 1	Physical association Colocalization, Association	bioluminescence resonance energy transfer, confocal microscopy, anti–bait coimmunoprecipitation, tandem affinity purification, pull down	(Deriziotis *et al.* 2014b; Li *et al.* 2015; Liu *et al.* 2018; Stellato *et al.* 2015; Yue *et al.* 2018)

Data from the table was curated by the EMBL–EBI IMEx database (Orchard *et al.* 2012) and various papers. A direct interaction is defined as two molecules in direct contact with each other. Physical association is defined as molecules that are in the same complex, they do not necessarily interact directly. Co–localisation is defined as molecules in the same subcellular location, the methodology used is not high resolution enough to confirm an interaction.

1.11. Autism Spectrum Disorder

Autism spectrum disorders (ASD) are characterised by multiple symptoms which include impairments or deficits in communication and social interaction. ASD often also presents with repetitive behaviour. ASD sufferers, however, do not all exhibit the same symptoms (Lord *et al.* 2000).

In recent years next generation sequencing has allowed insight into many genetic variations present in multiple suffers of ASD symptoms. Complete knockouts of a variety of genes via rare mutations have been discovered in suffers of ASD and provide insight into important proteins involved in ASD (De Rubeis *et al.* 2014). Wild type TBR1 has been shown to influence the expression of 24 genes associated with ASD including AUTS2, GRIN2B and RELN (Huang & Hsueh 2015). Furthermore, certain *de novo* TBR1 mutations have been identified as significant risk factors for ASD (Krumm *et al.* 2014).

FOXP2 is one of only a few proteins shown to interact with TBR1 in the cortex, the others are FOXP1, CASK and BCL11A (B–cell lymphoma/leukemia 11A) (den Hoed *et al.* 2018; Deriziotis *et al.* 2014b; Hsueh *et al.* 2000). Mutations in TBR1, CASK, BCL11A, FOXP1 and FOXP2 have all been associated with ASD or intellectual disability and it is possible they could form a regulatory complex (den Hoed *et al.* 2018; Deriziotis *et al.* 2014b; Dias *et al.* 2016; Sollis *et al.* 2017).

When TBR1 is present, CASK is imported to the nucleus where there is then an observed increase in RELN and GRIN2B gene expression (Hsueh *et al.* 2000). The interaction between TBR1 and FOXP2 is less clear although it has been substantially proven that mutations in both genes result in disruption of their interaction (Deriziotis *et al.* 2014b). It has been proposed that an interaction between FOXP2, TBR1 and CASK regulates the expression of genes such as RELN, GRIN2B, AUTS2 and possibly FOXP2 (**Figure 21**). Mutations in TBR1 have been shown to alter FOXP2 expression (Huang & Hsueh 2015) suggesting that they may form a feedback loop for FOXP2 regulation. Furthermore, GRIN2B's subunit NMDA regulates TBR1 expression suggesting this complex may also form part of a TBR1 feedback loop (Deriziotis *et al.* 2014b).

In addition to TBR1 interacting with CASK, FOXP2 and FOXP1, it also interacts with BCL11A. The interface for this interaction involves part of the TBR1 T–Box domain. Interestingly, BCL11A also interacts with NR2F1 and NR2F2, two proteins that have recently been shown to interact with FOXP2 (den Hoed *et al.* 2018; Estruch *et al.* 2018). Moreover, as previously discussed, it has also been shown that different combinations of FOXP homo– and heterodimerisation events affect how FOXP2 regulates a number of genes (Mendoza & Scharff 2017). It is thus likely that FOXP2, FOXP1, TBR1, CASK and BCL11A form a variety of large complexes that result in the tight regulation of genes such as RELN, AUTS2 and GRIN2B. Mis–formation of these complexes due to mutations is likely to initiate the underlying mechanisms which cause ASD and ID. Therefore, a more complete understanding of the mechanisms that control these pathways can

provide insight into developmental disorders such as ASD and how multi–protein complexes regulate gene expression.

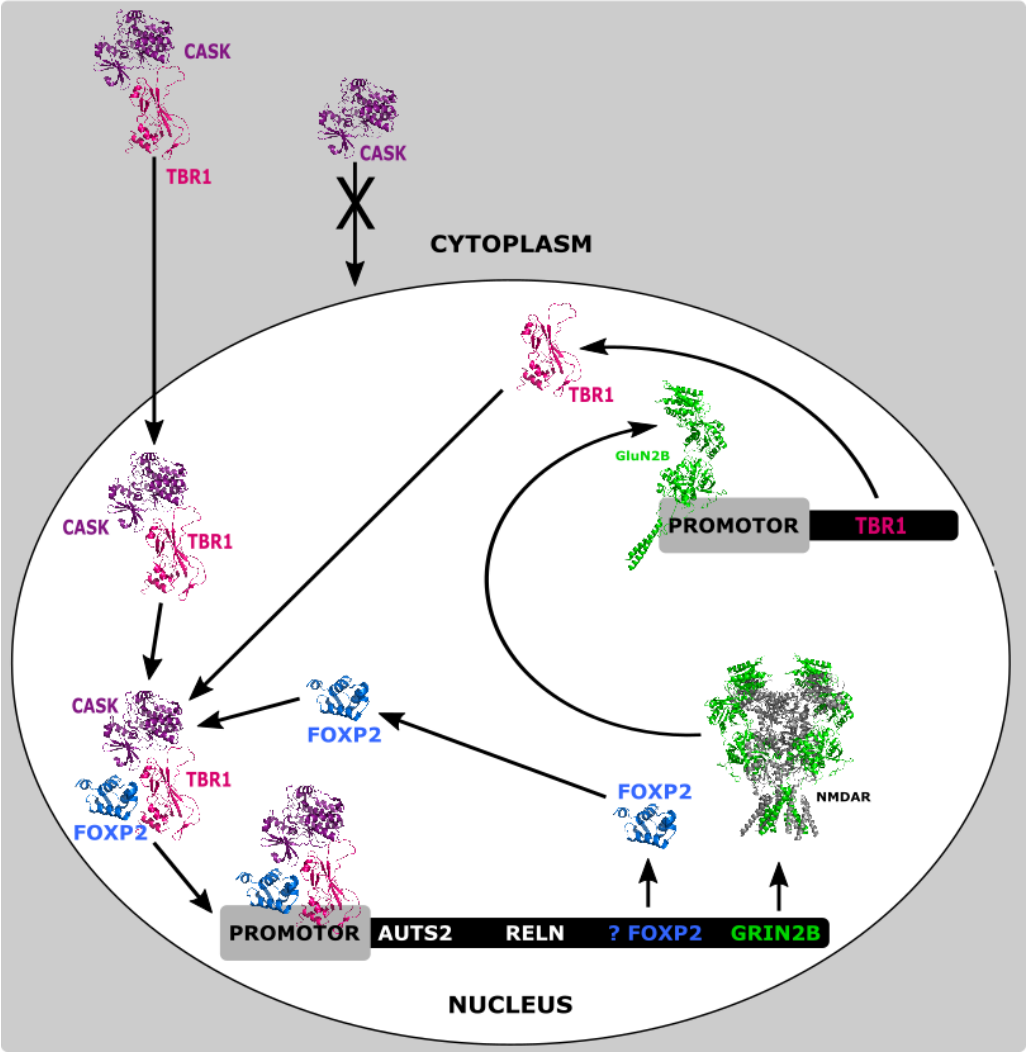


Figure 21: Proposed mechanism of gene regulation via a CASK–TBR1–FOXP2 interaction.

Hsueh *et al*, 2000 showed that in presence of TBR1 (pink), CASK (purple) is imported from the cytoplasm into the nucleus. FOXP2 (blue) interacts with TBR1 and does not distrupt the TBR1–CASK interaction. It has, therefore, been proposed that FOXP2, TBR1 and CASK form a ternary complex responsible for the regulation of genes such as RELN, GRIN2B (green) and AUTS2. TBR1 may also regulate the expression of FOXP2 (Huang & Hsueh 2015). The GluN2B (green) subunit of NMDAR (grey) upregulates TBR1 expression (Chuang et al, 2014). Therefore, the CASK–TBR1–FOXP2 interaction may also form a regulatory feedback loop for both TBR1 and FOXP2. The protein structures used to represent FOXP2, TBR1, GluN2B/NMDAR and CASK were the FOXP2 FHD (2A07) (Stroud *et al.* 2006), the T–Box domain of Xbra (1XBR) (Muller & Herrmann 1997), GluN1/GluN2B NMDA receptor (4TLL) (Lee *et al.* 2014) and CASK–CaMK (6LMN) respectively. Protein images rendered with PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC. Image created using Inkscape 0.92.4.

1.12. TBR1–FOXP2 INTERACTION

TBR1 and FOXP2 are co–expressed in the brain (layer 6 of the cortex) and these two transcription factors are exclusively located in the nucleus. *De novo* and rare mutations in both FOXP2 and TBR1 have been linked to speech defects, ASD and developmental delay (Deriziotis & Fisher 2017).

Deriziotis *et al*, 2014 have shown that TBR1 and FOXP2 interact using bioluminescence resonance energy transfer (BRET) assays. The BRET assays were performed between the T–Box domain of TBR1 and multiple FOXP2 isoforms. They found that isoform 10+ (lacking the C–terminal regions of FOXP2, including the FHD), did not interact with TBR1 (**Figure 22**). Additionally, the full length FOXP2 with the R553H mutation in the FHD was also unable to interact with TBR1, indicating that the FHD may be involved in the interaction with TBR1. They also tested FOXP2 variants lacking the polyglutamine tracts and determined the TBR1 binding site may lie between the polyglutamine tracts. It is therefore possible that there are two TBR1 binding sites on FOXP2, one in between the polyglutamine tracts and one within the FHD.

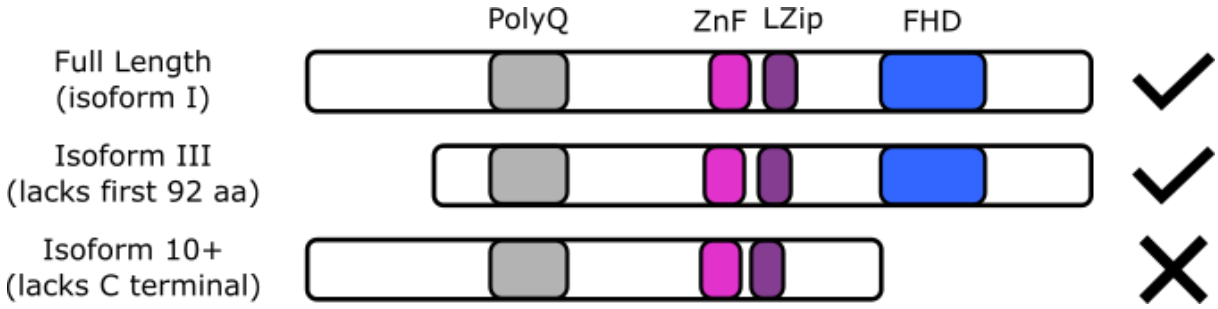


Figure 22: FOXP2 isoforms used in TBR1 BRET assays.

The full length FOXP2 (isoform 1) and isoform III (lacking the first 92 amino acids) interacted with TBR1. Isoform 10+ lacking the C–terminal region of FOXP2 including the FHD did not interact with TBR1. The polyglutamine rich region (grey), zinc finger (pink), leucine zipper (purple) and the forkhead domain (blue) are indicated. Image created using Inkscape 0.92.4.

Multiple mutations in TBR1 T–Box domain, which were discovered in ASD patients, were shown to lead to the disruption of the FOXP2–TBR1 interaction (**Table 3** and **Figure 23**) (Deriziotis *et al.* 2014b). The authors of this study concluded that the interaction between TBR1 and FOXP2 occurred between the T–Box domain of TBR1 and either the glutamine rich region or the C– terminal regions of FOXP2 (Deriziotis *et al.* 2014b).

Table 3: Effects of mutations in the TBR1 gene found in patients suffering from ASD

Mutation	Type	Region on TBR1	Effect on DNA biding	Protein–protein interactions		TBR1 dimerisation
				CASK	FOXP2	
K228E	*De novo*	T–Box	Not significant	Yes	No	Yes
N374H	*De novo*	T–Box	Not significant	Yes	No	Yes
S351X	*De novo*	Truncated missing part of the T–Box domain	No binding noted	yes	No	No
Q178E	Rare Inherited	N terminal	Not significant	Yes	Yes	Yes
Q418R	Rare Inherited	C terminal	Not significant	Yes	Reduced	Yes
P542R	Rare Inherited	C terminal	Not significant	Yes	Yes	Yes
V356M	Rare Inherited	T–Box	Not significant	Yes	Yes	Yes

This table displays the effects are various TBR1 gene mutations found in patients suffering from ASD (Deriziotis *et al.* 2014b). The coloured mutations are found in the T–Box domain and are displayed in **Figure 23**.

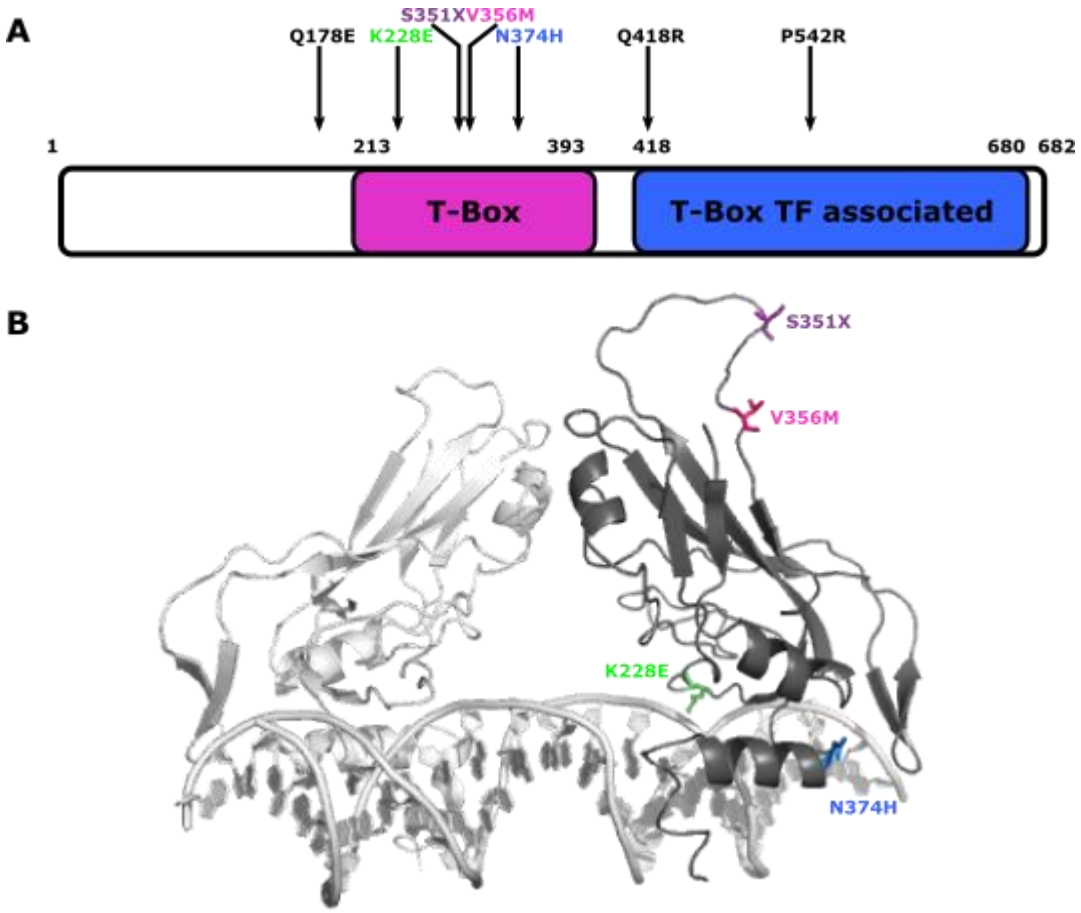


Figure 23: The domain architecture of TBR1 and a model of the TBR1 T–Box showing the ASD–causing mutations studied by Deriziotis *et al.* (2014b)**.**

A: The domain architecture of TBR1 is displayed as boxes. The approximate position of each mutation is indicated by the arrows. **B**: The TBR1 T–Box domain model (Black) is displayed alongside the XBra monomer and DNA (white) (PDB ID: 1XBR)(Muller & Herrmann 1997). The T–box mutations that were studied are indicated and coloured accordingly. Protein image rendered with PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC. Image created using Inkscape 0.92.4.

RATIONALE

Neural development is a complex, strictly regulated process that depends on an intricate network of interactions between various macromolecules. These include transcription factors which interact with DNA, other transcription factors, as well as transcription machinery. Mis–regulation of transcription, which can result in neurological conditions, is often a result of mutations that alter the interactions that a transcription factor forms with other macromolecules(Latchman 1997; Tootle & Rebay 2005). Both FOXP2 and TBR1 are transcription factors that are co–expressed in the brain. They have been, along with some of their gene targets, associated with autism spectrum disorder and other neurological disorders (Fisher & Scharff 2009; Gong *et al.* 2004; Huang & Hsueh 2015; Krumm *et al.* 2014; Mukame *et al.* 2011). It has been suggested that an interaction between them could be an important part of the gene expression required for development (Deriziotis & Fisher 2017). Deriziotis *et al*, (2014) have suggested that FOXP2 and TBR1 form an interaction *in vivo*. One of the shortfalls of the current evidence for this interaction is that it is unknown whether they interact directly or form part of the same complex. A complete picture of the interactions that FOXP2 and TBR1 form with each other and with DNA is required to understand conditions such as ASD and ID. Only once the network of interactions involved in the relevant developmental pathways is understood, can potential treatments to such conditions be designed. For this reason, the TBR1–FOXP2 interaction was investigated in three parts in this study as follows:

i) the homomeric association between TBR1 T–Box monomers.
ii) the interaction between FOXP2 forkhead domain and TBR1 T–box domain
iii) the interaction between DNA and either FOXP2 or TBR1, both in the absence and presence of the interacting protein partner.

1. Homomeric association between TBR1 T–Box monomers

   Dimerisation of transcription factors is known to affect gene regulation by altering DNA binding affinity and specificity, as well as cellular location (Amoutzias *et al.* 2008). There has been very little work done to characterise the quaternary structure of the TBR1 T–Box domain, and there are no crystal structures available for it. Other T–Box domain crystal structures have been found to be either dimeric or monomeric in both the apo and the DNA bound forms (Coll *et al.* 2002; El Omari *et al.* 2012; Muller & Herrmann 1997; Stirnimann *et al.* 2010). When bound to DNA, these T–Box domains interact as a monomer if only one DNA binding sequence is present, while if two palindromic binding sequences next to each other are present in the DNA oligomer, a dimer forms. There is some discrepancy about the T–Box dimers observed binding to this palindrome DNA in crystal structures, as the dimer interfaces are very small and the palindrome sequence itself has not been identified

in vivo (Wilson & Conlon 2002). TBX21, a T–Box protein in the same family as TBR1, is seen as a dimer in its crystal structure with a much more extensive dimer interface, and only occupies one of the binding sites on the palindrome DNA (Stirnimann *et al.* 2010). One of the loops involved in this TBX21 dimer interface is conserved in the TBR1 T–Box domain, which suggests that it is possible that the TBR1 T–Box domain might form a physiologically relevant dimer. Additionally, The TBR1 T–Box domain has two cysteine residues that are not conserved in other T–Box domains. The presence of these could suggest a dimer formed via a disulfide bridge.

2. Interaction between FOXP2 forkhead domain and TBR1 T–box domain

   Using *in* vivo studies, Deriziotis *et al.* (2014b) showed that FOXP2 and TBR1 interact. However, because they used a BRET assay, it cannot be concluded whether a direct interaction occurs between the two proteins, or if they simply form part of the same multi–molecular complex. Furthermore, they proposed that the interaction between FOXP2 and TBR1 occurs between the T–Box domain of TBR1 and either a region within the polyglutamine rich region or the FHD of FOXP2. This study aimed to confirm if the T–Box domain of TBR1 interacts with the FHD of FOXP2 directly.

3. Interaction between DNA and either FOXP2 or TBR1, both in the absence and presence of the interacting protein partner

   The interaction between FOXP2 and DNA has already been well characterised by previous studies (Morris *et al.* 2018; Morris & Fanucchi 2016; Stroud *et al.* 2006; Webb *et al.* 2017). There has, however, been almost no research done to characterise TBR1 T–Box DNA binding. For this reason, the thermodynamic and kinetic parameters of the TBR1 T–Box DNA interaction were investigated to establish the binding mechanism. Knowing how the TBR1 T-Box binds various DNA sequences, as well as the mechanism in which it binds, could be used to elucidate its function with respect to each binding site. Furthermore, knowing the binding mechanism could help to identify important residues in the interaction, which could potentially be used to find target areas for future treatments for conditions such as ASD. Furthermore, transcription factors function by forming multi–macromolecular complexes. It is of importance to know if an interaction between FOXP2 and TBR1 alters the function of either protein. This could give valuable insight into how these macromolecules function together. This is the first essential step in determining the roles they play in disorders such as ASD and how transcription factors function in general.

AIM AND OBJECTIVES

2.1. AIM

To characterise the FOXP2–TBR1 interaction and determine if the interaction regulates TBR1 or FOXP2 DNA binding function.

2.2. OBJECTIVES

2.2.1. Overexpress (in *E. coli*) and purify the FOXP2 FHD and TBR1 T–Box domain using immobilised metal affinity chromatography and size exclusion chromatography.

2.2.2. Perform low resolution structural characterisations (tertiary and quaternary structure) of the FOXP2 FHD and the T–Box domain of TBR1 using intrinsic tryptophan fluorescence and size exclusion chromatography.

2.2.3. Characterise the interaction between the T–Box domain of TBR1 and its single and palindrome DNA sequences using electrophoretic mobility shift assay, isothermal titration calorimetry and single molecule–Förster resonance energy transfer.

2.2.4. Determine whether the T–Box domain of TBR1 interacts directly with the FOXP2 FHD and characterise any interaction using a super shift electrophoretic mobility shift assay and fluorescence anisotropy.

2.2.5. Determine if the interaction between the FOXP2 FHD and TBR1 T–Box regulates FOXP2 DNA binding using fluorescence anisotropy.

METHODS AND MATERIALS

Although it has been shown that the TBR1 T–Box domain interacts with either the forkhead domain (FHD) or the glutamine rich region of FOXP2, the regulatory effect of the interaction is unclear. The aim of this study was, therefore, to determine whether the FOXP2 forkhead domain (FHD) and the TBR1 T–Box domain (T–Box) interact and if so, how that interaction affects TBR1 T–Box and FOXP2 FHD DNA binding. To achieve this, a complete picture of both of the interacting parties is required, which includes insight into both their structures and their DNA–binding functions.

Both the FOXP2 FHD and the TBR1 T–Box domain were over expressed in *E. coli* cells and purified using liquid chromatography. The structure of each protein was subsequently assessed using intrinsic tryptophan fluorescence and size exclusion chromatography (SEC) in order to confirm that each protein was folded. The interaction between TBR1 T–Box domain and DNA was studied using various techniques including electrophoretic mobility shift assay (EMSA), isothermal titration calorimetry (ITC) and single molecule–Förster resonance energy transfer (smFRET). Subsequently the interaction between TBR1 T–Box and the FOXP2 FHD was studied using a super shift EMSA and fluorescence anisotropy (FA). Finally, FA was used to determine if the FOXP2 FHD–TBR1 T–Box domain interaction alters either protein's DNA binding function.

3.1. MATERIALS

Table 4 contains the manufacturers of all specific kits and reagents. All other reagents were of analytical grade and solutions were prepared with distilled water and filtered with a 0.45 µM filter unless otherwise stated.

Table 4: The manufacturers of specific reagents, materials and kits used in this study.

Method	Material	Manufacturer
Mutagenesis, sequencing, and transformation	pET–lla plasmid	GenScript, USA
	XL10 Gold ultra–competent cells	New England Biolab, USA
	Site directed mutagenesis kit	Strategene, USA
	Ampicillin	Melford, UK
	T7 *E. coli* competent cells	New England Biolab, USA
	GeneJet mini prep kit	Thermo Fisher scientific, EU
	Primers for mutagenesis	Inqaba biotech, RSA
	Sybr Gold nucleic acid gel stain	Invitrogen, USA
	O'GeneRuler 1 kb DNA Ladder	Thermo Fisher Scientific, EU
Expression	Isopropyl β–D–1–thiogalactopyranoside	Melford, UK
	Lysozyme	Sigma–Aldrich, USA
	DNase	Merck, Germany
	Phenylmethylsulfonyl fluoride	Roche, Germany
Purification and cleavage	HisTrap column	GE healthcare, Sweden
	Benzamidine column	GE healthcare, Sweden
Size exclusion chromatography	Superdex 16/60 75 column	GE healthcare, Sweden
	Low range gel filtration calibration kit	GE healthcare, Sweden
EMSA and ITC	Duplex DNA	Integrated DNA technologies–WhiteSci scientific, RSA
Fluorescence Anisotropy	Atto 550–Nitrilotriacetic acid, complexed to Ni2+ (NTA–Atto)	Sigma–Aldrich, SA
	ROX labelled DNA	Integrated DNA technologies–WhiteSci scientific, RSA
SDS–PAGE	Thermo scientific PageRuler^{TM} Prestained Protein Ladder, 10 to 180 kDa	Thermo Fisher scientific, EU
	Thermo scientific low range unstained protein ladder, 100 to 3.4 kDa	Thermo Fisher scientific, EU
smFRET	Alexa Fluor™ 647 C2 Maleimide	Thermo Fisher scientific, EU
	Biotin and Cy3 labelled DNA	Integrated DNA technologies– WhiteSci scientific, UK
	Glucose oxidase	Sigma–Aldrich, UK
	Glucose catalase	Sigma–Aldrich, UK
	Trolox	Sigma–Aldrich, UK

3.2. Methods

3.2.1. Mutagenesis, Transformation and Expression

3.2.1.1. PET–11a Vector and Protein Constructs

This study was conducted on the FHD of FOXP2, as well as the T–Box domain of TBR1. The FHD of FOXP2 was used as opposed to the full length FOXP2, because it is the DNA binding domain and also one of the domains believed to interact with TBR1. Likewise, the T–Box domain of TBR1 was used, as the T–Box domain of TBR1 is already known to be the section of TBR1 that binds to FOXP2. Two TBR1 T–Box mutations were made, in which each of TBR1's two cysteine residues were mutated in turn. These mutations were done to ascertain if any T–Box dimer formation occurred as a result of disulfide bonding between either or both of these two un–conserved cysteine residues.

Codon optimised gene sequences for the FOXP2 FHD and the TBR1 T–Box domain (**Table 5**) were synthesised and incorporated into the pET–11a vector (GenScript, USA) (**Figure 24**). The pET–11a expression vector allows control via T7 polymerase promoter and contains the gene for ampicillin resistance (AmpR) for selection. It also contains a *Lac*I gene to prevent noninduced expression. The gene construct for each protein was inserted between the Lac Operator and the T7 terminator using the Nde1 and BamH1 restriction sites. Each protein construct had a tag containing six histidine residues to aid in purification via immobilised metal affinity chromatography. A thrombin cleavage site (LVPRGS) was inserted after the Hexahistidine–tag (His–tag). Thrombin cleaves between the arginine and glycine residues to remove the tag after purification. The WT TBR1 T–Box construct also contained a Strep–tag II for purification via streptavidin affinity chromatography, however, the protein was purified sufficiently without using it. The pET–lla plasmids containing the genes encoding the WT TBR1 T–Box and the WT FOXP2 FHD were obtained and codon optimised for expression in *E. coli* from GenScript (GenScript, USA).

Table 5: The various protein constructs used for the study.

Protein	Construct Name	Residues	Construct
FOXP2	WT FHD	502–585	MASMTGGQQMGMHHHHHHSSGLVPRGSFHD
TBR1	WT T–Box	203–396	MGSAWSHPQFEKGSSHHHHHHSSGLVPRGST–Box
	C209A T–Box	203–396	MGSAWSHPQFEKGSSHHHHHHSSGLVPRGST–Box C209A
	C274S T–Box	203–396	MGSAWSHPQFEKGSSHHHHHHSSGLVPRGST–Box C274S

The constructs of the various proteins are displayed. Grey regions are linker regions, the His–tags are displayed in purple. The Strep–tag is displayed in pink. Thrombin cleavage sites are shown in blue. The position of the actual protein sequence, C–terminal to the tags, is indicated in in green.

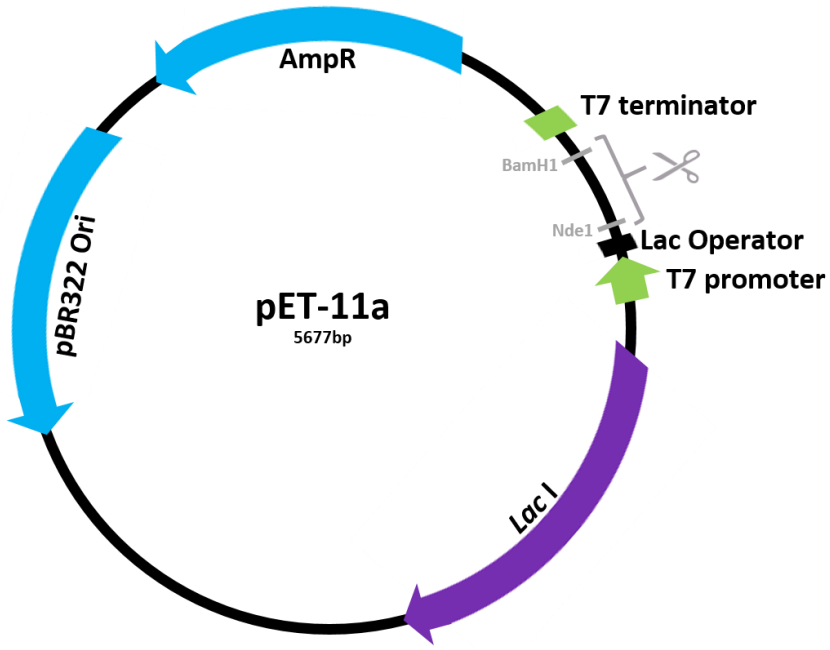


Figure 24: Schematic diagram of the pET–11a plasmid.

Diagram of the pET–11a plasmid. The plasmid contains a T7 promoter (green), the gene for ampicillin resistance (AmpR) (blue) and a LacI gene (purple) to prevent non–induced expression. The gene encoding the protein constructs were inserted between the lac operator and the T7 terminator (green) using Nde1 and BamH1 restriction sites.

3.2.1.2. Site Directed Mutagenesis

The TBR1 T–Box domain contains two cysteine residues (Cys 209 and Cys 274). Each residue was mutated in turn to establish if any dimer formation observed with the WT T–Box was a result of disulphide bonding. In order to obtain the C209A and C274S TBR1 T–Box mutations, site directed mutagenesis was performed using the WT TBR1 T–Box plasmid as a template. Site–directed mutagenesis is a polymerase chain reaction (PCR) designed to produce specific point mutations in a plasmid. This is achieved by using long primers which encode the desired mutation meaning that after each cycle of PCR, more mutant plasmids are produced. The WT template is digested using Dpn1 after the PCR is finished resulting in a product containing only intact mutant plasmid (Braman *et al.* 1996). Primers (**Table 6**) encoding the mutations were designed using PrimerX (http://www.bioinformatics.org). Complementary primers with the mutation in the centre were designed with melting temperatures between 75 and 85 °C and with a GC content between 40 and 60%. The primers were terminated with either a C or G. The primers were synthesised at Inqaba Biotec (RSA) and provided as a lyophilised pellet. The primers were resuspended in milli Q water at a concentration of 10 µM.

Table 6: Primers designed for mutagenesis using PrimerX.

	Primer Sequence	CG content (%)	TM (°C)	Length	Number of base changes
TBR1 T–Box C274S Forward Primer	5' GCAAGTGGGTTCCGTCTGGTAAAGCGGACACC 3'	59.38%	78.6 °C	32 bp	2
TBR1 T–Box C274S Reverse Primer	5' GGTGTCCGCTTTACCAGACGGAACCCACTTGC 3'				
TBR1 T–Box C209A Forward Primer	5' CGCAGGTGTACCTGGCCAACCGTCCGCTG 3'	68.97%	79.5 °C	29 bp	2
TBR1 T–Box C209A Reverse Primer	5' CAGCGGACGGTTGGCCAGGTACACCTGCG 3'				

The CG content, melting temperature (TM), length and number of bases changed are displayed. The changed bases are coloured in blue.

Mutagenesis was performed using the QuikChange site directed mutagenesis kit (Stratagene, USA) according to the manufacturer's protocol. The components and cycling conditions used are displayed in (**Table 7**) and (**Figure 25**). The PCR products obtained were run on a 1% agarose gel and stained with 1:10 000 dilution Sybr Gold DNA stain (Invitrogen, Life technologies). The PCR products were used to transform XL–10 gold ultracompetent cells (using the same protocol as below), cultured and sequenced in the same manner as the WT plasmids.

Table 7: Setup of site–directed mutagenesis reactions.

Reagent	Sample
Reaction buffer (10x)	5 µl
Ds DNA template (10–100 ng)	50 ng
Forward primer	125 ng
Reverse primer	125 ng
dNTP mix	1 µl
QuickSolution reagent	1.5 µl
Milli Q water	Volume made up to 50 µl
PfuUltra HF DNA polymerase (2.5 U/µl)	1 µl

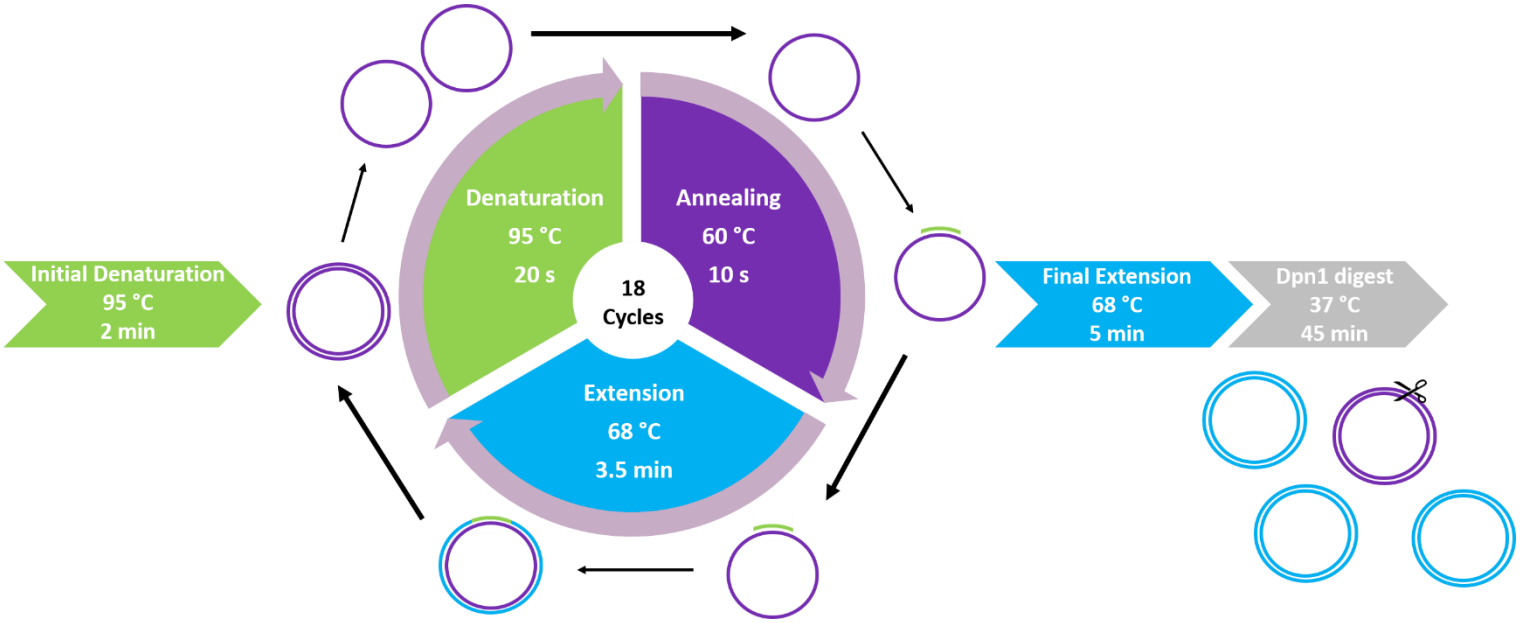


Figure 25: Schematic diagram indicating the PCR cycles used for all mutagenesis reactions.

Methylated WT template plasmid (purple) was denatured. Primers (green), encoding the mutation, bound to denatured template. The rest of the plasmid was built from the primer by polymerase, which resulted in the synthetic strand containing the mutation (blue). These new plasmids (not methylated) containing the mutation also acted as templates in subsequent PCR cycles. After 18 cycles of PCR Dpn1 was used to digest methylated plasmid (initial WT template) leaving a non–methylated product which contained the mutation.

3.2.1.3. TRANSFORMATION

The protein constructs obtained from GenScript (WT TBR1 T–Box and the WT FOXP2 FHD) were used to transform competent T7 *E. coli* cells (New England Biosciences) for expression, while the PCR products (C209A TBR1 T–Box and C274S TBR1 T–Box) were first used to transform ultracompetent XL–10 gold *E. coli* cells. These cells were chosen as they are ultra–competent (the transformation will likely work even if there is only a small portion of plasmid, which is likely when using a PCR product) and they contain a high copy number of the plasmid. All transformations were done using the following method.

Between 40 and 80 ng/µl of plasmid was added to 50 µl of competent cells which were thawed on ice. The cells were then incubated for 30 minutes on ice followed by heat shock at 42 °C for 45 seconds. Cells were placed on ice again for 2 minutes and 950 µl of super optimal broth with catabolite repression (SOC) media was added. SOC consists of 2% (w/v) tryptone, 0.5% (w/v) yeast, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$ and 20 mM glucose. The transformed cells were cultured at 37 °C for 1 hour while shaking at 220 rpm. The cells were subsequently grown overnight at 37 °C on agar plates containing 0.1 mg/ml of ampicillin for selection. Colonies were picked and cultured in 2X YT media

(1.6% (w/v) tryptone, 1% (w/v) yeast extract and 0.5% (w/v) NaCl) containing 0.1 mg/ml ampicillin at 37 °C for 16 hours with vigorous shaking. The culture was used to make up a 1:1 glycerol stock (cell culture: 80% (v/v) glycerol) and a mini prep was done to extract the plasmid (GeneJet plasmid mini prep kit). The isolated plasmids were sent for sequencing (Inqaba Biotec, RSA) to ensure the gene sequences were correct. The C209A TBR1 T–Box and C274S TBR1 T–Box mutant plasmids were subsequently used to transform T7 *E. coli* for expression once the sequence was confirmed using the above protocol.

3.2.1.4. EXPRESSION

T7 *E. coli* cells were chosen for protein overexpression as these cells are engineered for good regulation of target protein overexpression (**Figure 26**). This is achieved using the Lac operon to induce expression upon the addition of Isopropyl β–D–1–thiogalactopyranoside (IPTG).

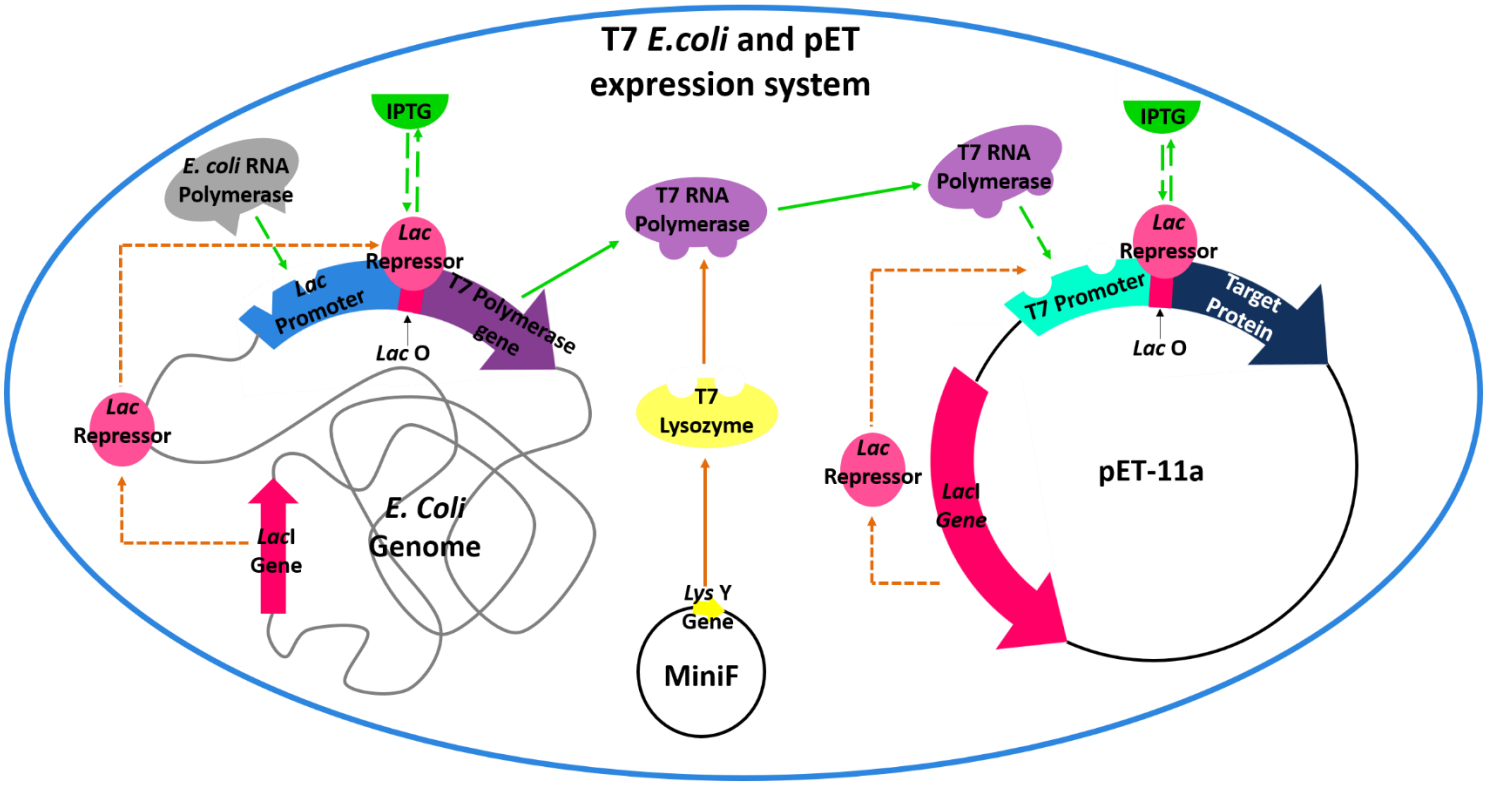


Figure 26: T7 *E. coli* and pET–11a expression system.

This expression system is highly regulated to prevent target protein expression before the desired time. The orange arrows represent the systems inhibiting target gene expression and the green arrows represent the systems activating target gene expression. The T7 *E. coli* genome contains the LacI gene (dark pink) which encodes the Lac Repressor (light pink). The Lac Repressor binds the Lac Promoter (light blue) and inhibits the *E. coli* RNA polymerase (grey) from transcribing the T7 RNA Polymerase gene (dark purple) downstream of the Lac promoter. When IPTG (green) is added it displaces the Lac repressor and *E. coli* RNA polymerase transcribes the T7 RNA polymerase gene, forming T7 RNA polymerase (light purple). The pET–11a plasmid also contains a LacI gene (dark pink) so the T7 RNA polymerase can only bind the T7 promoter (cyan) once the IPTG (green) has displaced the Lac repressor (light pink). T7 *E. coli* Lys Y cells also contain a miniF plasmid which encodes the Lys Y gene (yellow). The Lys Y gene encodes a T7 lysozyme (light yellow), which lacks amidase activity, that binds T7 RNA polymerase (light purple) and inhibits it. This is an extra regulatory measure that prevents any basal expression of the target gene. When IPTG is added, large quantities of T7 RNA polymerase are produced allowing high target gene expression after induction. (New England BioLab C3010I Datasheet).

The FHD and T–Box variants (WT and both mutants) were overexpressed using the same basic protocol as previously described (Blane & Fanucchi 2015), An overnight culture consisting of 50 ml 2X YT medium with 0.1 mg/ml of ampicillin was inoculated with 1 ml of glycerol stock and was grown up overnight at 37 °C. Larger volumes of 2X YT media with 0.1 mg/ml ampicillin were inoculated with a 1 in 50 dilution of overnight culture. Cells were grown at 37 °C with shaking until an O.D 600 between 0.6 and 0.7 was reached. The cultures were then cooled to 20 °C before IPTG was added to a

concentration of 0.5 mM (FHD) or 0.2 mM (T–Box variants). The cultures were incubated at 20 °C with shaking for 20 – 22 h. The cells were pelleted by centrifugation at 5000 xg for 10 minutes at 4 °C and the medium was discarded. The pellet was resuspended in 50 ml of equilibration buffer (500 mM NaCl, 20 mM Imidazole and 20 mM Tris–HCl pH 7.5) per litre of culture and frozen at –20 °C. When needed, cells were thawed at 20 °C and thereafter 1 mM phenylmethylsulfonyl fluoride (PMSF) and 0.1 mg/ml of lysozyme were added. The lysate was then sonicated (55% amplitude) 5 times for 30 seconds on ice. The lysate was left to incubate at 20 °C for 30 minutes after the addition of 0.01 mg/ml DNase 1. The lysate was then centrifuged at 18 000 xg for 25 minutes at 4 °C to separate the soluble and insoluble fractions. The supernatant was used in the subsequent purifications.

3.2.2 Purification

3.2.2.1. Immobilised Metal Affinity Chromatography

Immobilised metal affinity chromatography (IMAC) was the first purification step used in purification of all the proteins. In this technique, the recombinantly expressed protein, with a Hexahistidine tag (His– tag), specifically binds a column with either nickel or cobalt embedded (**Figure 27**). This is because the imidazole side chain of histidine has a high affinity to chelated metals. In this way, the His–tagged protein should interact with the column while the rest of the cell lysate should not form an interaction and therefore, elute from the column. At this point multiple washing steps, such as salt or detergent, were used to remove any DNA, proteins or lipids bound to the protein that was immobilised on the column. The pure protein was then eluted from the column with imidazole since imidazole competes with the His–tag to bind the nickel or cobalt column (Arnau *et al.* 2006).

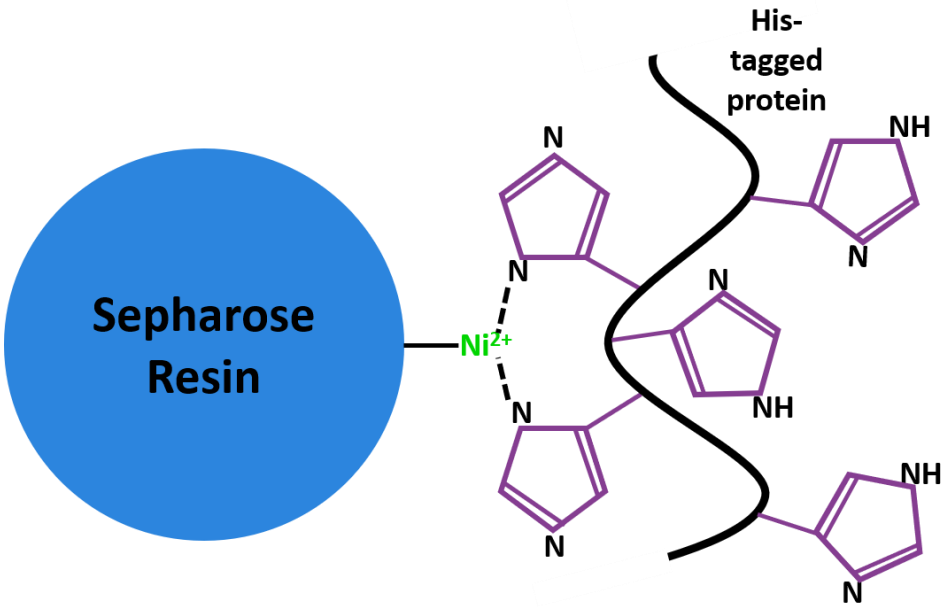


Figure 27: The interaction between the nickel charged IMAC column and the His–tag.

The column consists of a Sepharose resin (blue) charged with nickel (green). Each nickel ion has the potential to form two hydrogen bonds (black dotted line) with histidine residues (purple).

Each protein was purified on its own IMAC 5 ml column (1.6 X 2.5 cm HisTrap column (GE healthcare, USA)) charged with nickel (FHD) or cobalt (T–Box variants). The purification protocol for each protein was the same and as follows.

The column was equilibrated with 5 to 10 column volumes of equilibration buffer (20 mM Tris–HCl, 500 mM NaCl and 30 mM imidazole, pH 7.5). The supernatant was loaded on to the column and the flow through was collected. After the flow through passed through the column, the column was allowed to re–equilibrate with 5 to 10 column volumes of equilibration buffer. For the FHD purification, a salt wash buffer (20 mM Tris–HCl, 1.5 M NaCl and 30 mM imidazole, pH 7.5) was used to wash the column for at least 5 column volumes to remove any DNA bound to the protein. For purification of the T–Box variants, a detergent salt wash was done in the same manner using the detergent salt wash buffer (20 mM Tris–HCl, 1.5 M NaCl, 0.1% Tween 20 and 30 mM imidazole, pH 7.5). After the wash step, the column was re–equilibrated with 5 to 10 column volumes of equilibration buffer. The His–tagged target protein was then eluted using elution buffer (20 mM Tris–HCl, 500 mM NaCl and 300 mM imidazole, pH 7.5). Fractions of the various eluted peaks were collected along with samples of the supernatant, pellet, and flow through. All the fractions and samples collected were resolved on a 16% tricine gel (FHD) or a 12.5% glycine gel (T–Box variants) as described in section **3.2.2.4. SDS–PAGE**.

3.2.2.2. Thrombin Cleavage

The His–tag was subsequently cleaved off the FOXP2 FHD using thrombin since a thrombin cleavage site was incorporated into the recombinant FHD construct (**Figure 24**). Fractions from IMAC containing the FHD were pooled and dialysed against thrombin cleavage buffer (20 mM Tris–HCl, 2 mM KCl and 100 mM NaCl, pH 8). Protein was centrifuged at 10 000 xg for 15 minutes to remove aggregates. Approximately 1 unit of thrombin was added per mg of protein and the solution was left to incubate at 20 °C for 3 to 4 hours. Once cleavage was completed, the cleaved protein was separated from any tag, thrombin and uncleaved protein by filtering the solution through a 0.2 µM filter and loading it on a pre–equilibrated 1 ml (0.7 X 2.5 cm) benzamidine (GE healthcare, USA) and IMAC column connected in series. The thrombin binds to the benzamidine column while the His–tag binds to the IMAC thus the cleaved protein elutes independently (Young *et al.* 2012). The cleaved FHD was collected in fractions as it flowed through the columns. Any uncleaved protein was eluted from the IMAC column using elution buffer. Samples of all fractions collected were resolved on a 16% tricine gel. The tag was not cleaved off of the T–Box domain as the protein aggregated in the low salt environment required for active thrombin. The FOXP2 FHD was stored in equilibration at 4 °C until required.

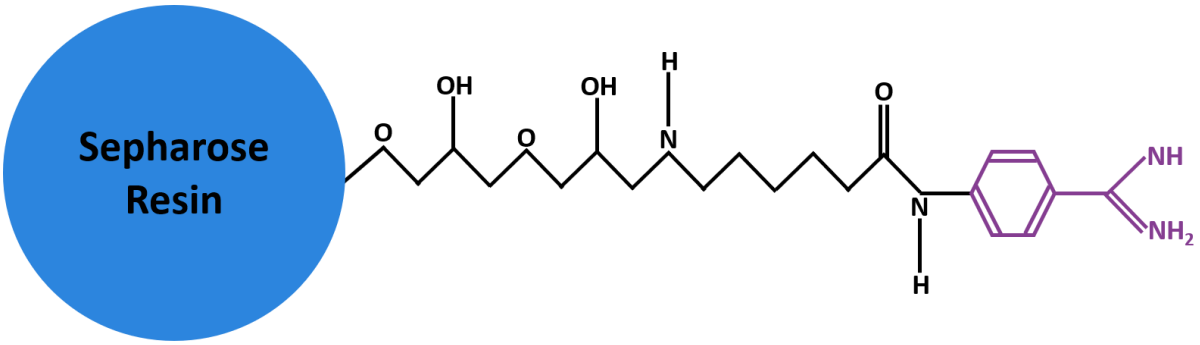


Figure 28: Benzamidine column resin.

A benzamidine molecule (purple) is attached to the Sepharose resin (blue) via a carbon chain (black). Benzamidine is an inhibitor for most serine proteases including thrombin. Thrombin binds the benzamidine and is trapped in the column allowing the protein with the cleaved off tag to elute (Terpe 2003).

3.2.2.3. Preparative Size Exclusion Chromatography

Size exclusion or gel filtration is a form of column chromatography in which molecules are separated based on their Stokes radius. This technique was originally described by Lathe & Ruthven, 1956. The Stokes radius is determined by a molecule's molecular mass and size. As molecules are passed through the column, molecules with a smaller Stokes radius may enter the pores which slows them down. Larger molecules that cannot fit through the pores, pass through the column quickly thus eluting faster. There are now

a variety of resins commonly used for size exclusion chromatography which include dextran (Sephadex/Sephacryl), agarose (Sepharose) and polyacrylamide polymers. Superdex resin is made up of a matrix consisting of agarose with covalently linked dextran chains. It is often used as it offers high resolution and is generally non–reactive with biological samples (Blurnenfeld & Gardner 1985). For each type of resin, the pore size is determined by adjusting the concentration of the cross–linking monomers which in turn allows the separation of macromolecules of various sizes. An advantage of using gel filtration is that it can be used to determine the size of molecules when used in conjunction with molecular weight markers (Blurnenfeld & Gardner 1985).

Size exclusion chromatography was used as a second purification step for the T–Box variants, as well as to observe the quaternary structure of each protein. A HiLoad™ 16/60 Superdex™ 75 prep grade column (GE healthcare, USA) was equilibrated with 1 to 2 column volumes of equilibration buffer (10 mM Tris–HCl, 500 mM NaCl and 30 mM imidazole, pH 7.5). Imidazole was included in the equilibration buffer as it tends to help prevent the protein from interacting with the column resin, resulting in sharper peaks. A maximum of 2 ml of protein was loaded on to the column at a time and all the peaks were collected in fractions. Samples were resolved on either a 12.5% glycine gel. The proteins were all stored in equilibration at 4 °C until required.

3.2.2.4. SDS–PAGE

Sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS–PAGE) is a versatile technique which is commonly used to estimate sizes of proteins and observe if there are other protein contaminants in a sample. In this technique, molecules are separated through a gel matrix via an electric charge. The protein samples are denatured (by heat) and the disulfide bonds are reduced (by β–mercaptoethanol). Additionally, the proteins are exposed to SDS (Sodium dodecyl sulfate), a denaturing anionic surfactant, which binds the protein and not only gives it a uniform charge but also forces it into a rod–like shape. This ensures that the samples are fully denatured and are thus separated by size and not by charge or shape as they are drawn through a polyacrylamide matrix by an electric current. There are multiple methods of SDS–PAGE, however, the glycine (Laemmli 1970) and tricine (Schägger 2006) are two of the most commonly used. These two methods differ in the buffering system (glycine–Tris and tricine–Tris) and subsequently the trailing ions. The p*K*a of the trailing ion affects its mobility and therefore, the rate at which it migrates in the gel. Tricine is ideal for separations of smaller proteins and peptides (Schägger & Von Jagow 1987). The FOXP2 FHD was resolved on 16% Tricine gels whereas all the T–Box domains were resolved on 12.5% glycine gels.

Samples of 50 µl were prepared for gels by diluting them 2X in glycine reducing sample buffer or 4X in tricine reducing sample buffer. Samples were subsequently boiled for 5

minutes, sonicated 3 times for 1 second and vortexed for 15 seconds to ensure the samples were fully denatured and thoroughly mixed.

The Bio–Rad Mini PROTEAN^{TM} Tetra Cell electrophoresis set was cleaned, and the casting stand assembled. First the separating gel was prepared according to **Table 8** and applied in between the glass plates and allowed to polymerise. The stacking gel was subsequently prepared according to **Table 8** and applied on top of the separating gel and allowed to polymerise with the comb in.

Once the gels were polymerised, the gels were placed in the tanks with the appropriate buffers (see **Table 8**). Samples volumes of 10 µl were loaded into the wells and the gels were initially run at 50 V until the samples reached the separating gel. Then gels were then run at 160 V for approximately 1 to 2 hours until the dye front reached the bottom of the gel. The gels were removed and stained for 1 hour with Coomassie stain solution (0.1% (w/v) Coomassie dye in 1:5:4 (v/v/v) acetic acid–methanol–water) and then de–stained overnight with de–stain solution (1:5:4 (v/v/v) acetic acid–methanol–water). Additional de–staining steps were repeated as necessary.

Table 8: Preparation of both tricine and glycine sodium dodecyl sulfate–polyacrylamide gel electrophoresis gels and their respective running buffers.

Tricine				Glycine			
Reagent Stocks		**16% Separating Gel**	**4% Stacking Gel**	**Reagent Stocks**		**12.5% Separating Gel**	**4% Stacking Gel**
Monomer Solution	Acrylamide/Bis–Acrylamide Stock (49.5% T and 3% C)	5 ml	0.34 ml	Monomer Solution	Acrylamide/Bis–Acrylamide Stock (30% T and 2.7% C)	6.25 ml	0.94 ml
3X Gel Buffer	3 M Tris, 1 M HCl and 0.3% SDS, pH 8.45	5 ml	1 ml	Separating Gel Buffer	1.5 M Tris–HCl, pH 8.8	3.75 ml	
Glycerol	80% glycerol	1.5 ml		Stacking Gel Buffer	500 mM Tris–HCl, pH 6.8		1.75 ml
dH_2O		3.5 ml	2.7 ml	SDS	10% (w/v)	150 µl	70 µl
Ammonium per sulphate	10% (w/v)	75 µl	34 µl	dH_2O		4.75 ml	4.3 ml
TEMED		25 µl	7 µl	Ammonium per sulphate	10% (w/v)	75 µl	35 µl
Cathode Buffer	1 M Tris, 1 M Tricine and 1% SDS, ~pH 8.25 (pH not adjusted)			TEMED		7.5 µl	15 µl
Anode Buffer	1 M Tris, pH 8.9 (pH adjusted with HCl)			Tank Buffer	250 mM Tris–HCl, 192 mM glycine, 0.1% (w/v) SDS, pH 8.3		
Reducing Sample Buffer	12% (w/v) SDS, 6% (w/v) β–mercaptoethanol, 30% (w/v) glycerol, 0.005% Coomassie blue G–250 and 150 mM Tris–HCl, pH 7.5			Reducing Sample Buffer	125 mM Tris–HCl, 4% (w/v) SDS, 20% (v/v) glycerol, 10% (v/v) β–mercaptoethanol, pH 6.8		

This table shows how two tricine and glycine gels were prepared according to (Schägger 2006) and (Laemmli 1970) respectively.

3.2.2.5. Protein Purity and Concentration

Once each protein was found to be sufficiently pure of other protein contaminants based on the SDS–PAGE gels, each protein was dialysed against 10 mM sodium phosphate, 100 mM NaCl, pH 7.4. The proteins were removed from dialysis and centrifuged at 4 °C at 10 000 xg for 10 minutes to remove any aggregates. Either a JASCO V–630 spectrophotometer or a Thermo Scientific™ NanoDrop™ One Spectrophotometer was used to take an absorbance spectrum from 240 to 350 nm in triplicate. The 340 nm reading was subtracted from the 280 nm reading to account for aggregation. The Beer–Lambert law (**Equation 1**) was used to calculate the protein concentration. The theoretical extinction coefficients obtained from ExPASy ProtParam (Gasteiger *et al.* 2005) that were used for each protein are displayed in **Table 9**. The extinction coefficients are calculated using **Equation 2** based on the primary structure using extinction coefficients for tyrosine, tryptophan and cysteine in water (Gasteiger *et al.* 2005). The A_{280}/A_{260} ratio was used to detect the presence of any nucleic acid contamination.

Equation 1: $A = \varepsilon.c.l$

Where A is the absorbance at a specified wavelength (A.U), ε is the extinction coefficient at wavelength ($M^{-1}.cm^{-1}$), c is the concentration of protein (M) and l is the path length (cm).

Table 9: The theoretical extinction coefficients for each protein

Protein	ε_{280} reduced ($M^{-1}.cm^{-1}$)	ε_{280} oxidised ($M^{-1}.cm^{-1}$)
FOXP2 FHD	22 460	22 460
WT TBR1 T–Box	36 440	36 565
C209A TBR1 T–Box	36 440	36 440
C274S TBR1 T–Box	36 440	36 440

The extinction coefficients for the reduced proteins at 280 nm are displayed and were obtained from ExPASy ProtParam (Gasteiger *et al.* 2005) using **Equation 2**. The reduced extinction coefficient was used when determining protein concentrations. The oxidised extinction coefficient assumes that all pairs of Cys residues form cysteines.

Equation 2: $\varepsilon(prot) = n(Tyr) \times \varepsilon(Tyr) + n(Trp) \times \varepsilon(Trp) + n(Cys) \times \varepsilon(Cys)$

Where n refers to the number of amino acid residues, ε is the molar extinction coefficient of the amino acid and $(prot)$ refers to the predicted molar extinction coefficient of the protein.

3.2.3 PROTEIN LABELLING

For fluorescence anisotropy (FA) and single molecule Förster resonance energy transfer (smFRET), protein conjugated to a fluorophore is necessary for detection. For FA studies to observe the interaction between T–Box and FHD, T–Box was labelled with Atto 550 via the His–tag. For smFRET studies of the interaction between T–Box and DNA, T–Box was labelled via its cysteine residues with Alexa Fluor 647. Atto 550 was used for FA rather than Alexa Fluor 647 as the fluorescence lifetime of Atto 550 is ideal for anisotropy measurements whereas the fluorescence lifetime of Alexa Fluor 647 is too short (**Table 10**). Alexa Fluor was used for smFRET as it forms a good FRET pair with Cy3 (used to label the DNA). Cy3 was used to label the DNA as the excitation laser used in smFRET was the correct wavelength to excite Cy3. The protocols used to label T–Box with each dye are described in the sections to follow and the properties of each dye are displayed below in **Table 10**.

Table 10: Properties of the various fluorophores used for protein labelling

	Conjugation method	Ɛ at λmax ($M^{-1}.cm^{-1}$)	λmax: absorption (nm)	λmax: emission (nm)	Correction Factor	Molecular weight (Da)	Fluorescence lifetime (ns)
Alex Fluor™ 647	Maleimide	265 000	651	671	0.03	~1250	~1
NTA–Atto	NTA	120 000	554	576	0.1		~3.6

Ɛ at λmax is the extinction coefficient of the fluorophore at the wavelength of its maximum absorbance. λmax: absorption is the wavelength of its maximum absorbance (nm). λmax: emission is the wavelength of its maximum emission after excitation at the wavelength of its maximum absorbance. The correction factor is a constant used to account for any spectra overlap of the dye fluorophore at 280 nm (the wavelength used to calculate protein concentration).

Once T–Box was labelled with each dye, the protein concentrations were calculated using and labelling efficiencies **Equation 3** and **Equation 4**.

Equation 3: $$[Protein] = (A_{280} - (CF \times A_{\lambda\varepsilon}))/\varepsilon_{Protein}$$

Where A_{280} is the absorbance reading at 280 nm, CF is the dye correction factor, $A_{\lambda\varepsilon}$ is the absorbance at the wavelength of the absorbance peak of the fluorophore and $\varepsilon_{Protein}$ is the extinction coefficient of the protein at 280 nm.

Equation 4: $$Labelling\ efficiency\ (\%) = [Dye]/[Protein] \times 100$$

3.2.3.1. NTA–ATTO 550

The WT TBR1 T–Box domain was labelled via the His–tag with Atto 550–Nitrilotriacetic acid, complexed to Ni^{2+} (NTA–Atto) (Sigma– Aldrich, SA) for fluorescence anisotropy.

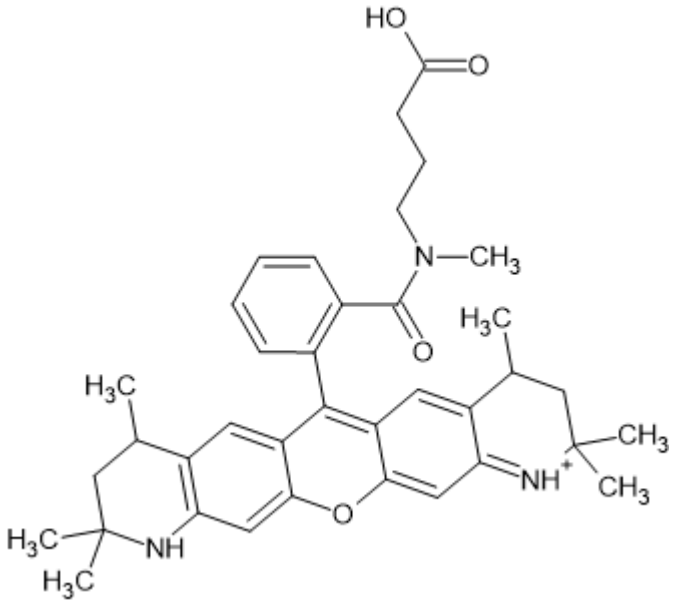


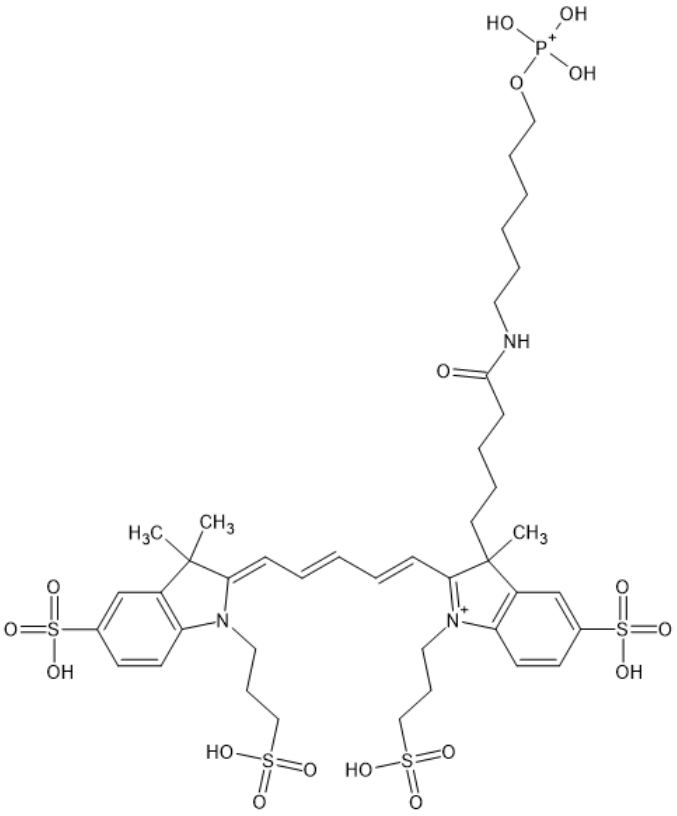
Figure 29: The structure of Atto 550

The structures of Atto 550. Chemical structures generated using ACD ChemSketch 2019.2.0.

The purified WT TBR1 T–Box was dialysed against 10 mM sodium phosphate, pH 7.4 containing either 100 or 500 mM NaCl. The protein was then dialysed against a final 500 ml of buffer containing 10 mM sodium phosphate, pH 7.5, with either 100 or 500 mM NaCl and 3 mM Tris (2–carboxyethyl) phosphine (TCEP). After dialysis, the protein was centrifuged at 10 000 xg for 10 minutes at 4 °C and the solution was filtered with a 0.2 µM filter. The concentration was checked as described previously. Equimolar NTA–Atto was added to the freshly dialysed WT TBR1 T–Box and was left to incubate, protected from light, at 20 °C for 1 hour. Any unbound dye was removed using the Pierce™ Dye Removal Columns (ThermoFisher Scientific, USA) according to the manufacturer's instructions. The concentration of the protein and the dye were checked again using a Thermo Scientific™ NanoDrop™ One Spectrophotometer. The concentration of the protein was determined using **Equation 3**, where the CF was 0.1 and $A_{\lambda\varepsilon}$ was 554 nm. The concentration of the NTA–Atto 550 was determined using the Beer–Lambert law **Equation 1**, where the wavelength was 554 nm and the extinction coefficient was 120 000 ($M^{-1}.cm^{-1}$) (**Table 10**). The labelling efficiency was subsequently determined using **Equation 4**.

3.2.3.2. ALEXA FLUOR™ MALEIMIDE

WT TBR1 T–Box was labelled with Alex Fluor™ 647 (**Figure 30**) for smFRET rather than the cystine mutants (C209A and C274S) as a poor labelling efficiency was obtained for the mutants.



Alex Fluor 647

Figure 30: The structures of Alex Fluor 647.

The structure of Alex Fluor 647 is displayed above. Chemical structure generated using ACD ChemSketch 2019.2.0.

The Alexa Fluor 647 fluorophore was conjugated to a maleimide group in order to label the protein via thiol groups located on Cys residues. The reaction is depicted in **Figure 31** below.

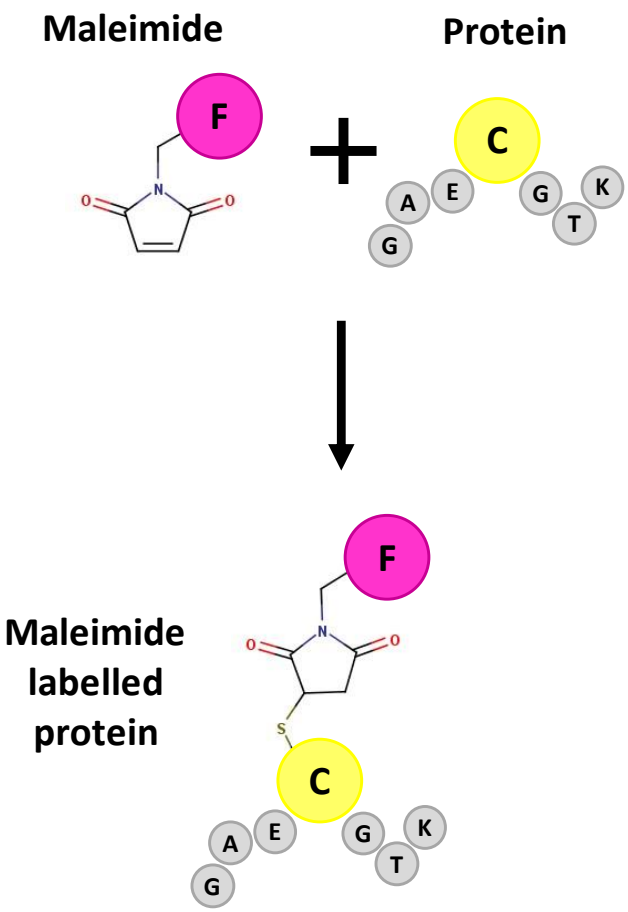


Figure 31: The labelling reaction of a maleimide conjugated fluorophore to a Cys residue.

The position of the fluorophore is indicated by the pink circle. The Cys residue is represented by the yellow circle on a polypeptide chain. Chemical structures generated using ACD ChemSketch 2019.2.0.

The WT TBR1 T–Box was dialysed against 10 mM HEPES, 500 mM NaCl, pH 7.5. After dialysis, the protein was centrifuged at 10 000 xg for 10 minutes at 4 °C and the solution was filtered with a 0.2 µM filter. A 50 µl labelling reaction was set up consisting of 2.5 nmol of WT T–Box, 20X excess Alex Fluor maleimide conjugate and 500 µM TCEP. The reaction was left at 4 °C overnight protected from light. Any unbound dye was removed using the Pierce™ Dye Removal Columns (ThermoFisher Scientific, USA) according to the manufacturer's instructions. The concentration of the protein and the dye were checked again using a Thermo Scientific™ NanoDrop™ One Spectrophotometer. The concentration of the protein was determined using **Equation 3** and the CF and absorbance wavelengths from **Table 10**. The concentration of the Alexa Fluor was determined using the Beer–Lambert law (**Equation 1)** and the extinction coefficient and absorbance wavelengths from **Table 10**. The labelling efficiency was subsequently determined using **Equation 4**.

3.2.4 DNA PREPARATION

3.2.4.1. DNA SEQUENCES

The FOX family of proteins has a defined DNA consensus sequence which is 5'–[A/G][C/T][A/C]AA[C/T]A–3' (Carlsson & Mahlapuu 2002). The core binding sequence specifically for FOXP2 was defined as 5'–A[C/T]AAATA–3' (Wang *et al.* 2003). The FOXP2 2A07 crystal structure (Stroud *et al.* 2006) contains the Wang sequence 5'–AACTATGAAACAAATTTTCCT–3'. More recently, however, Nelson *et al.* (2013) found the sequence 5'–TGTTTAC–3' to which FOXP2 binds with a higher affinity. In this study, DNA containing the Nelson binding site 5'–AGGTGTTTACTTTCATAG–3' was used for FOXP2 DNA binding studies and termed Nelson DNA (Nel). Additionally, a second DNA sequence containing the Nelson binding site was used. This sequence was termed Nel C as the flanking regions were obtained from the promoter region of the CNTNAP2 gene, a gene FOXP2 is known to regulate.

All proteins in the T–Box family identified thus far, are capable of binding a consensus DNA sequence 5'–TCACACCT–'3 (Wilson & Conlon 2002). There are crystal structures of various T–Box domains bound to a palindromic DNA sequence 5'–AATTTCACACCT*AGGTGTGA*AATT–3'. This sequence has two DNA binding sites orientated in a perfect palindrome and was determined *in vitro* by nucleotide selection (Muller & Herrmann 1997). TBX5 has also been crystallised in complex with DNA containing only a single binding site (Stirnimann *et al.* 2010). In this study T–Box palindrome DNA (Pal) with the sequence 5'–ATTTCACACCT*AGGTGTGA*AAT–3' and T–Box single site long DNA (SSL) 5'–CCCAATTTCACACCTTCCTCA–3' were used for DNA binding studies involving T–Box proteins. Furthermore, DNA from the Adeno–associated virus type 2 P5 promoter (5'–AGGGTCTCCATTTTGAAGCG–3') was used in EMSA to ensure that the proteins are binding DNA specifically. DNA binding of the FOXP2 FHD to DNA is well characterised (Morris *et al.* 2018; Morris & Fanucchi 2016; Stroud *et al.* 2006; Webb *et al.* 2017), so FHD DNA binding studies were only conducted to confirm protein integrity and for comparative purposes. T–Box DNA binding was the focus of the DNA binding section as no work has been done on TBR1 T–Box DNA binding *in vitro*.

Table 11 displays all of the DNA sequences used in this study. For EMSA and ITC, unlabelled DNA was used. However, for FA, Nel C and SSL DNA labelled with ROX was required. For smFRET, DNA labelled with both Cy3 and biotin was used. Unlabelled DNA and ROX labelled DNA were purchased as purified duplex DNA. The Cy3 labelled and biotinylated DNA, however, were purchased as single stranded oligos and were subsequently hybridised and purified. All DNA labelled with any fluorophore was always protected from light. All DNA was purchased from Integrated DNA technologies as a lyophilised pellet.

Table 11: The various DNA sequences used in these studies.

DNA	Sequence	Number of binding sites	Origin	Number of base pairs	Known Binding site for	Extinction Coefficient at 260 nM ($M^{-1}cm^{-1}$)	Techniques	Labels
T–Box Palindrome (Pal)	5'–ATTTCACACCT*AGGTGTGAA*AT–3' 3'–TA*AAGTGTGGA*TCCACACTTTA–5'	2	T–Box Crystal structures	22	T–Box domains	383828	EMSA ITC FA smFRET	Unlabeled Unlabeled Unlabeled and ROX Cy3, Biotin
T–Box Single Site Long (SSL)	5'–CCCAATTTCACACCTTCCTCA–3' 3'–GGGTTAAAGTGTGGAAGGAGT–5'	1	Auts2 Promoter	21	T–Box domains	342046	EMSA ITC FA smFRET	Unlabeled Unlabeled Unlabeled and ROX Cy3, Biotin
T–Box Single Site Short (SSS)	5'–TTTCACACCTT–3' 3'–AAAGTGTGGAA–5'	1	Crystal structure of	11	T–Box domains	176984	EMSA	Unlabeled
Nelson (Nel)	5'–AGGTGTTTACTTTCATAG–3' 3'–TCCACAAATGAAAGTATC–5'	1	microfluidic chip analysis	18	FOXP2 FHD	256016	EMSA	Unlabeled
Nelson CNTNAP2 (Nel C)	5'–TTTGATTGTTTACTTTGTTC–3' 3'–AAACTAACAAATGAAACAAG–5'	1	CNTNAP2 promoter	20	FOXP2 FHD	310960	EMSA FA	Unlabeled Unlabeled and ROX
Nelson short (Nel S)	5'–GGTGTTTACTT–3' 3'–CCACAAATGAA–5'	1	Shorter version of Nelson	11	FOXP2 FHD	177962	EMSA	Unlabeled
AAV2 P5 Promoter (AAVP5)	5'–AGGGTCTCCATTTTGAAGCG–3' 3'–TCCCAGAGGTAAAACTTCGC–5'	1	Adeno–associated virus type 2 P5 promoter	20	Yin–Yang 1	325715	EMSA	Unlabeled

The protein for which the DNA contains a known binding site for is shown in the last column. The techniques each DNA sequence was used for and whether it was labelled or not are displayed in the last two columns

3.2.4.2. Hybridisation and Purification of DNA for smFRET

The forward strands of DNA containing T–Box binding sequences (SSL and Pal) were ordered from Integrated DNA technologies with Cy3 conjugated to the 3' ends. The corresponding Pal and SSL reverse oligos were ordered with and without biotin conjugated to the 3' ends. By hybridising the Cy3 labelled forward strands with the corresponding biotinylated and non–biotinylated reverse strands, biotinylated and non–biotinylated Cy3 labelled SSL and Pal DNA was produced (**Figure 32**).

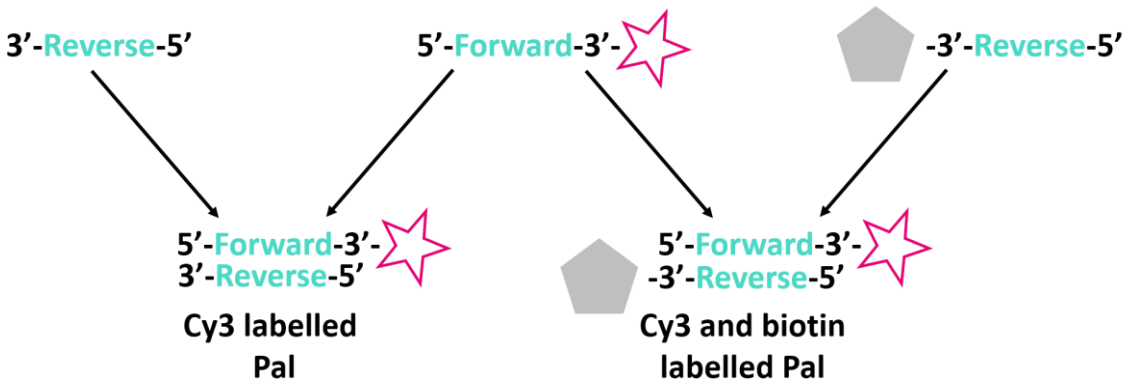


Figure 32: Schematic showing the hybridisation of single stranded DNA oligos to obtain the Cy3 and biotinylated duplex DNA required for smFRET.

Single stranded forward strands of SSL and Pal DNA were conjugated to Cy3 (pink star). Single stranded reverse strands of SSL and Pal DNA were obtained with and without biotin (grey pentagon). The forward strands were hybridised to both the biotinylated and non–biotinylated reverse strands. As a result, both DNA labelled with only Cy3 and DNA labelled with Cy3 and biotin were obtained.

The DNA oligos were purchased as lyophilised pellets from Integrated DNA technologies. The DNA oligos were resuspended in in 50 mM Tris–HCl pH 7.5 at ~ 500 µM. The various combinations of oligos were annealed by setting up 50 µl reactions containing 20 µM forward strands, 20 µM reverse strands, 100 mM NaCl and 50 mM Tris–HCl, pH 7.5. The reactions were incubated at 90 °C for 5 minutes in a water bath. The water bath was switched off and the reactions were left to slowly cool in the water bath overnight. The samples were resolved on a 10% (w/v) polyacrylamide gel containing 1X TBE buffer and 100 mM NaCl. The tank buffer consisted of 1X TBE buffer and 100 mM NaCl. The samples were prepared by the addition of 3% (v/v) ficoll. The gel was first run at 180 V for about 10 minutes (until the samples had entered the wells) and thereafter at 120 V for an additional 4 hours. The gel was removed, and the bands were cut from the gel and placed in separate 1.5 ml sample tubes. The gel pieces in each sample tube were mashed and enough buffer (50 mM Tris–HCl, pH 7.5) was added to cover the gel fragments. The tubes were left on a rotator overnight at 4 °C to allow the DNA to diffuse from the gel. The buffer containing the DNA was removed and placed in new sample tubes. The DNA was precipitated by the addition of 3 M acetic acid and 100% (v/v) chilled ethanol in a

ratio of 1 DNA: 0.1 acetic acid: 2.5 ethanol. The samples were incubated at –20 °C overnight. The tubes were then centrifuged at 4 °C for 10 minutes at 12 000 xg and the supernatant was immediately discarded. The pellet was resuspended in 50 mM Tris–HCl, pH 7.5 and the concentration was calculated using the theoretical extinction coefficients in **Table 11** and **Equation 1**.

3.2.4.3. Resuspension of Purchased Duplex DNA

Purchased duplex DNA was resuspended in filtered and autoclaved MilliQ water at a concentration of 100 µM with the exception of DNA for ITC. The DNA for ITC was resuspended at a concentration of 1000 µM to prevent buffer mismatch during ITC titrations. All DNA was stored at –20 °C when not in use and multiple freeze thaw cycles were avoided.

3.2.5 Structural Characterisation

3.2.5.1. Quaternary Structure

Size exclusion chromatography was repeated with each protein individually. It was used not only as a preparative step but also analytically in order to observe the quaternary structure of the FHD and T–Box variants. This was important to ensure that each protein existed predominantly as a single oligomeric species, because a mixture would convolute data analysis in other techniques. Furthermore, it was unknown whether or not the TBR1 T–Box domain exists as a dimer in solution.

A HiLoad™ 16/60 Superdex™ 75 prep grade column (GE healthcare, USA) was equilibrated with IMAC equilibration buffer (10 mM Tris–HCl, 500 mM NaCl and 30 mM imidazole, pH 7.5). One ml of protein at concentrations of 25 and 50 µM was loaded on to the column in the presence and absence of 2 mM DTT. Protein standards (GE healthcare low range gel filtration calibration kit) were separated according to the manufacturer's instructions in equilibration buffer. The standards consisted of conalbumin (75 kDa), ovalbumin (44 kDa), carbonic anhydrase (29 kDa), ribonuclease A (13.7 kDa) and aprotinin (6.5 kDa). A standard curve was used to approximate the size of the protein in each peak. It was constructed by plotting the log Mw of each standard against the ratio of their respective elution volumes (Ve) and void volume (Vo). A linear regression was fitted to the curve and the subsequent equation was used to calculate the sizes of the proteins based on their elution volumes.

3.2.5.2. Tertiary Structure

Fluorescence spectroscopy makes use of the fluorescence phenomenon where a molecule (the fluorophore) absorbs light at a specific wavelength. This light excites electrons to a higher energy level and light is then emitted when the electrons move back to their ground state (**Figure 33**). The emitted light is of a lower frequency/ longer wavelength than the light used to excite the fluorophore (Lakowicz 2006).

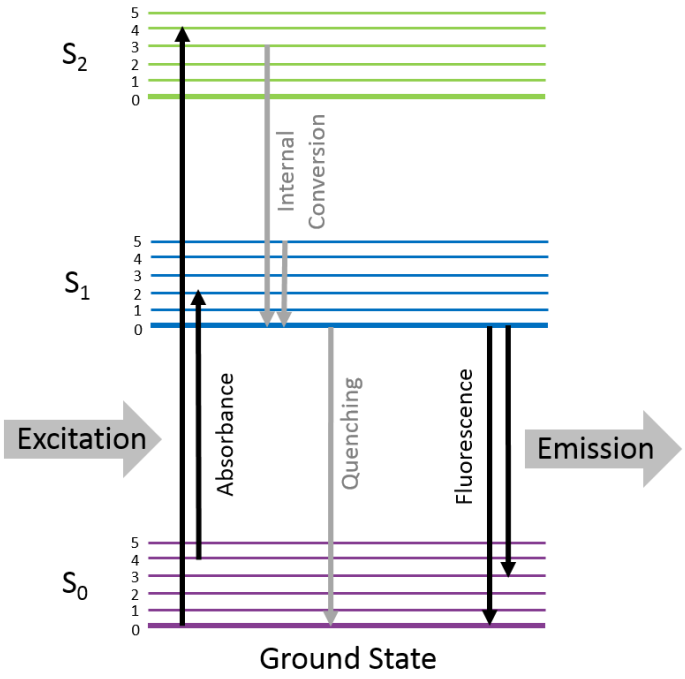


Figure 33: Jablonski diagram showing the fluorescence phenomenon.

S0, S1 and S2 are energy levels and the lines indicate different vibrational energy levels. Light of a specific wavelength is used to excite electrons and the light is absorbed. Electrons fall back to the lowest vibrational energy level with energy loss due to heat and vibration (internal conversions). As the electron drops back down to the ground state the energy can either be emitted via light (fluorescence) or lost due to quenching (Lakowicz, 2006).

Proteins may contain amino acids which act as fluorophores (tryptophan, tyrosine, phenylalanine and cysteine). Cysteine and phenylalanine exhibit extremely weak signals and are thus not used as intrinsic fluorophores. Tryptophan is most often used as it has a good signal and reduces the quantum yield of nearby tyrosine residues. Both tyrosine and tryptophan are excited at 280 nm and tryptophan can be exclusively excited at 295 nm. These intrinsic fluorophores can be used to probe local tertiary structural changes in the protein. This is due to the sensitivity of the fluorophore to the polarity of its environment which alters the emission spectrum. A hypsochromic or blue shift indicates a decrease in the polarity which often indicates burying of the fluorophore within the protein interior and thus reducing its exposure to the polar solvent. A bathochromic or red shift indicates an increase in polarity which occurs when the fluorophore becomes more exposed to the polar solvent (Lakowicz 2006).

Intrinsic fluorescence of the FHD and all the T–Box variants was used to observe the differences in spectra between native and denatured protein in order to deduce if the native protein is folded or not. The spectra of native WT T–Box was compared to the mutants to determine if there are any tertiary structural changes caused by the mutations.

Fluorescence spectra of 2 µM protein in 10 mM sodium phosphate, 100 mM NaCl, 2 mM TCEP at pH 7.5 were determined in triplicate using a JASCO FP–6300 spectrofluorometer. Both 295 and 280 nm were used to excite the fluorophores (tryptophan at 295 nm and tryptophan and tyrosine at 280 nm) and the emission spectra were recorded from 280 to 450 nm with a scanning speed of 500 nm/min. The excitation and emission band widths were set at 5 and 2.5 nm, respectively.

3.2.6 Binding Studies

3.2.6.1. Electrophoretic Mobility Shift Assays

Electrophoretic mobility shift assay (EMSA) is an electrophoresis–based technique that is used to observed protein–DNA interactions. After allowing time for the protein to bind DNA, the solution is run on a 10% polyacrylamide gel. Protein–DNA complexes are larger than the free DNA and thus are retarded by the gel. This should result in the formation of two distinct bands visible on the gel after staining with a nucleic acid stain as depicted in **Figure 34** (Garner & Revzin 1981). EMSAs, however, are often of low resolution because of the smearing of bands. These smears are a result of dissociation and re–association of the complex as the gel is run as equilibrium is constantly shifted due to the separation of the various species (Sidorova *et al.* 2010). The technique, however, is suitable to indicate whether or not the complex associates and may give an indication of the dissociation constant for the binding interaction. EMSAs, however, are low resolution techniques and are thus only usually used to obtained qualitative data (Sidorova *et al.* 2010).

Firstly, EMSAs were used to assess if each protein binds DNA containing its cognate binding site. This indicates whether or not the protein has adopted its correct DNA binding fold and is active as a result. Secondly, EMSAs were used to determine which sequences each protein could bind. This was important to ensure that WT FOXP2 FHD and WT TBR1 T–Box domain bind their own DNA sequences specifically rather than binding any DNA sequence.

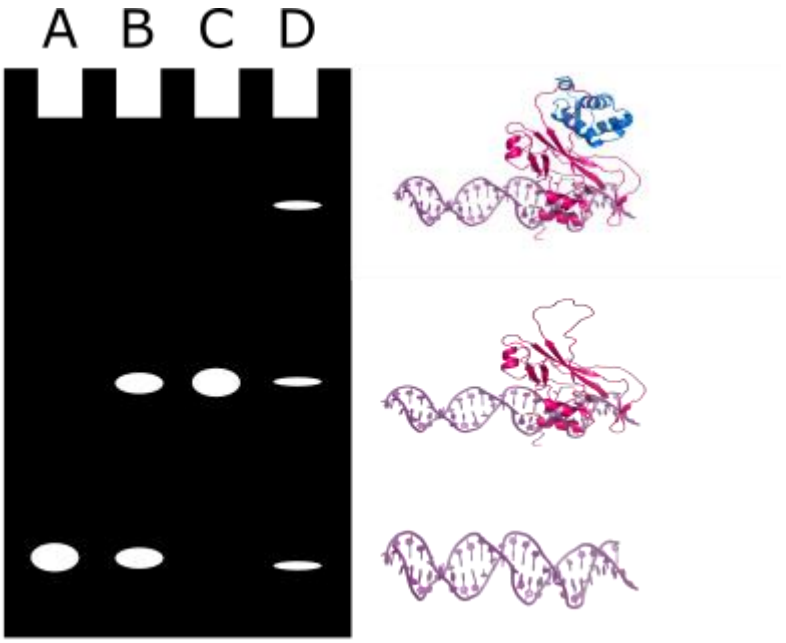


Figure 34: Diagram depicting an electrophoretic mobility shift assay (EMSA) and a super shift EMSA.

An EMSA is a technique where the interaction of DNA and a protein is observed by running a sample containing both DNA and the protein in an acrylamide gel. Once the gel has run, the DNA is stained by a fluorescent intercalating agent and observed under a UV light. Lane A contains DNA only and runs as a single band near the bottom of the gel. In lane B protein is added and a higher up band consisting of the protein DNA complex forms. In lane 3, when a saturating concentration of protein is added all the DNA is in complex and the free DNA band disappears all together. In Lane D, a second binding partner is added, and a super shift occurs as the DNA, protein A and protein B form the largest complex. The protein structure represented in pink and the DNA is from the crystal structure of the T domain from *Xenopus Laevis* bound to DNA (1XBR) (Muller & Herrmann 1997) and the protein represented in blue is the FOXP2 FHD (2A07) (Stroud *et al.* 2006). Image created using Inkscape 0.92.4

Firstly, traditional EMSAs were performed with the WT FHD and WT T–Box using various DNA sequences including the SSL, SSS, Pal, Nel, Nel S, Nel C and AAVP5 DNA sequences (**Table 11**). EMSAs on the T–Box mutants were only performed with the T–Box DNA (Pal and SSL). Reactions of 50 µl were set up containing increasing protein concentration from 0 to 20 µM (0 to 40X in excess of DNA), 0.5 µM DNA in EMSA binding buffer (10 mM HEPES, 100 mM KCl, 1 mM Ethylenediaminetetraacetic acid (EDTA), 1 mM $MgCl_2$, 0.1 mg/ml bovine serum albumin, 10% (v/v) glycerol, pH 7.5) and MilliQ water. The samples were left to incubate at 4 °C for at least 30 minutes. An 8% polyacrylamide gel was prepared with 1X Tris–HCl Borate EDTA (0.089 M Tris, 0.089 M Borate and 2 mM EDTA, pH 8.2 to 8.4) buffer and 20% triethylene glycol using the Bio–Rad Mini PROTEAN™ Tetra Cell (8.3 x 7.3 cm gels). The gel along with the running buffer (TBE buffer) was left to equilibrate to 4 °C for 1 hour. The gel was run at 4 °C for 1.5 hours immediately after loading the samples. Once the electrophoretic run was over the gel

was removed and stained with a 1:10 000 dilution of SYBR gold in TBE buffer for 5 minutes. The gels were visualised under UV light using the Bio–Rad gel doc system.

Secondly, super–shift EMSAs were performed to observe if TBR1 T–Box could bind both DNA and FOXP2 FHD simultaneously. A super shift–EMSA applies the same principles as a classic EMSA, however, an additional binding partner, such as an antibody, is included. This means that when the third binding partner binds the original complex (Protein A and DNA), the complex will appear as a band shifted higher on the gel as seen in lane D in **Figure 34** (Buratowski & Chodosh 2001; Pares– Matos 2014). A super shift–EMSA was used in this case to determine if T–Box can bind both the FHD and DNA simultaneously.

A 16 cm X 16 cm 8% continuous EMSA (as described above) was prepared using the Bio–Rad PROTEAN™ II xi Cell. Reactions of 50 µl were set up containing increasing concentrations of each protein as displayed in **Table 12**. The samples also contained 0.5 µM DNA (Either Pal or SSL) and EMSA binding buffer (10 mM HEPES, 100 mM KCl, 1 mM Ethylenediaminetetraacetic acid (EDTA), 1 mM $MgCl_2$, 0.1 mg/ml bovine serum albumin, 10% (v/v) glycerol, pH 7.5). The samples were left to incubate overnight at 4 °C along with the running buffer (1X TBE buffer). The entire 50 µl samples were loaded and the gel was run at 4°C for 10 hours. The gel was stained with a 1:10 000 dilution of SYBR gold in TBE buffer for 5 minutes. The gels were visualised under UV light using the Bio–Rad gel doc system.

Table 12: The concentrations of DNA, FOXP2 FHD and TBR1 T–Box used in the super–shift electrophoretic mobility shift assays

Lane→	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
DNA (µM)	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
FHD (µM)	0	0	0	0	2.5	2.5	2.5	2.5	5	5	5	5	10	10	10
T–Box (µM)	0	1	2.5	5	0	1	2.5	5	0	1	2.5	5	0	1	2.5

3.2.6.2. Isothermal Titration Calorimetry

Isothermal titration calorimetry (ITC) is a versatile technique that measures the heat absorbed or released when one reactant is titrated into another. The isotherm obtained can be used to determine thermodynamic properties such as the binding affinity (K_a), enthalpy, entropy, and free energy. The K_a and enthalpy (ΔH) can be derived directly from the titration (**Figure 35**). This can be used to calculate the Gibbs free energy (ΔG) and entropy (ΔS) (Ladbury & Chowdhry 1996) using **Equation 5** and **Equation 6**.

Equation 5: $\Delta G = RTln\,(K_a)$

Where ΔG refers to Gibbs free energy (kJ/mol), R refers to the gas constant, T refers to temperature (K) and K_a refers to the association constant.

Equation 6: $\Delta G = \Delta H - T\Delta S$

Where ΔG refers to Gibbs free energy (kJ/mol), ΔH refers to the enthalpy (kJ/mol), T refers to the temperature (K) and ΔS refers to the entropy (kJ/mol).

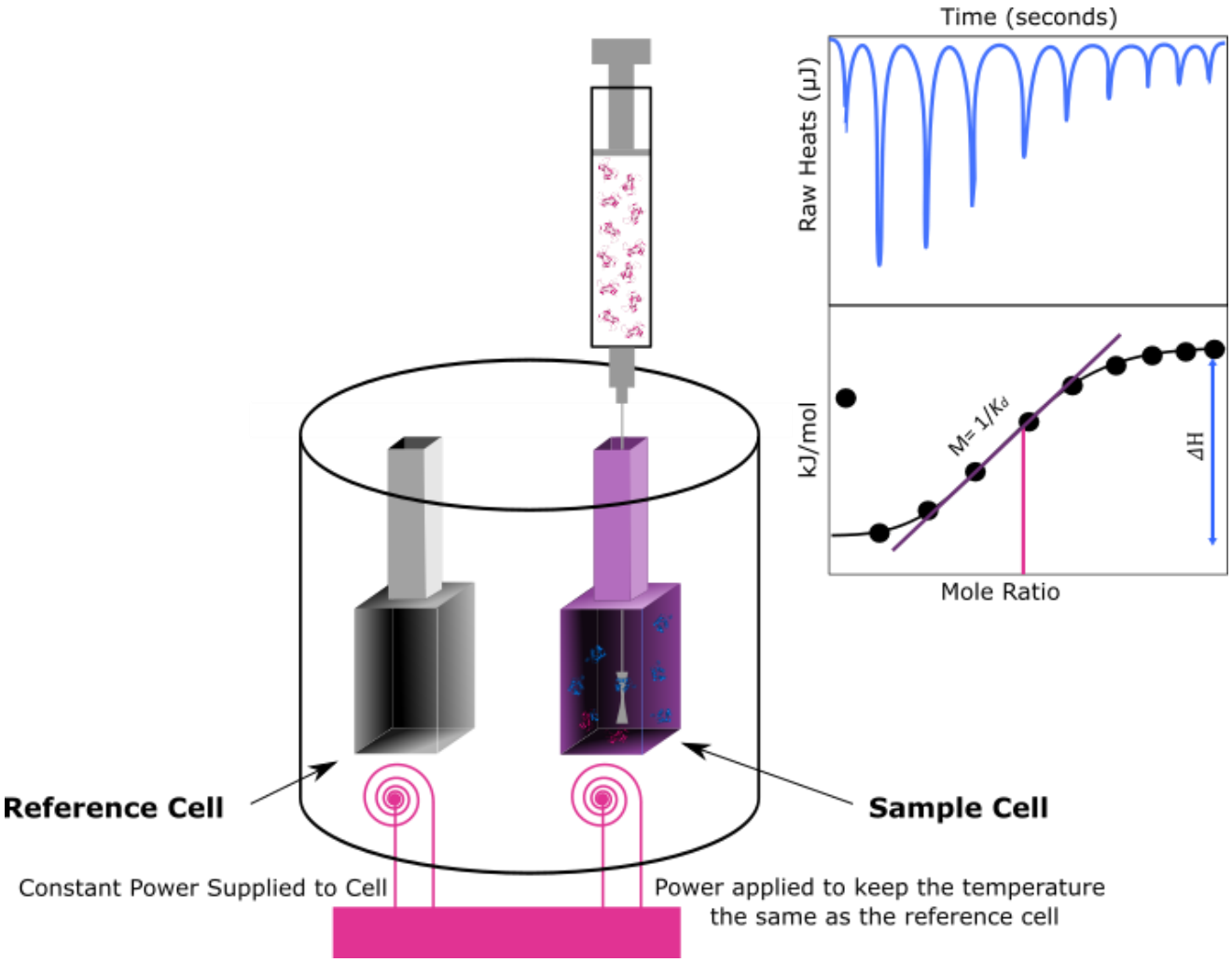


Figure 35: The basic concept of isothermal titration calorimetry (ITC).

An isothermal calorimeter typically has two cells, a reference cell (grey) and a sample cell (purple). Constant power is applied to the reference cell. As the titrant is titrated into the sample cell and the result is the release or absorbance of heat as the reaction occurs. Power is applied to the sample cell to keep the temperature the same as the reference cell. The output from this is used to create the raw heat thermogram. The area under the curve for each injection is used to calculate the raw heats (kJ/mol) for each titration which is plotted against the mole ratio. A model is then fitted to the curve in order to determine the thermodynamic parameters. Image created using Inkscape 0.92.4

Isothermal titration calorimetry was used to characterise the DNA binding activity of the TBR1 T–Box domain with its palindrome and single site DNA sequences. This was important in establishing a TBR1 T–Box DNA binding mechanism and to observe any

differences in binding between the Pal and SSL sequences, which would be indicated by a change in thermodynamic parameters.

All ITC data was generated with the Nano isothermal titration calorimeter (TA instruments, USA) and was done in triplicate. Purified WT TBR1 T–Box was dialysed freshly against 10 mM Sodium Phosphate, 3 mM TCEP and either 100 or 250 mM NaCl at pH 7.5. The protein was centrifuged at 4 °C for 10 minutes at 10 000 xg to remove any aggregates and final dialysis buffer was filtered through a 0.2 nM filter. Protein of a concentration of 60 µM, was loaded into the syringe. The DNA was solubilised at ~1000 µM in filtered and autoclaved MilliQ water. The DNA was diluted to 5 µM in the dialysis buffer and loaded into the sample cell. This was important so as to ensure there was no buffer mismatch. Once the DNA and protein were loaded, the calorimeter was left to equilibrate to 20 °C with stirring at a rate of 250 rpm. Between 45 and 50 five µl injections were done with an integration time of 300 seconds. The isotherms obtained were finally fitted using a one site independent model using TA instruments NanoAnalyze software. This model was chosen for both the single site DNA as well as the palindrome DNA because preliminary data suggested that binding to the palindrome sequence is not a cooperative process and that the two binding sites are identical. Furthermore, fitting the data to a two–site model resulted in unrealistic affinity and stoichiometry parameters.

3.2.6.3. Single Molecule–Förster Resonance Energy Transfer

Förster Resonance Energy Transfer (FRET) is a phenomenon whereby there is a radiationless transmission of energy from a donor molecule to an acceptor molecule that are in close proximity (usually between 10 to 100 Å). The donor molecule in this case is a fluorophore which absorbs light of a specific wavelength and transmits it to the acceptor molecule via a dipole–dipole interaction, which then emits the light at a longer wavelength. This transfer of energy results in a reduction in the donor's fluorescence intensity and lifetime along with an increase in the acceptor's fluorescence intensity. In order for two molecules to form a FRET pair, the absorption spectrum of the acceptor must overlap with the emission spectrum of the donor (**Figure 36**) (Lakowicz 2006; Roy *et al.* 2008; Schuler 2013).

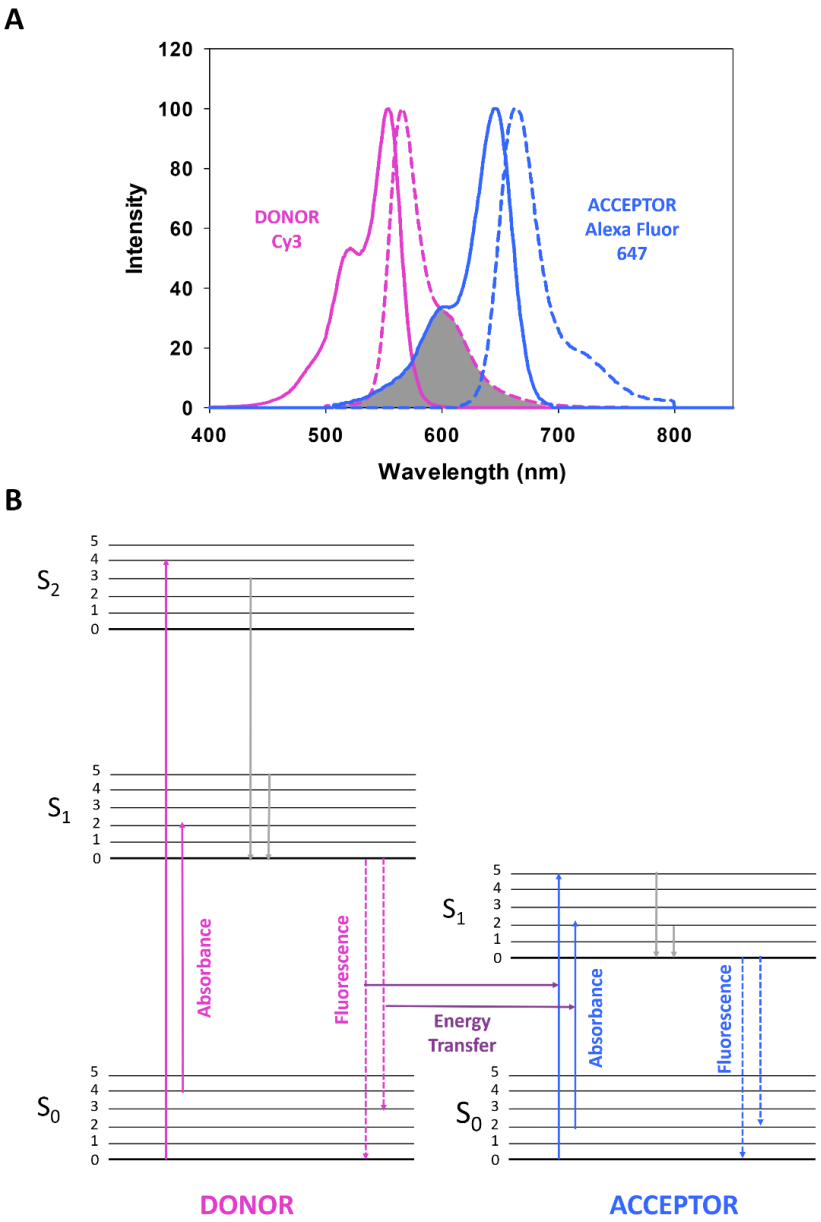


Figure 36: A: The absorption and emission spectra of Cy3 and Alexa Fluor 647. B: A Jablonski diagram showing the FRET energy transfer

A: The absorption (solid line) and the emission (dotted lines) of the FRET pair Cy3 (pink) and Alexa Fluor 647 (blue) are displayed above. The part where the donor's (Cy3) emission spectrum overlaps with the acceptor's (Alexa Fluor 647) absorbance spectrum is coloured in grey. **B**: S0, S1 and S2 are energy levels and the lines indicate different vibrational energy levels. Light of a specific wavelength is used to excite electrons and the light is absorbed by the donor (pink solid lines). Electrons fall back to the lowest vibrational energy level with energy loss due to heat and vibration (internal conversions). As the electron drops back down to the ground state there is a transfer of energy (purple) to the acceptor (blue). The electrons in the acceptor fluorophore are excited. As the electron drops back down to the ground state the energy is emitted as light (fluorescence)(Lakowicz 2006). Image created using Inkscape 0.92.4

The extent of the transfer is highly dependent on the distance between the fluorophore as well as the degree of spectral overlap. The Förster radius (R_0) is the distance between the two fluorophores where the efficiency of energy transfer is 50%, which is dependent of the overlap of the donor and acceptor emission and absorption spectra (**Equation 7**). The efficiency of the energy transfer of a compatible FRET pair is, therefore, indicative of the distance between them. Since the Förster radii for many FRET pairs are comparable in size to macromolecules, FRET can be used to accurately measure distances between sites on macromolecules (Lakowicz 2006; Roy *et al.* 2008; Schuler 2013; Stryer & Haughland 1967)

Equation 7: $E = R_0^6/(R_0^6 + r^6)$

Where E is the FRET efficiency, R_O is the Förster radius and r is the actual distance between the donor and the acceptor fluorophores.

Single molecule Förster resonance energy transfer (smFRET) makes use of the same FRET phenomenon as ensemble FRET experiments but monitors individual molecules. This is advantageous as there is no need to synchronise events and individual sub–populations in a mixture can be observed. SmFRET has been used for a wide variety of applications and has been used to measure distances, stoichiometries, conformational changes as well as frequency and duration of binding events (Roy *et al.* 2008). In a smFRET experiment the FRET efficiency is calculated using **Equation 8**.

Equation 8: $E = I_A/(I_D + \alpha I_A)$

Where I_A and I_D are the acceptor and donor fluorescence intensities and α is a correction for excitation of the acceptor and leakage of donor into the acceptor channel.

SmFRET requires labelling of the macromolecules with fluorophores that must have a high enough extinction coefficient and quantum yield, good photostability, be water soluble and be conjugatable to biological molecules (Roy *et al.* 2008). Cyanine dyes Cy3 and Cy5 have been used as a favourite FRET pair although newer dyes such as Alexa Fluor (Thermo Fisher Scientific) and Atto dyes (ATTO–TEC) are now available and have comparable properties to the Cy dyes (Roy *et al.* 2008).

SmFRET is commonly performed in two formats: immobilised and free diffusion where background noise is reduced by either limiting the illumination to a shallow depth (as with total internal reflection (TIR)) or to a very small volume (as in fluorescence correlation spectroscopy). Methods involving immobilisation have the advantage that interactions can be monitored for longer periods which is useful as many biological interactions can occur over timescales in the milliseconds to seconds range (Joo *et al.* 2008).

In TIR (**Figure 37**) a thin layer (~100 nm) of the sample on a quartz slide is illuminated by an evanescent wave. Snell's law describes how a wave is refracted based on the angle of incidence as it passes from one medium to another with different refractive indices. The evanescent wave is generated when the beam is directed at the interface of the quartz slide and the water at a critical angle which causes the beam to be completely reflected. The evanescent wave intensity decreases exponentially with distance, which means that only a small depth of the sample becomes illuminated, resulting in an increase in signal to noise ratio required to illuminate few enough molecules to resolve them individually (Axelrod *et al.* 1984; Holden *et al.* 2010).

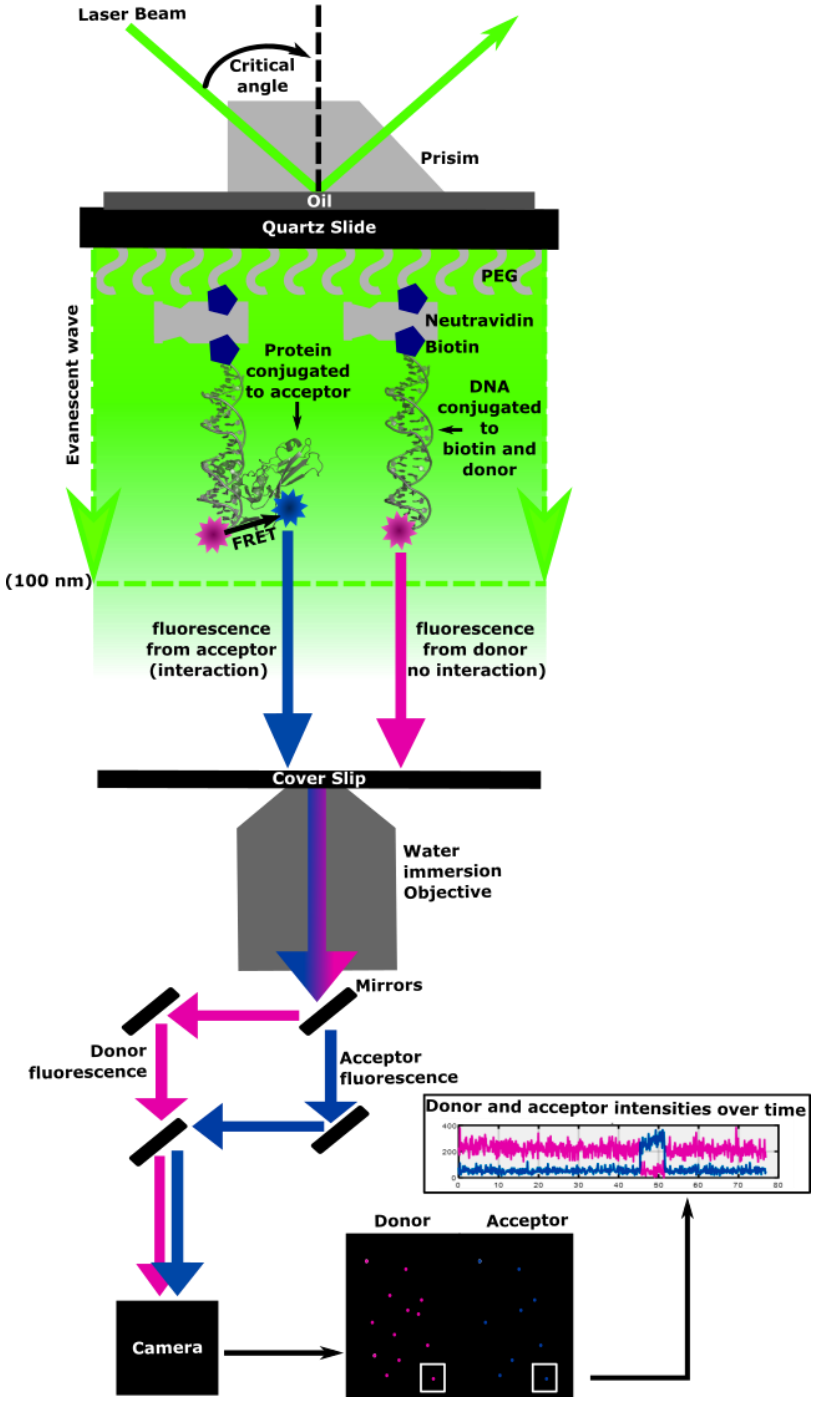


Figure 37: Schematic of the setup of a TIR smFRET microscope

A quartz slide is pegylated with a 1 in 100 dilution of biotinylated PEG in non–biotinylated PEG. The channels are washed with neutravidin which binds the biotin. A macromolecule, such as DNA, conjugated to biotin and a donor fluorophore is immobilised to the slide by binding the neutravidin via the biotin. A laser is directed at a prism which directs the beam onto the quartz slide at the critical angle which generates the evanescent wave. The evanescent wave illuminates a thin layer (~100 nm) of sample. The donor fluorophores in this thin layer are excited by the evanescent wave and either transfer energy to an acceptor fluorophore (if the two macromolecules are bound in close enough proximity) or else they fluoresce. The

fluorescent light is collected by an objective. A series of mirrors and filters separate the light emitted from the donors and acceptors and direct the two beams onto two halves of a camera. The result is a series of dots corresponding to individual immobilised molecules. The intensities are recorded and traces of the donor and acceptor intensities for each spot are displayed. The protein structure used in this figure is Xbra bound to DNA (1XBR) (Muller & Herrmann 1997). Image created using Inkscape 0.92.4

SmFRET studies were performed at the University of St Andrews in Scotland with the assistance of Dr Juan Carlos Penedo. SmFRET studies were conducted for TBR1 T–Box binding to Pal and SSL DNA to obtain kinetic binding parameters. Unfortunately, due to the Covid 19 pandemic, a full dataset was not obtained for the SSL DNA.

Quartz slides with holes drilled (**Figure 38**) and coverslips were cleaned extensively using the following procedure. The slides were first soaked for 30 minutes in 50% (v/v) hydrogen peroxide and 50% (v/v) concentrated HCl. They were then sonicated in alconox (15 minutes), MilliQ (5 minutes), acetone (15 minutes), MilliQ (5 minutes), 3M KOH (15 minutes), MilliQ (5 minutes), methanol (15 minutes), MilliQ (5 minutes) and 3M KOH (15 minutes). All the slides and coverslips were then dried with nitrogen. The slides were burned in a blue flame on each side for 2 minutes whereas the coverslips were quickly passed through the flame and checked for discoloration or deformation. An opaque copulin staining jar was filled with ~60 ml methanol, 2.5 ml of acetic acid and 200 µl aminosilane. The slides were sonicated in this solution for no more than 10 minutes. Once sonicated the solution was removed and the slides and coverslips were rinsed with methanol, MilliQ and methanol again. The slides and coverslips were dried with nitrogen and stored in a clean environment protected from light.

Biotinylated PEG (1–2 mg) and non–biotinylated PEG (100 mg) was dissolved in 320 µl of PEG buffer (100 mM sodium bicarbonate, pH 8.5). Sixty microliters of PEG solution was pipetted onto each slide. A coverslip was then placed over the drop of PEG solution on each slide, so it spread over the surfaces of the sides and coverslips. The slides and coverslips were balanced inside tip boxes and left protected from light for 1.5 h. The coverslips were removed from the slides keeping note of which sides were covered in PEG. The coverslips and slides were subjected to methanol, MilliQ and methanol rinsing steps and were subsequently dried with nitrogen once again. Double–sided tape was applied to the pegylated side of the slides to form channels. The coverslips (PEG side down) were placed on top of the double–sided tape and pressed with a pipette tip to ensure a seal. The slides were vacuum packed and stored at –80 °C until used. Before use of a slide, it was allowed to thaw protected from light for ~15 minutes and then the sides of the channels were sealed with epoxy (**Figure 38**). The channels were washed with HPLC grade water and slides were viewed under the microscope to check if they were clean.

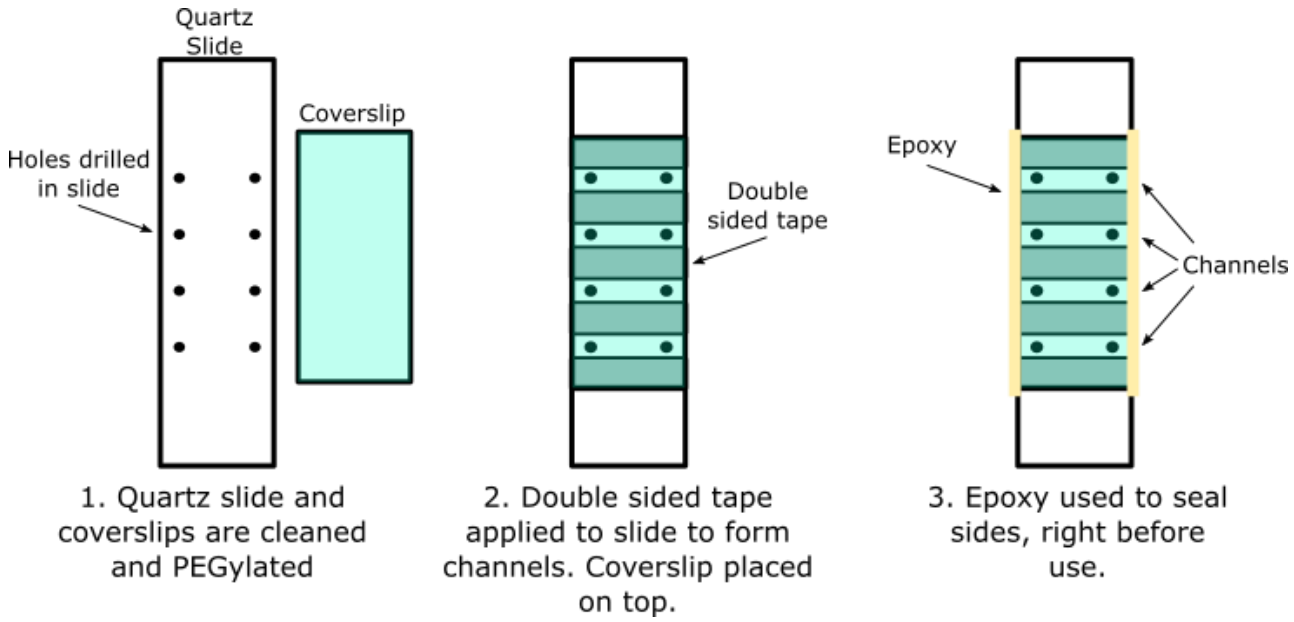


Figure 38: Preparation of slides for smFRET.

Quartz slides with holes drilled in them and coverslips were thoroughly cleaned and PEGylated. Doubled–sided tape was used to create channels on the slides and the coverslips were stuck on top. Just before use, the sides of the coverslip was sealed with epoxy resin (yellow). Image created using Inkscape 0.92.4.

Once a channel, on a slide, was deemed clean, 60 µl of 0.2 mg/ml neutravidin was applied to the channel and allowed to incubate for 15 minutes at room temperature. The channel was then washed with smFRET binding buffer (50 mM Tris–HCl, 50 mM NaCl, pH 7.5).

For the study of TBR1 T–Box binding to Pal DNA, first 20 pM Cy3 and biotin labelled Pal DNA (Cy3 Bio Pal) were applied to the channel and allowed to incubate for 5 minutes. The channel was then washed with imaging buffer (50 mM Tris–HCl, 50 mM NaCl, 6% (w/w) glucose, 0.1 mg/ml glucose oxidase, 0.02 mg/ml glucose catalase, 200 µM Trolox, pH 7.5) and data was collected. WT TBR1 T–Box labelled with Alexa Fluor 647 (T–Box Alexa 647) was diluted to 20 pM in imaging buffer and 60 µl was applied to the channel and data was collected. The channel was washed with increasing concentrations of T–Box Alexa 647 (up to 75 nM) diluted in imaging buffer and data was collected for each concentration.

For the study of TBR1 T–Box binding to SSL DNA, first 20 pM Cy3 and biotin labelled Pal DNA (Cy3 Bio Pal) were applied to the channel and allowed to incubate for 5 minutes. The channel was then washed with imaging buffer and data was collected. WT TBR1 T–Box labelled with Alexa Fluor 647 (T–Box Alexa 647) was diluted to 20 pM in imaging buffer and 60 µl was applied to the channel and data was collected. The channel was washed with increasing concentrations of T–Box Alexa 647 (up to 50 nM) diluted in imaging buffer and data was collected for each concentration.

Data was collected on a custom–built prism type TIR single molecule setup. It consisted of an Olympus IX71 microscope with a 1.2NA water–immersion objective (UPLAPO60XW, Olympus, Melville, USA). Fifty mW continuous–wave lasers emitting at 532 nm and 642 nm were used as excitation light sources (Crystalaser, Reno, USA). Long–pass dichroic mirrors (DCRLP, 645 nm) (Chroma Technology Corp, Bellows Falls, USA) were used to split the donor and acceptor signals and direct then to a back illuminated iXon electron–multiplying CCD camera (Andor, Belfast, UK). Carboxylated fluorescent polystyrene beads (Crimson Fluorescent Fluospheres, 5 µm) were used to correlate the donor and acceptor signals. The raw data movies were obtained and processed using custom–built software in IDL 6.0 (Exelis, Boulder, USA). The data was then further analysed using laboratory written software in MatLab (Mathworks, Natick, USA).

3.2.6.4. FLUORESCENCE ANISOTROPY

Fluorescence anisotropy (FA) is a technique whereby an interaction between two macromolecules is observed by monitoring a change in tumbling rate as the two molecules interact. In order to observe the change in tumbling rate, one molecule is labelled with a fluorescent molecule which is excited by vertically plane polarised light. Only the fluorophores whose dipoles are aligned with the polarisation of the light will be excited. The labelled molecule is smaller than the complex and will thus have a fast tumbling speed. When the interacting partner binds the labelled molecule, the complex is larger and thus will have a much slower tumbling speed. The greater the tumbling of the complex the more the emission will be depolarised **Figure 39**. Only the fluorophores whose dipoles are aligned with the polarisers' emissions are detected. A ratio of the vertically and horizontally polarised light is used to determine the anisotropy value which is related to the proportion of complex formed (LiCata & Wowor 2008).

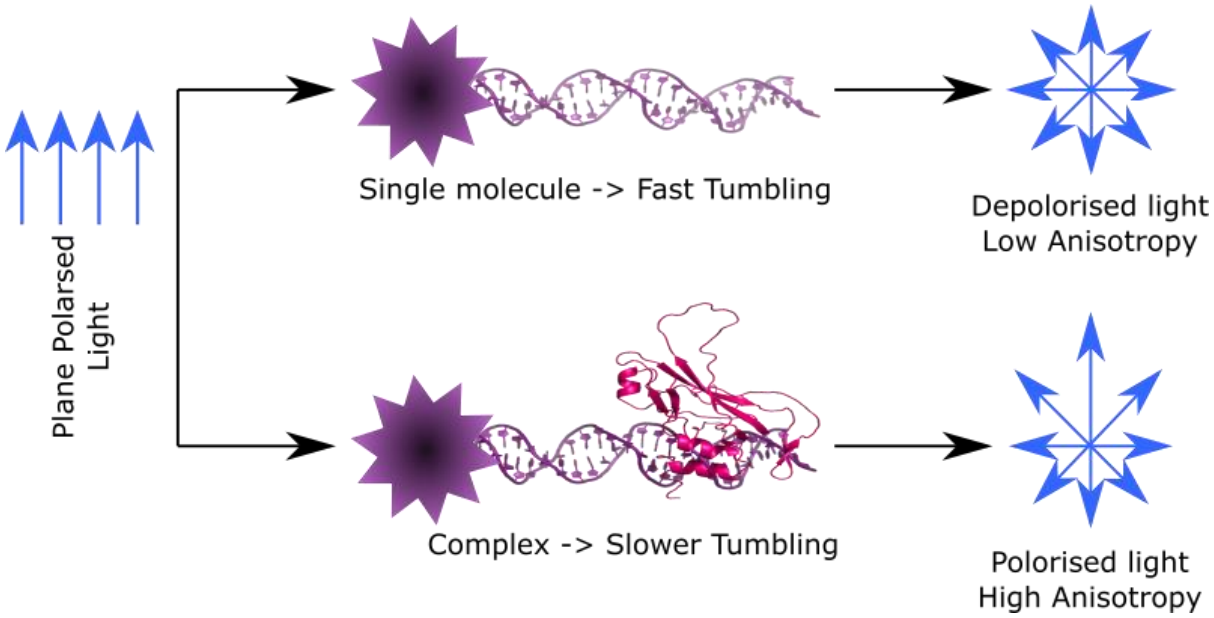


Figure 39: A diagram depicting the basic principle of fluorescence anisotropy.

Vertically plane polarised light is used to excite molecule A (purple) labelled with a fluorophore. This small complex has a fast tumbling volume and thus depolarises the light. When molecule B (pink) binds molecule A the larger complex has a slower tumbling volume and depolarises the light less. The anisotropy value is calculated and is proportional to the average tumbling volume which is indicative of the fraction of molecule A interacting with molecule B. Pink protein and DNA is the T domain from *Xenopus Laevis* bound to DNA (1XBR) (Muller & Herrmann 1997). Image created using Inkscape 0.92.4

Fluorescence Anisotropy was used for 3 purposes, firstly, to determine the binding affinity of protein–DNA interactions. Secondly, to determine the binding affinity of protein–protein interactions and finally, to observe if a ternary complex forms when both proteins and DNA are present.

The anisotropy measurements were performed on a Perkin Elmer LS50B fluorimeter and the experiments followed the same basic design. A sample of the labelled macromolecule was prepared and was used to calculate the grating factor (G Factor). The G Factor accounts for machine bias and is automatically determined by the instrument using **Equation 9**. The labelled sample was then titrated with a high concentration of its binding partner. The exact concentrations were optimised for each pair depending on the K_d. The anisotropy value of the first sample (labelled macromolecule with no binding partner) was subtracted from each value in the replicate to get the relative anisotropy value, which was plotted to get the curve. The data for the protein–DNA and protein–protein interactions were then fitted using SigmaPlot version 12 to a one site saturation, ligand binding curve (**Equation 10**).

Equation 9: $G = I_{hv}/I_{hh}$

Where, G is the G factor, I_{hh} is the intensity of light detected when both the emissions and excitations polarisers are horizontal and I_{hv} is the intensity of light detected when the emissions polariser is horizontal, and the emissions polariser is vertical.

Equation 10: $Y = (B_{max})(x)/(K_D + x)$

The one site saturation, ligand binding model used to fit the fluorescence anisotropy data.

Fluorescence anisotropy was used to determine the binding affinity of the TBR1 T–Box and FOXP2 FHD to SSL and Nel C DNA. In order to achieve this, the duplex DNA sequence was synthesised and conjugated to Rhodamine X (ROX) (**Figure 40**) by Integrated DNA technologies–WhiteSci scientific, RSA. The excitation and emission wavelengths were 580 and 605 nm respectively. Both the excitation and emission band widths were set to 5 nm and the integration time was 3 seconds. A 0.2 µM ROX labelled DNA sample was prepared in 10 mM sodium phosphate, 100 mM NaCl, 3 mM TCEP, pH 7.5 buffer. A sample of 0.2 µM ROX labelled DNA was titrated with a high concentration of protein (>100 µM).

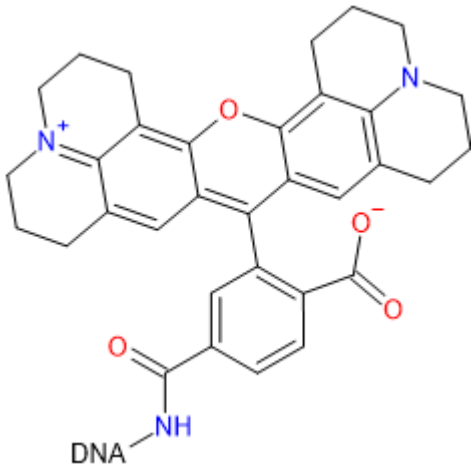


Figure 40: The structure of Rhodamine X (ROX).

The NHS ester of Rhodamine X was conjugated to the duplex DNA sequences by Integrated DNA technologies– WhiteSci scientific, RSA. Image generated using ACD ChemSketch 2019.2.0.

Fluorescence anisotropy was used to determine the binding affinity of the TBR1 T–Box to the FOXP2 FHD. In order to achieve this, uncleaved TBR1 T–Box was labelled with NTA–Atto 550 as previously described. Two 0.5 µM NTA–Atto labelled T–Box samples were prepared in 10 mM sodium phosphate, 3 mM TCEP and 100 or 500 mM NaCl, pH 7.5 buffer. The excitation and emission wavelengths were 558 and 608 nm respectively. Both the excitation and emission band widths were set to 5 nm and the integration time

was 3 seconds. The samples were titrated with a high concentration (>100 µM) FHD starting from a concentration of 0.25 µM up to 20 µM.

Fluorescence anisotropy was used to attempt to monitor binding events that occur in the presence of DNA and both FOXP2 FHD and TBR1 T–Box. For the SSL DNA, containing a single T–Box binding site, 0.2 µM of ROX labelled SSL was used to measure the G Factor and take the first anisotropy reading. Then 500 nM TBR1 T–Box was added to the DNA, this concentration was chosen as it is the concentration where the T–Box–SSL binding curve started to reach saturation. A high concentration of FHD was titrated into the sample and readings were obtained. An additional buffer control was also performed in order to ensure any changes in anisotropy values were not due to the dilution of the DNA and T–Box. The anisotropy value of DNA alone was subtracted from each other value in the curve in order to plot relative anisotropy.

For the Nel C DNA, containing a single FHD binding site, 0.2 µM of ROX labelled Nel C was used to take the first reading. Then 8 µM FOXP2 FHD was added to the DNA, this concentration was chosen as it is the concentration where the FHD Nel C binding curve started to reach saturation. A high concentration of T–Box was titrated into the sample and readings were obtained. An additional buffer control was also performed in order to ensure any changes in anisotropy values were not due to the dilution of the DNA and FHD. The anisotropy value of DNA alone was subtracted from each other value in the curve in order to plot relative anisotropy.

3.2.7 *In Silico* Studies

3.2.7.1 T–Box Structure Predictions and Modelling

Unfortunately, while there are a few T–Box structures in the protein data bank, there are no structures of the TBR1 T–Box specifically. As structure pertains to function, a model of the TBR1 T–Box structure was constructed in order to better understand how it interacts with DNA and with FOXP2, as well as to interpret the structural information obtained *in vitro*. The I–TASSER server (https://zhanglab.ccmb.med.umich.edu/I–TASSER/) was, therefore, used to generate a model of the TBR1 T–Box.

The I–TASSER algorithm first takes the user submitted amino acid sequence and finds template proteins with similar folds. I–TASSER then excises fragments from the various templates and reassembles them into full–length structures. Any regions that do not align to any known structures are built using *ab initio* modelling. The lowest energy states, or near native models are determined from the pool of protein structure decoys by clustering them into centroids. Fragments are then reassembled from each cluster to remove steric clashes and improve the global topology for the models. These models are re–clustered and the lowest energy models are further refined. Finally, the models are clustered again and scored (Roy *et al.* 2010; Yang *et al.* 2014; Yang & Zhang 2015). Once the models were generated, they were energy minimised firstly using the

ModRefiner algorithm, which runs refinement simulations where both the backbones and sidechains are flexible (Xu & Zhang 2011). A second energy minimisation step was done using Chimera 1.7 (Pettersen *et al.* 2004).

Nel and SSL DNA were modelled using the build nucleic acid function in Accelrys Discovery studio 4.1, the Pal DNA model was obtained from chains C and D of the TBX21 crystals structure (5T1J). The WT TBR1 T–Box model obtained from I–TASSER was docked to the Nel and Pal DNA models using the easy interface HADDOCK (High ambiguity driven docking approach) server (Dominguez *et al.* 2003). HADDOCK makes use of ambiguous interaction restraints (AIRS) which contain biophysical information to aid in the docking process. The docked structures are subsequently ranked based on their intermolecular energy which considers electrostatic and Van der Waals energy along with the AIRs. The AIRs in the easy interface, which was used, were automatically generated based on user defined active residues (residues involved directly in the interaction) and passive residues (the surface residues around the binding site). The active residues used for each docking simulations are displayed in **Table 13** and were chosen based on other T–Box crystal structures. The docked models are grouped into clusters based on their similarity by HADDOCK. The best scored structure of the cluster with the highest HADDOCK score was analysed for each docking simulation. Docking simulations were run, where the TBR1 T–Box model (monomeric) was simply docked to Pal DNA, SSL DNA and Nel DNA.

Table 13: The active residues defined for each docking simulation

Docking simulation	Protein Active residues	DNA Active residues
T–Box monomer to Pal DNA	Ile 226, Arg 232, Thr 370, Ala 371, Phe 387, Gly 390, Phe 391	ATTTCACACCT**AGGTGTGAA**AT
T–Box to SSL DNA	Ile 226, Arg 232, Thr 370, Ala 371, Phe 387, Gly 390, Phe 391	CCCAAT**TTCACACCT**TCCTCA
T–Box to Nel	Ile 226, Arg 232, Thr 370, Ala 371, Phe 387, Gly 390, Phe 391	AGG**TG****TTTAC**TTTCATAG

RESULTS

The interactions that transcription factors form is what determines how genes are transcribed and thus expressed. This in turn regulates important processes such as cell differentiation and homeostasis. The aim of this project was to obtain a comprehensive picture of the interactions that FOXP2 and TBR1 form with each other and with DNA, to aid in the understanding of how these proteins are involved in conditions such as ASD. This was achieved firstly by overexpression and purification of functionally folded FOXP2 FHD and TBR1 T–Box domain. Structural characterisations (intrinsic tryptophan fluorescence and SEC) were performed firstly to ensure the overexpressed proteins were folded. Additionally, comparisons were made with the WT T–Box and the mutants to ensure there were no significant changes in tertiary structure upon mutation. SEC was used to ensure that each protein existed in a single predominant oligomeric state. The T–Box mutants (each of which had one of the two cysteines in the T–Box domain replaced with either Ala or Ser) were used to determine if disulfide bonds were involved in TBR1 T–Box dimerisation. Finally, DNA binding studies (EMSA, FA, ITC and smFRET) and protein–protein interaction studies (EMSA and FA) were done to understand any functional changes upon the interaction of these two proteins as well to obtain the TBR1 T–Box DNA binding mechanism. By confirming and characterising each interaction, not only would functional information about the TRB1–FOXP2 interaction be found, but some of the interacting network involved in neural development would also be elucidated.

4.1. MUTAGENESIS AND SEQUENCING

pET–11a plasmids with inserts encoding the gene sequences of WT FOXP2 FHD and WT TBR1 T–Box domain were obtained from GenScript (USA). These vectors were used to transform T7 *E. coli* cells. The plasmids were isolated using the GeneJet mini prep kit (Thermo Fisher scientific, EU) and were sequenced at Inqaba Biotech (RSA) to confirm the sequences (**Figure 42** and **Figure 43**).

Furthermore, the TBR1 T–Box C209A and C274S mutants needed to be produced so primers for mutagenesis (**Table 6**) were designed using PrimerX (www.bioinformatics.org). The pET–11a plasmid incorporated with the WT TBR1 T–Box gene was used as a template for mutagenesis. The PCR products obtained from the mutagenesis reactions were resolved on a 1% agarose gel (**Figure 41**). A large band was present in the first lane which is likely the intact C209A mutant plasmid. Lane two contains many bands which shows that the PCR product of the C274S mutant had a large amount of degraded plasmid. The product was, however, successfully used to transform XL–10 Gold cells which means there was enough intact plasmid present in the sample. The cells were grown up and the plasmids were extracted using a mini prep kit. The extracted plasmids were sequenced, and the correct mutations were found to have occurred (**Figure 42**).

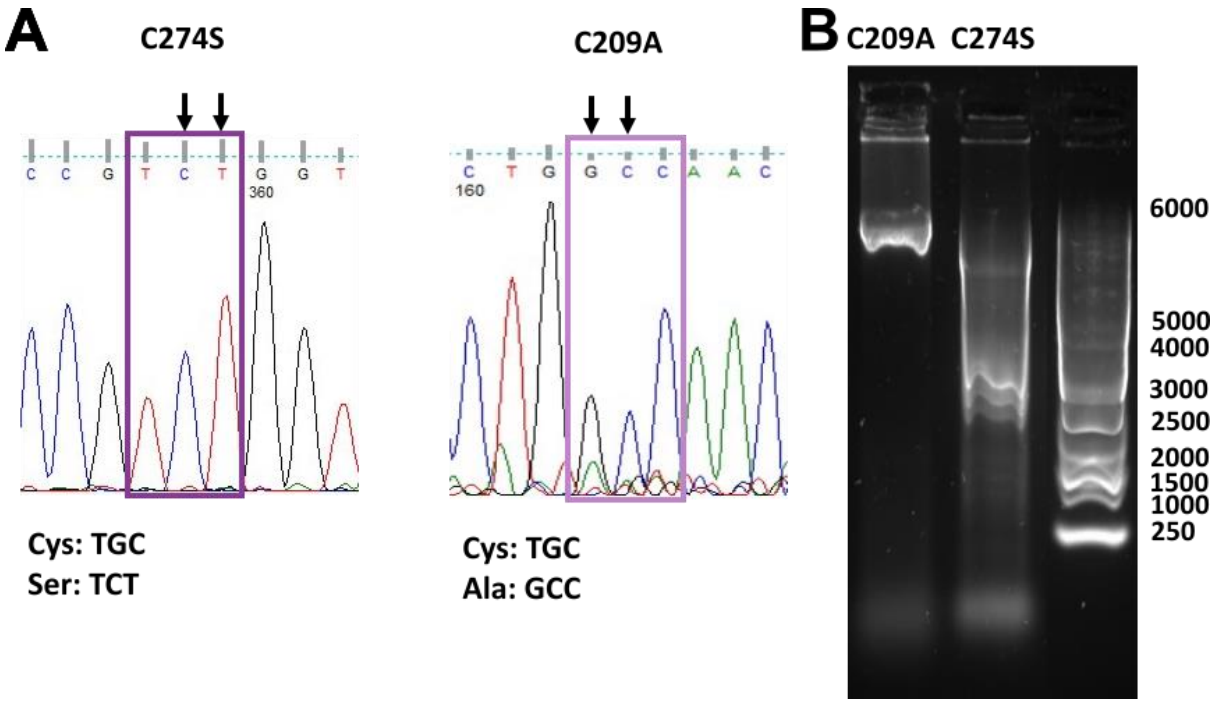


Figure 41: Sequencing results (A) and the PCR product resolved on an agarose gel (B) of the T–Box C209A and C274S mutations.

A: The purple blocks indicate the mutated codons. The arrows above show which nucleotides were changed**. B**: PCR product obtained from PCR reaction resolved on a 1% agarose gel stained with SYBR gold nucleic acid stain. The intact plasmids are visible as clear bands at the top of the gel. Lane 1 and 2 contains the PCR product of the C274S and C209A PCR reactions. The C274S PCR product contains a lot of degraded plasmid; however, enough intact plasmid was present for the transformation. Lane 3 contains the O'GeneRuler 1 kb DNA Ladder (Thermo fisher scientific, EU).

Figure 42: Multiple sequence alignment of the sequencing results for WT TBR1 T–Box, C209A TBR1 T–Box and C274S TBR1 T–Box.

The vectors containing WT TBR1 T–Box, C209A TBR1 T–Box domain and C274S TBR1 T–Box domain inserts were grown up and the plasmids were extracted and sent for sequencing. The sequencing results are displayed (**WT T–Box**: pink, **C209A T–Box**: lilac and **C274S T–Box**: purple). The translated sequencing results were aligned to the translated codon optimised sequence (Gasteiger *et al.* 2005) using ClustalW (Larkin *et al.* 2007) and visualised using Jalview version 2 (Waterhouse *et al.* 2009). The mutations are indicated in black boxes. Image edited using Inkscape 0.92.4.

Figure 43: Sequence alignment of the sequencing results of the WT FOXP2 FHD.

The vector containing WT FOXP2 FHD insert was grown up and the plasmid was extracted and sent for sequencing. The translated sequencing result is aligned to the translated codon optimised sequence (Gasteiger *et al.* 2005) using ClustalW (Larkin *et al.* 2007) and visualised using Jalview version 2 (Waterhouse *et al.* 2009). Image edited using Inkscape 0.92.4.

4.2. Purification

In order to obtain sufficient quantities of each protein to perform structural characterisations and binding studies, the proteins were expressed using the pET–11a expression vector in T7 *E. coli* cells. The FHD was subsequently purified using IMAC and thereafter the His–tag was cleaved. The T–Box variants were purified using IMAC followed by size exclusion chromatography. The following sections show an example of how each protein was purified.

4.2.1. FHD Purification

The WT FHD was expressed for 22 hours at 20 °C after induction with 0.5 mM IPTG. The soluble fraction of the cell lysates was loaded on to the IMAC columns and fractions were collected of the eluted proteins (**Figure 44**). Fractions 2 and 3 contained large quantities of FOXP2 FHD and were thus pooled and the His–tag was cleaved using thrombin. There were still large quantities of contaminants in the pooled fractions, however, these containments mostly aggregated during dialysis against thrombin cleavage buffer and were removed by centrifugation at 10 000 xg for 10 minutes.

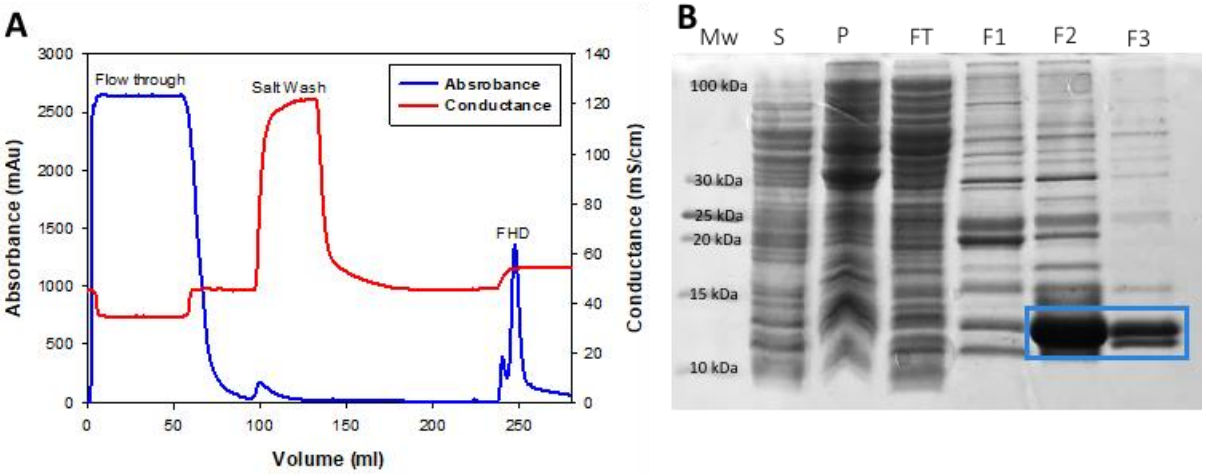


Figure 44: Immobilised metal affinity chromatography (IMAC) purification of the WT FOXP2 FHD.

A: The IMAC elution profile of the WT FHD. The blue line is the absorbance (mAu), and the red line is the conductance (mS/cm). The protein was eluted using imidazole as indicated by the absorbance peak at 250 ml. **B**: 16% tricine SDS–PAGE gel of samples collected from the purification. S: supernatant. P: pellet. FT: flow through from the IMAC column. F1–3: fractions collected of the eluted WT FHD peak. Fractions F2 and F3 were pooled for thrombin cleavage. The blue block indicates bands that correspond to the expected size of the FOXP2 FHD. The protein eluted from IMAC (F2 and F3) contained large quantities of FHD, however, many protein contaminants were still present. The Thermo scientific low range unstained protein ladder (100–3.4 kDa) was used in lane 1. Data was plotted using SigmaPlot version 12.

Thrombin was added to the dialysed protein and a sample was taken every half an hour and resolved on an 16% tricine SDS–PAGE gel. In order to remove the cleaved–off His–tag and thrombin the protein solution was passed through a benzamidine column (to bind the thrombin) and an IMAC column (to bind the His–tag). In the gel in **Figure 45** after 3 hours there is still some uncleaved protein. Samples of the fractions of the cleaved protein eluted from the IMAC and benzamidine columns are indicated on the gel. The protein in these two lanes was free from protein contaminants and could be used for further study.

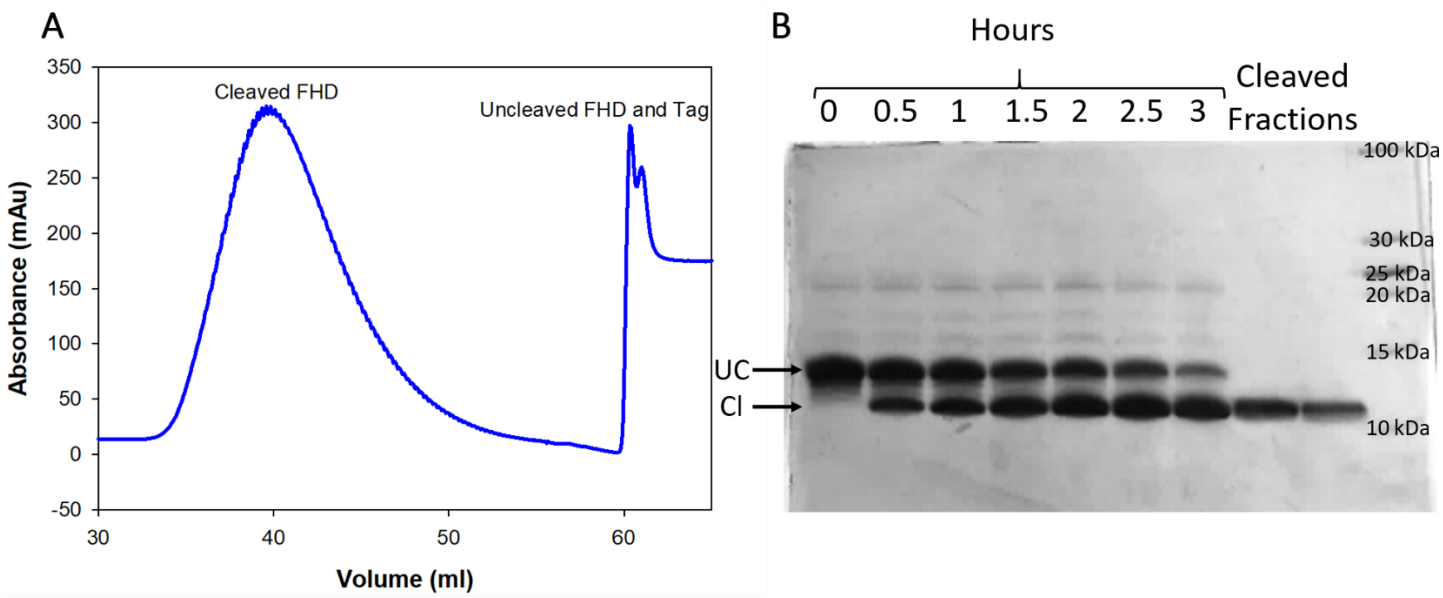


Figure 45: WT FOXP2 FHD His–tag cleavage.

A: The elution profile of cleaved protein from the IMAC and benzamidine column. The first peak which eluted contains the cleaved FHD which passed through the IMAC and benzamidine columns. The second peak contains uncleaved FHD which bound the IMAC column and was eluted using IMAC elution buffer. **B**: 16% tricine SDS–PAGE gel showing samples taken at time intervals during cleavage and samples of the cleaved protein eluted from the IMAC and benzamidine columns (F1 and F2). The uncleaved (UC) and cleaved (Cl) proteins are indicated on the gel. The protein in the cleaved fractions was pure and used for downstream studies. The Thermo scientific low range unstained protein ladder (100–3.4 kDa) was used. Data was plotted using SigmaPlot version 12.

4.2.2. T–Box Purifications

The WT TBR1 T–Box, C209A TBR1 T–Box and C274S TBR1 T–Box were expressed for 22 hours at 20 °C after induction with 0.2 mM IPTG. The soluble fraction of the cell lysates was loaded on to IMAC columns and fractions were collected of the eluted protein (**Figure 46**). Unlike the FHD purification, with the T–Box purifications, a peak is clearly visible from the 1.5 M NaCl and 0.1% Tween 20 wash (W1 and W2 in **Figure 46**) indicating the removal of non–specifically bound contaminants and/or DNA from the purified sample. Fractions 1 to 4 in **Figure 46** contained significant quantities of each T–

Box proteins, however, large amounts of contamination was still present, so size exclusion chromatography was used as a second purification step for each protein (**Figure 47**).

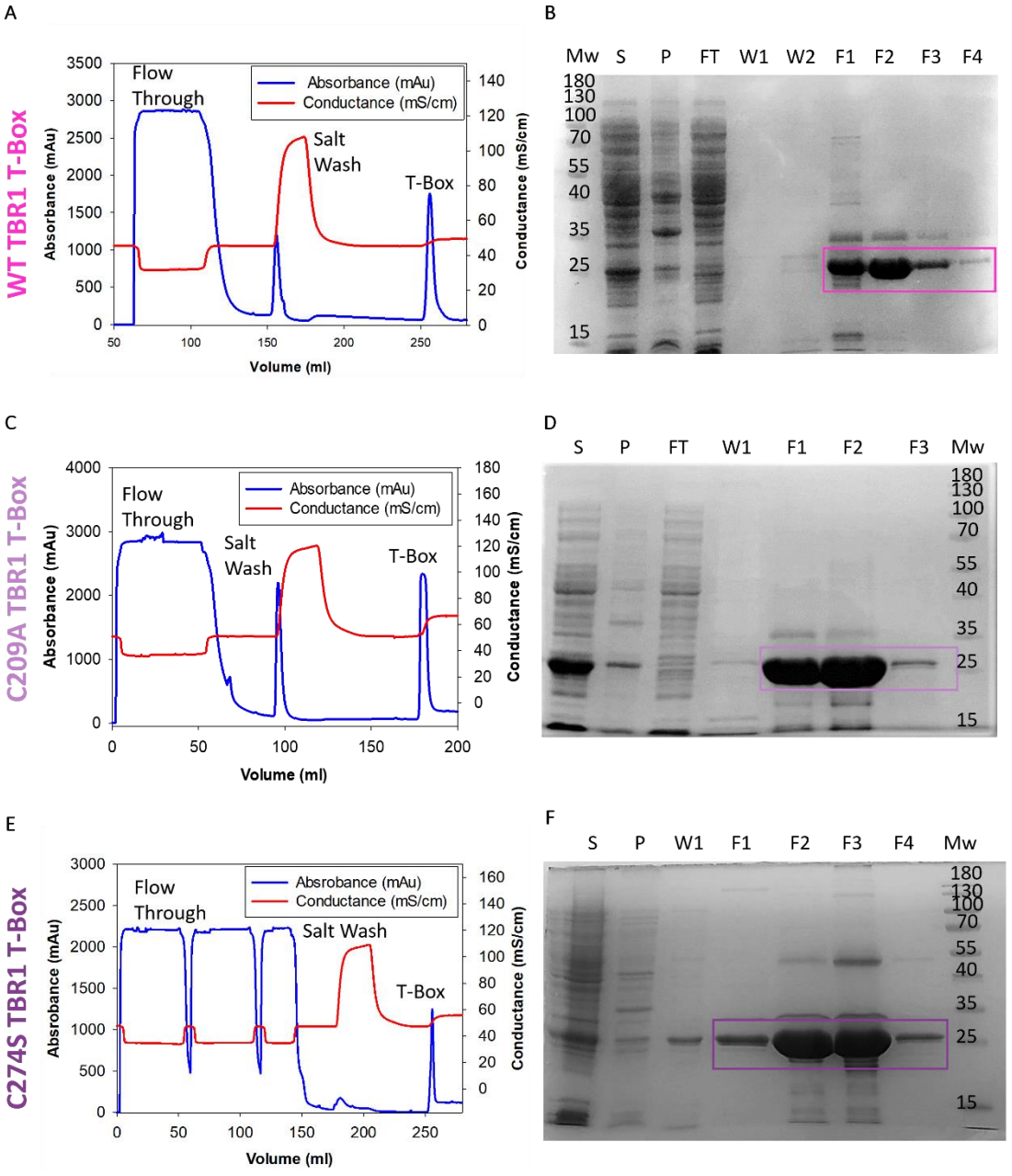


Figure 46: Immobilised metal affinity chromatography (IMAC) purification of the WT TBR1 T–Box, C209A TBR1 T–Box and C274S TBR1 T–Box domains.

The IMAC elution profiles for WT T–Box (**A**), C209A T–Box (**C**) and C274S T–Box (**E**) are displayed. The blue lines are the absorbance (mAu), and the red lines are the conductance (mS/cm). After the flow through was collected, wash steps consisting of 1.5 M NaCl and 0.1% Tween 20 were done. The proteins were eluted using imidazole as indicated by the labelled peaks. Reducing glycine SDS–PAGE (12.5%) samples were collected from the purifications of WT T–Box (**B**), C209A T–Box (**D**) and C274S T–Box (**F**). S: supernatant. P: pellet. FT: flow through from the IMAC columns.

W1–2: samples from the salt and Tween washes. F1–4: fractions collected of the eluted protein peaks. The fractions with the coloured blocks for each protein were pooled and further purified using size exclusion chromatography. The Thermo scientific PageRuler^{TM} Prestained Protein Ladder, 10 to 180 kDa was used. Data was plotted using SigmaPlot version 12.

The pooled fractions, obtained from each IMAC purification of the various T–Box proteins, were concentrated down to a volume of less than 6 ml and 1.5 ml was loaded on to the SEC column for each run. Each sample was centrifuged at 10 000 xg for 10 minutes and reduced with 2 mM DTT 10 minutes prior to loading onto the size exclusion column (pre–equilibrated with IMAC equilibration buffer). (**Figure 47**). All the T–Box domains eluted predominantly as a monomer with a small dimer peak. The small amount of dimer present is likely either a result of incomplete reduction of disulfide linked dimers or the formation of non-disulfide linked dimers. A very small amount of T–Box eluted along with contaminants between the void volume and the dimer peak. Protein in the fractions circled in the boxes in **Figure 47** were deemed sufficiently pure of other protein contaminants, pooled and used for further applications.

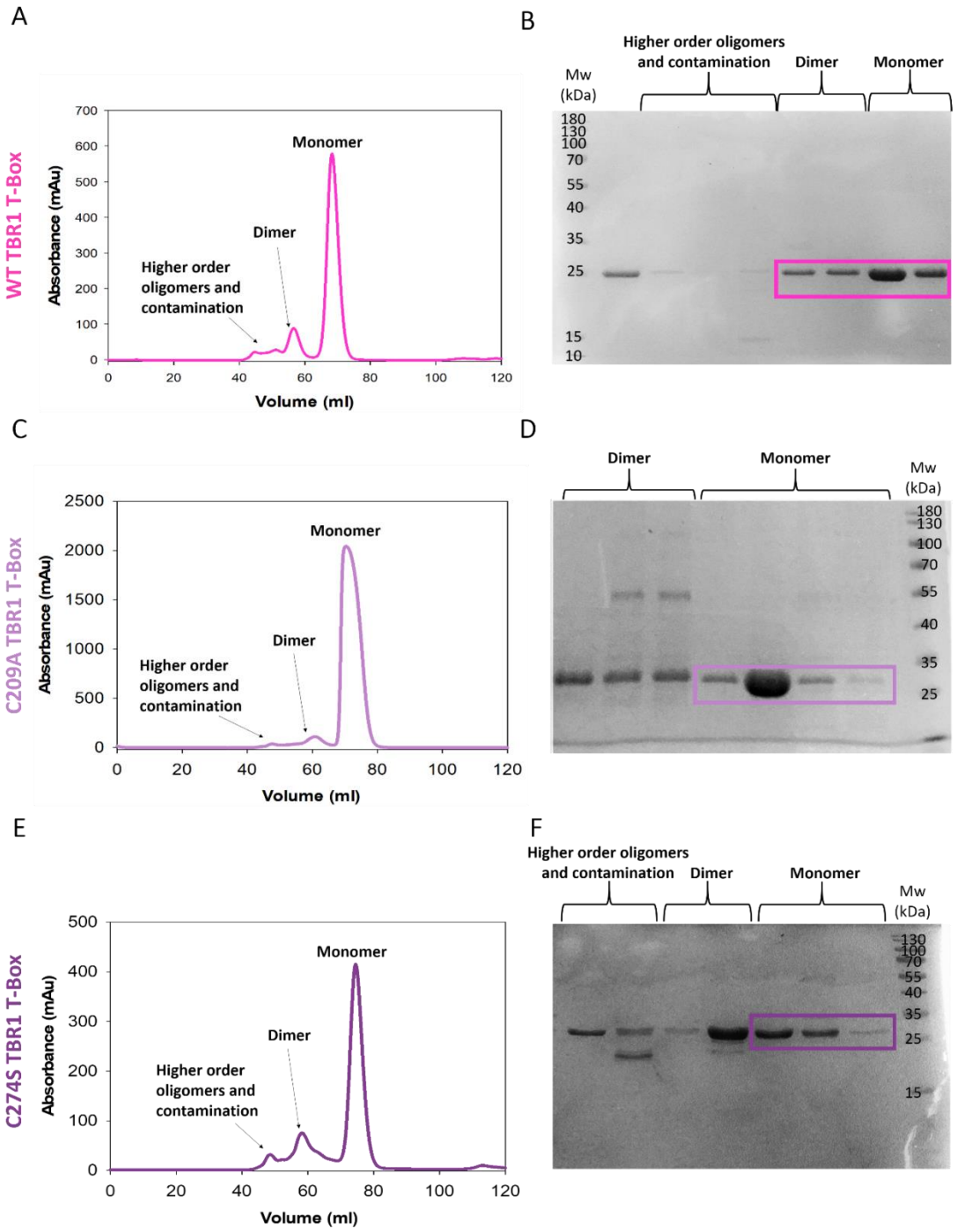


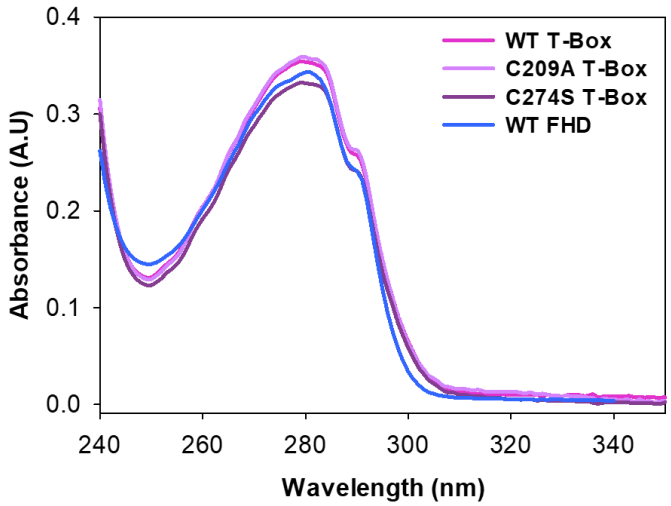
Figure 47: Size exclusion chromatography of the WT TBR1 T–Box, C209A TBR1 T–Box and C274S TBR1 T–Box domains.

The SEC elution profiles for WT T–Box (**A**), C209A T–Box (**C**) and C274S T–Box (**E**) are displayed. The Glycine SDS–PAGE (12.5%) with samples collected from the purifications of WT T–Box (**B**), C209A T–Box (**D**) and C274S T–Box (**F**). Samples were collected from each peak. The oligomeric state of each TBR1 T–Box is indicated on the elution profiles. The boxes on each gel indicate the fractions of protein that were pooled and used for downstream studies. Thermo scientific PageRuler^{TM} Prestained Protein Ladder, 10 to 180 kDa was used. Data was plotted using SigmaPlot version 12.

4.3. PROTEIN QUANTIFICATION

The pooled fractions of each protein (WT FHD, WT T–Box, C209A T–Box and C274S T–Box) were dialysed against binding buffer (10 mM sodium phosphate 100 mM NaCl, pH 7.5) and aggregates were removed by centrifugation. The concentration of each protein was determined by taking an absorbance spectrum in triplicate (**Figure 48**), the Beer–Lambert law (**Equation 1**) and each protein's extinction coefficient (**Table 9**). Each protein showed a typical absorbance spectrum for a protein where peaks at 280 nm and 295 nm indicate the absorbance of tyrosine and tryptophan residues. No peak at 340 nm suggests that there were no protein aggregates. The A_{280}/A_{260} ratio for all the proteins was ~ 1.7 which indicates that they were free from nucleic acid contamination. The concentration and A_{280}/A_{260} ratio obtained from these spectra is displayed in **Figure 48**.

A



B

Protein	A_{280}	A_{260}	Concentration (µM)	A_{280}/A_{260}
WT T-Box	0.35356	0.2044	9.7	1.73
C209A T-Box	0.3587	0.2038	9.8	1.76
C274S T-Box	0.3319	0.1919	9.1	1.73
WT FHD	0.3430	0.2014	15.3	1.7

Figure 48: Absorbance spectra of WT FHD, WT T–Box, C209A T–Box and C274S T–Box.

A: An absorbance spectrum of each protein WT T–Box (pink), C209A T–Box (lilac), C274S T–Box (purple) and WT FHD (blue). The spectra were taken on a JASCO V–630 spectrophotometer from 240 to 350 nm. **B:** The concentration and A_{280}/A_{260} ratios calculated from the protein spectra. The protein concentrations were calculated using the Beer– Lambert law (**Equation 1**), the theoretical extinction coefficients (**Equation 2**) for each protein were obtained from (**Table 9**). The A_{280}/A_{260} ratios suggest that all the proteins are free from DNA contamination.

The sizes of each protein were determined by constructing a standard curve from each protein's final purification gel. All the proteins ran to approximately the correct size on SDS–PAGE (**Figure 49**). The WT FHD, WT T–Box, C209A T–Box and C274S T–Box were sized as 10.2 kDa, 26.5 kDa, 25.9 kDa and 26.2 kDa, respectively.

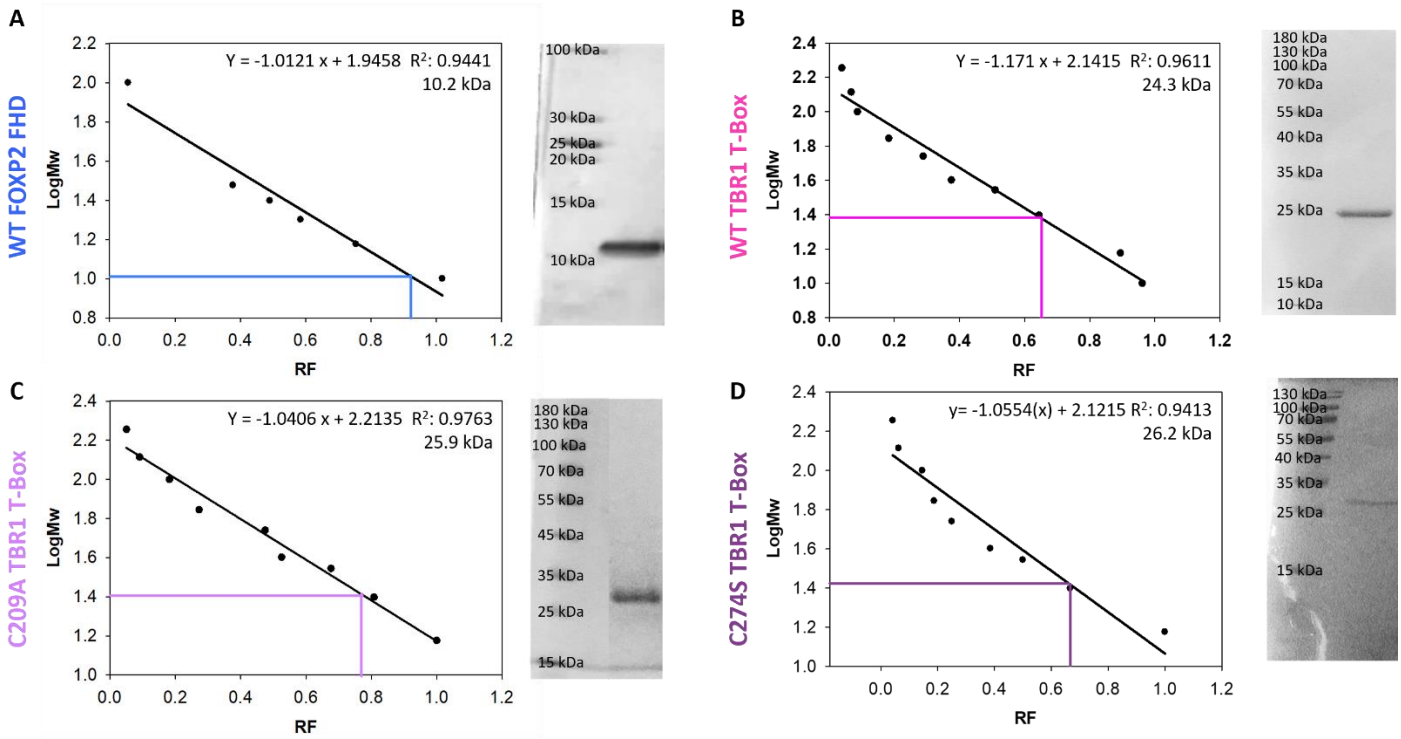


Figure 49: The standard curves used to calculate the sizes of WT FHD, WT T–Box, C209A T–Box and C274S T–Box using SDS–PAGE.

The sizes of protein WT T–Box (pink), C209A T–Box (lilac), C274S T–Box (purple) and WT FHD (blue). The RF for each band in the molecular weight marker was plotted against the log of its molecular weight (kDa). A linear regression was fitted, and the equation was used to determine the size of the protein on each gel. The WT FHD (A) was 10.2 kDa, WT T–Box (B) was 24.3 kDa, C209A T–Box (C) was 25.9 kDa and C274S T–Box (D) was 26.2 kDa which were all very close to the proteins predicted sizes. The Thermo scientific low range unstained protein ladder (100–3.4 kDa) was used for the FHD gel. The Thermo scientific PageRuler^{TM} Prestained Protein Ladder, 10 to 180 kDa was used for the T–Box gels. Data was plotted using SigmaPlot version 12.

After the final purification steps for WT FHD, WT T–Box, C209A T–Box and C274S T–Box, each collected fraction of protein was resolved using SDS–PAGE. These proteins were observed as clear single bands which were free from other contaminating proteins. The cleaved FHD is predicted to be 10.2 kDa based on its primary structure, which corresponds exactly to the size obtained using SDS–PAGE (**Figure 49**). The TBR1 T–Box domain's predicted size is 25.6 kDa, which is very similar to the sizes obtained for the WT T–Box (24.3 kDa), C209A T–Box (25.9 kDa) and the C274S T–Box (26.2 kDa) domains (**Figure 49**). Furthermore, each protein's absorbance spectrum had the characteristic peak at 280 nm attributed tryptophan and tyrosine residues within the proteins'

sequences. The A_{280}/A_{260} ratios for all the proteins were ~1.7 indicating that they were not contaminated with nucleic acid, which is a concern when purifying nucleic acid binding proteins such as transcription factors (**Figure 48**). Optimisation of purification revealed that the high salt wash included in the IMAC purification step helps remove nucleic acids bound to the immobilised protein.

The overexpression and purification protocols for each protein were relatively consistent, resulting in yields of protein of similar quantity and purity. Typically, ~11 mg of T–Box (WT and mutants) were obtained per litre of culture, whereas 10 to 14 mg of FHD was obtained per litre of culture.

4.4. Structural Characterisation

Structural characterisations (intrinsic tryptophan fluorescence and analytical size exclusion chromatography) were performed on WT FHD, WT T–Box, C209A T–Box and C274S T–Box in order to confirm that each protein was folded correctly. Additionally, there is no TBR1 T–Box domain crystal structure and structural analysis has never been conducted on the TBR1 T–Box domain. Therefore, it was important to conduct these studies to compare them to a predicted T–Box model and other related crystal structures.

4.4.1. Tertiary Structure

Intrinsic tryptophan fluorescence is a commonly used technique to observe the local tertiary structure of a protein around tryptophan residues. The fluorescence spectra of the native protein were compared with protein denatured with 8 M urea (**Figure 50**) to ascertain if the proteins were all folded. Furthermore, each of the cysteine residues in the TBR1 T–Box were mutated in turn to observe if any dimerisation events are disulfide linked. The spectra of the mutants can be compared to the WT to determine if the mutation caused any major tertiary structural change in the protein.

The emission spectra of both the native T–Box mutants overlay with the WT spectra suggesting that there are no significant tertiary structural changes around the tryptophan residues upon mutation of each cysteine residue. This is important as any differences in quaternary structure observed in the mutants when compared to the WT are likely not to be a result of the mutation disrupting the core structure of the protein.

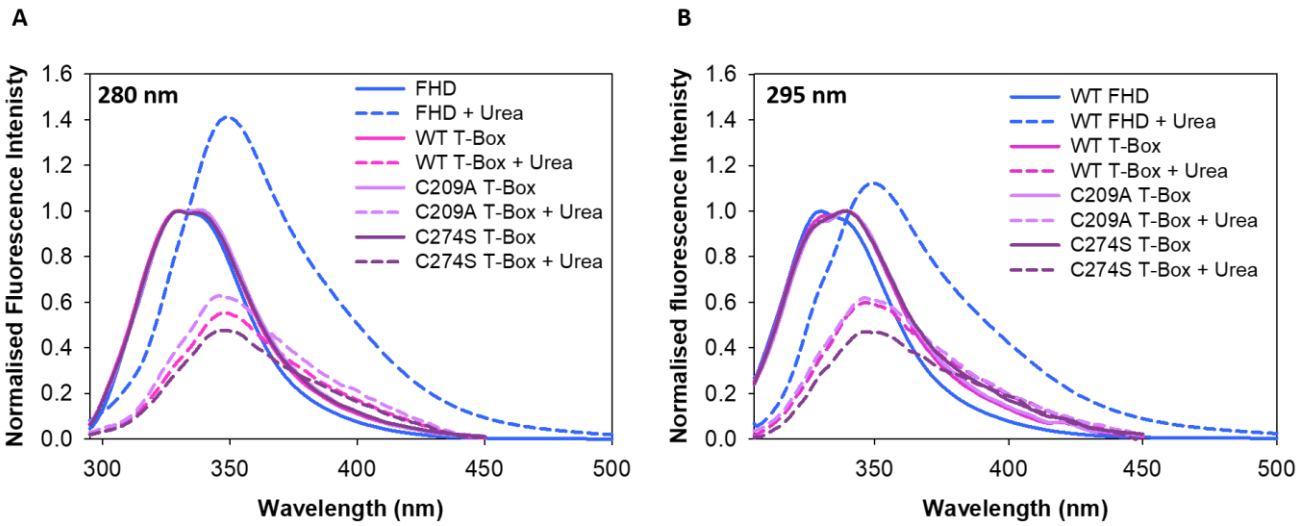


Figure 50: Fluorescence spectra of native and urea denatured WT FHD, WT T–Box, C209A T–Box and C274S T–Box at excitation wavelengths of 280 nm (A) and 295 nm (B).

The fluorescence spectra of native (solid lines) and urea denatured (dashed lines) of WT T–Box (pink), C209A T–Box (lilac), C274S T–Box (dark purple) and WT FHD (blue). Excitation wavelength were (**A**) 280 nm, where tryptophan and tyrosine are excited, and (**B**) 295 nm, where only tryptophan is excited. The native spectra for each protein was normalised with respect to one another and the denture spectra were normalised with respect its corresponding native spectrum. The native spectra for each protein exhibit a shouldered peak suggesting that there are tryptophans in varying polar environments. Data was plotted using SigmaPlot version 12.

Tryptophan residues in a polar environment typically emit at 350 nm, whereas tryptophans in a completely hydrophobic environment emit at 320 nm (Lakowicz 2006). All the non–denatured proteins, at both excitation wavelengths, emitted a fluorescence peak at 329 nm with a shoulder at 338 nm. This indicates there are tryptophan residues in two different polar environments. **Figure 51** displays the model of TBR1 T–Box and a crystal structure of FOXP2 FHD with the tryptophan and tyrosine residues highlighted. TBR1 T–Box has 4 tryptophan residues with an additional one in the tag and 6 tyrosine residues. Of these, Trp264, Trp214, Tyr333 and Tyr395 were calculated to be more solvent exposed than the other tryptophan and tyrosine residues in the model (Fraczkiewicz & Braun 1998). The fluorescence data indicate that there are tryptophan and tyrosine residues in two different polar environments which corresponds to the T–Box model.

When an excitation wavelength of 295 nm is used, the tryptophan residues are exclusively excited. However, when 280 nm is used, both the tryptophan and tyrosine residues are both excited. Tryptophan residues nearby tyrosine residues quench tyrosine fluorescence (Lakowicz 2006). The intensity of a peak is not only due to the number of tryptophan and tyrosine residues in each environment but also the presence quenchers such as charged amino acids and even the solvent itself, so it is difficult to

assign a number of tryptophan residues to each peak (Christov *et al.* 2004; Lakowicz 2006; White 1959). The emission spectra for T–Box when excited at 295 nm is shifted and the signal at the shoulder at 338 nm (more solvent exposed) is greater than that at 329 nm (more buried). When excited at 295 nm there appears to be more fluorescence signal from tryptophan residues in the more polar environment. The smaller signal of the polar peak when excited at 280 nm could be a result of quenching of a greater number of nearby tyrosine residues from tryptophans in the more apolar environment. Based on the T–Box model in **Figure 51**, this could be attributed to the quenching of Tyr207 by relatively buried Trp299.

The FHD has 3 tryptophan residues and 4 tyrosine residues. According to the crystal structure in **Figure 51**, only one tryptophan and 3 of the tyrosine residues are solvent exposed. The data corresponds well to the crystal structure and previous studies (Blane *et al.* 2018) suggesting the FHD is in its native conformation.

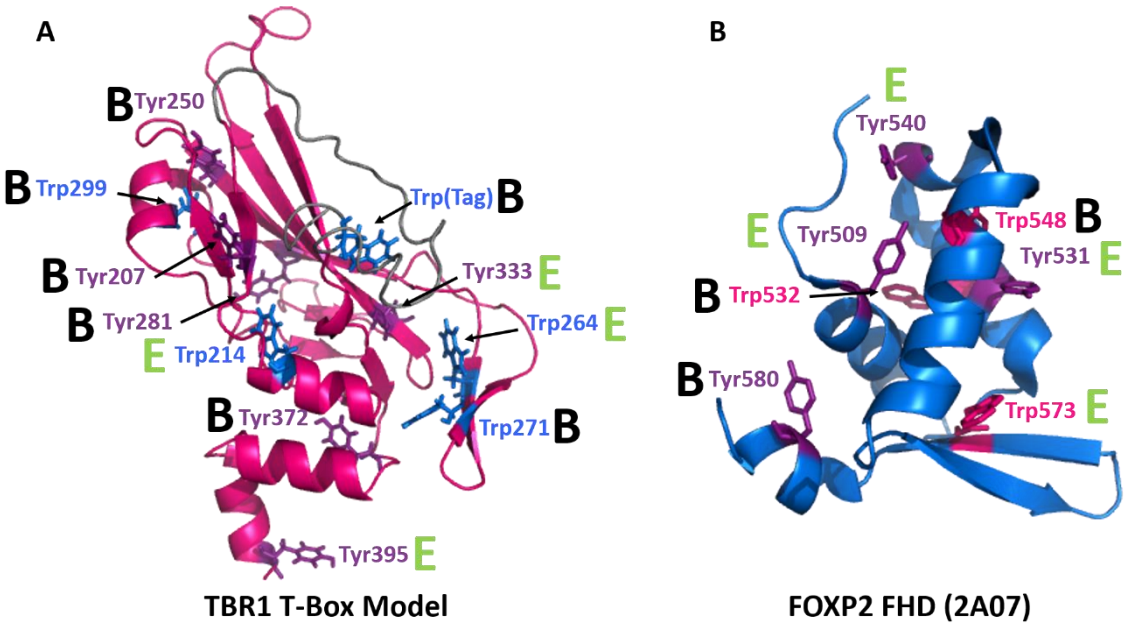


Figure 51: Model and crystal structure of the TBR1 T–Box and FOXP2 FHD displaying the tryptophan and tyrosine residues.

A: The model of the TBR1 T–Box showing the tryptophan residues (blue) and the tyrosine residues (purple). The purification tag is coloured in grey. **B**: Chain K of the crystal structure of FOXP2 FHD (2A07)(Stroud *et al.* 2006) showing the tryptophan residues (pink) and the tyrosine residues (purple). The B (black) indicates a more buried residues, while E (green) indicates a more solvent exposed residue. Protein images rendered with PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.

In the presence of urea for all proteins there is a shift of at least 10 nm and a significant change in fluorescence intensity which shows that the native proteins are folded. The denatured FHD exhibits an increase in fluorescence signal upon denaturation suggesting

that the tryptophan residues in the folded protein are nearby amino acids which can quench fluorescence (Christov *et al.* 2004; Lakowicz 2006; White 1959). The denatured T–Boxes all show a significant amount of quenching in comparison to the native spectra as the protein is unfolded the tryptophan residues' emissions are quenched by the water. The drastic shifts that occurs upon denaturation further exhibits that the proteins were successfully expressed and purified in a folded conformation.

4.4.2. Quaternary Structure

Quaternary structure analysis was performed on the WT FHD, WT T–Box, C209A T–Box and C274S T–Box proteins. Firstly, this was done to ensure for downstream studies that the proteins existed in a single oligomeric state, as a mixture of species would convolute the data. Secondly it is unknown if the TBR1 T–Box domain forms a dimer in solution as seen with the TBX21 T–Box domain; or if it is instead monomeric in solution as seen with other T–Box domains (Liu *et al.* 2016). To further investigate potential TBR1 T–Box dimerisation, each of its two cysteine residues were mutated to an Ala (C209A) and Ser (C274S) in turn to assess if any TBR1 T–Box dimerisation observed is due to the formation of disulfide bonds, and more specifically which cysteine residue is involved. In order to assess the quaternary structure of the WT FHD, WT T–Box, C209A T–Box and C274S T–Box proteins, size exclusion chromatography was used. Analytical size exclusion chromatography was done using a HiLoad™ 16/60 Superdex™ 75 prep grade column (GE healthcare, Sweden). Each of the proteins was analysed at two different concentrations under both reducing and non–reducing conditions.

GE healthcare standards were run through the column under the same conditions as the protein. Blue Dextran was applied to the column separately as it is known to interact with many other proteins which would result in inaccurate sizing (Marshall 1970). Blue dextran was used to determine the void volume of the column as it is much larger than the fractionation range. The blue dextran eluted in a sharp peak with a small amount of tailing which is a good indication that the column was performing well (Hong *et al.* 2012). In **Figure 52A** the elution profile of the size exclusion standards is displayed. The standards eluted in well resolved sharp peaks. The log Mw of each standard was plotted against the ratio of their elution volume (V_e) and the void volume (V_o). A linear regression was fitted with an equation of $Y = 3.3574 – 1.2396\ x$ and an R^2 of 0.9962.

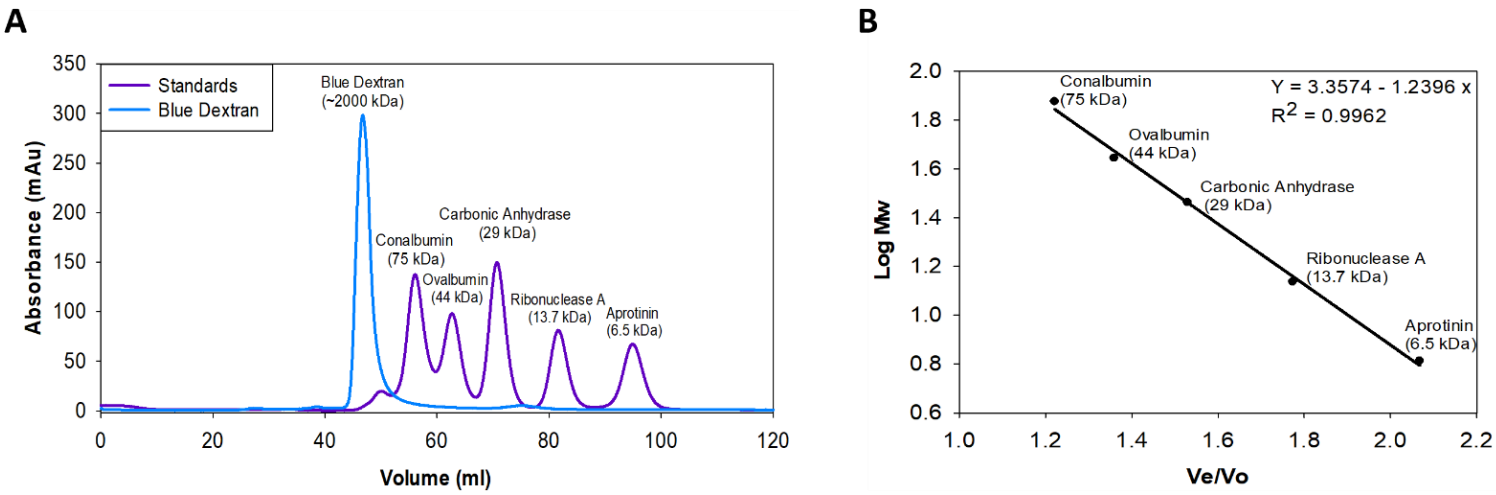


Figure 52: Size exclusion chromatography standards elution profile (A) and calibration curve (B).

GE Healthcare gel filtration standards were run through a HiLoad™ 16/60 Superdex™ 75 prep grade column (GE healthcare, Sweden). In panel **A** the blue dextran (blue) was run separately to the other standards (purple). The GE healthcare standards consisted of conalbumin (75 kDa), ovalbumin (44 kDa), carbonic anhydrase (29 kDa), ribonuclease A (13.7 kDa) and aprotinin (6.5 kDa). In panel **B** V_e/V_o for each standard was plotted against the log of its molecular weight. A linear regression was fitted with an equation of Y = 3.3574 – 1.2396 x and an R^2 of 0.9962. Data was plotted using SigmaPlot version 12.

The WT FHD, WT T–Box, C209A T–Box and C274S T–Box were resolved on the same size exclusion column at 25 and 50 µM under reducing (2 mM DTT) and non–reducing conditions (**Figure 53**). The FOXP2 FHD was almost completely monomeric at both concentrations under reducing and non–reducing conditions as previously described (Blane *et al.* 2018; Blane & Fanucchi 2015; Morris & Fanucchi 2016; Perumal *et al.* 2015).

Under non–reducing conditions, the WT T–Box exists as a mixture of both monomer and dimer. However, under reducing conditions, it is almost exclusively monomeric. This suggests that the small portion of dimeric TBR1 T–Box domain observed is likely a disulfide linked dimer. The TBR1 T–Box domain has only 2 cysteine residues, neither of which is conserved in other T–Box domains. Each cysteine was mutated in turn to give each of the T–Box mutants (C209A and C274S). Both of the T–Box mutants exist predominantly as a monomeric species under both reducing and non–reducing conditions, which indicates that the disulfide linked dimer formed by the WT occurs between Cys 209 in one monomer and Cys 273 in the other monomer. TBR1 is a transcription factor and therefore functions in the nucleus – a reducing environment. This means the disulfide linked dimer is unlikely to be physiologically relevant. Therefore, all other experiments were conducted in the presence of a reducing agent. The sizes of each protein in both monomeric and dimeric form were calculated using

their elution volumes and the calibration curve in **Figure 52**. Each protein was very close to its expected size and is displayed in **Table 14**.

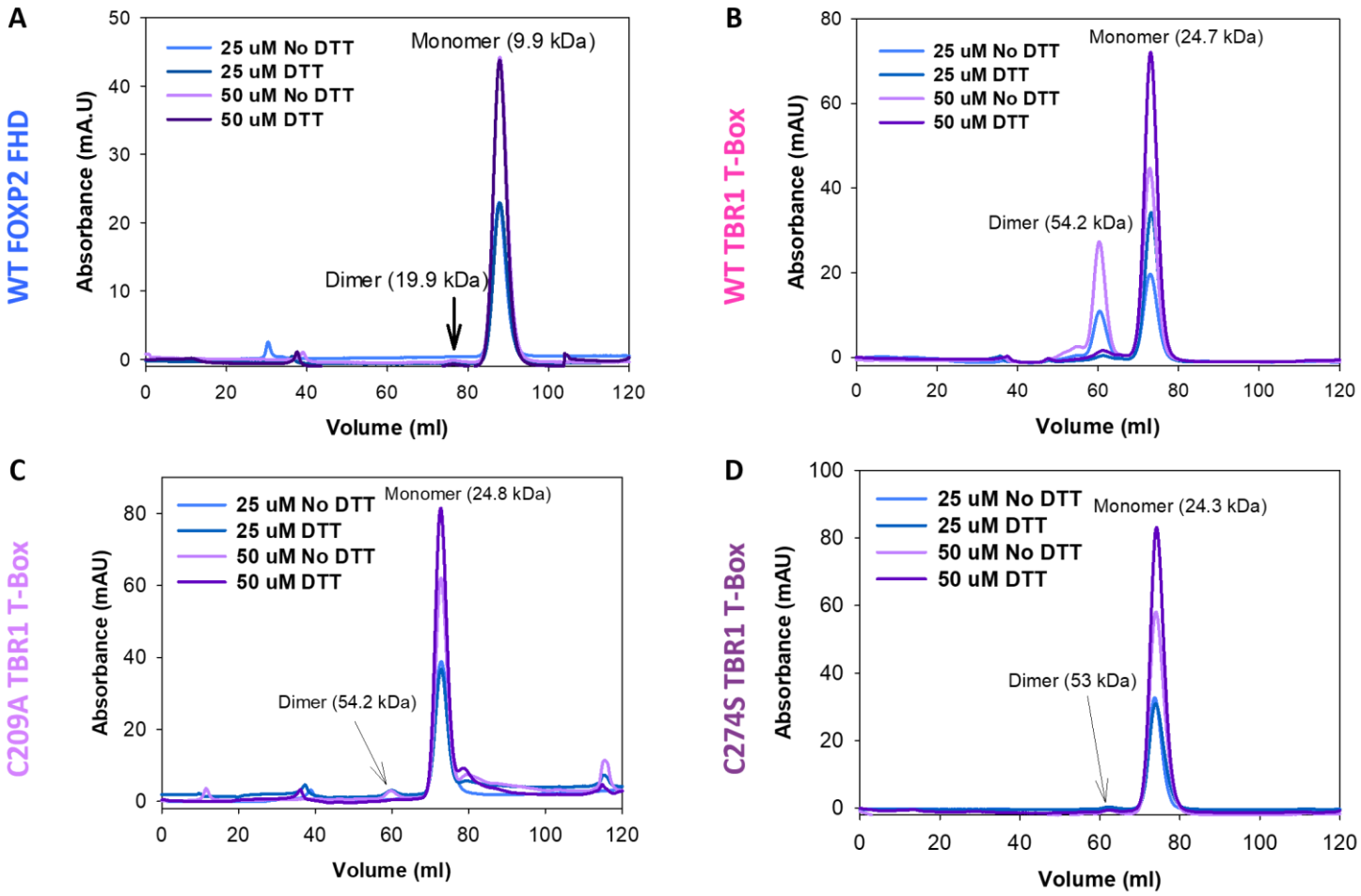


Figure 53: Analytical size exclusion chromatography of WT FHD, WT T–Box, C209A T–Box and C274S T–Box at 25 and 50 µM in reducing and non–reducing conditions.

The WT FHD (blue), WT T–Box (pink), C209A T–Box (lilac) and C274S T–Box (purple) were resolved at 25 and 50 µM on a HiLoad™ 16/60 Superdex™ 75 prep grade column in a buffer consisting of 20 mM Tris–HCl, 500 mM NaCl and 30 mM Imidazole, pH 7.5 with the addition of 2 mM DTT under reducing conditions. The calculated size of each oligomeric state is displayed above and correspond well to the predicted sizes in literature. The FHD and both T–Box mutants were predominantly monomeric is reducing and non–reducing conditions. The WT T–Box exists as mixture of monomer and dimer under non–reducing conditions but is predominantly monomeric in reducing conditions. Data was plotted using SigmaPlot version 12.

Table 14: The calculated size of the WT FHD, WT T–Box, C209A T–Box and C274S T–Box monomer and dimer from analytical size exclusion chromatography

Protein	Monomer (kDa)	Dimer (kDa)	Expected monomer (kDa)
WT FHD	9.9	19.9	10
WT T–Box	24.7	54.2	25
C209A T–Box	24.8	54.2	25
C274S T–Box	24.3	53	25

The size calculated from size exclusion and the expected sizes of the WT FHD, WT T–Box, C209A T–Box and C274S T–Box.

4.4.3. Molecular Modelling

Unfortunately, there are no structures of the TBR1 T–Box domain available in the protein databank. For this reason, the I–TASSER server was used to model our TBR1 T–Box construct. Having a reliable model of the TBR1 T–Box domain could help identify the structural differences and similarities between the TBR1 T–Box domain and other T–Box domains, which in turn could allow insight into the interactions it may form and also aid in interpreting wet lab results.

The I–TASSER sever firstly gives the predicted secondary structure for each residue along with the confidence value using the Protein Secondary Structure Prediction (PSSpred) algorithm (Yan *et al.* 2013). The confidence score ranges between 0 and 9 with a higher score indicating more confidence in the predicted secondary structure. **Figure 54** shows a model of TBR1 T–Box coloured according to the confidence score of the secondary structure prediction. It is clear that most of the structure actually scores quite high. The residues with the lowest confidence score are mostly found on the ends of each structural element. Furthermore, the predicted secondary structure does seem to match up with the secondary structure of T–Box domains.

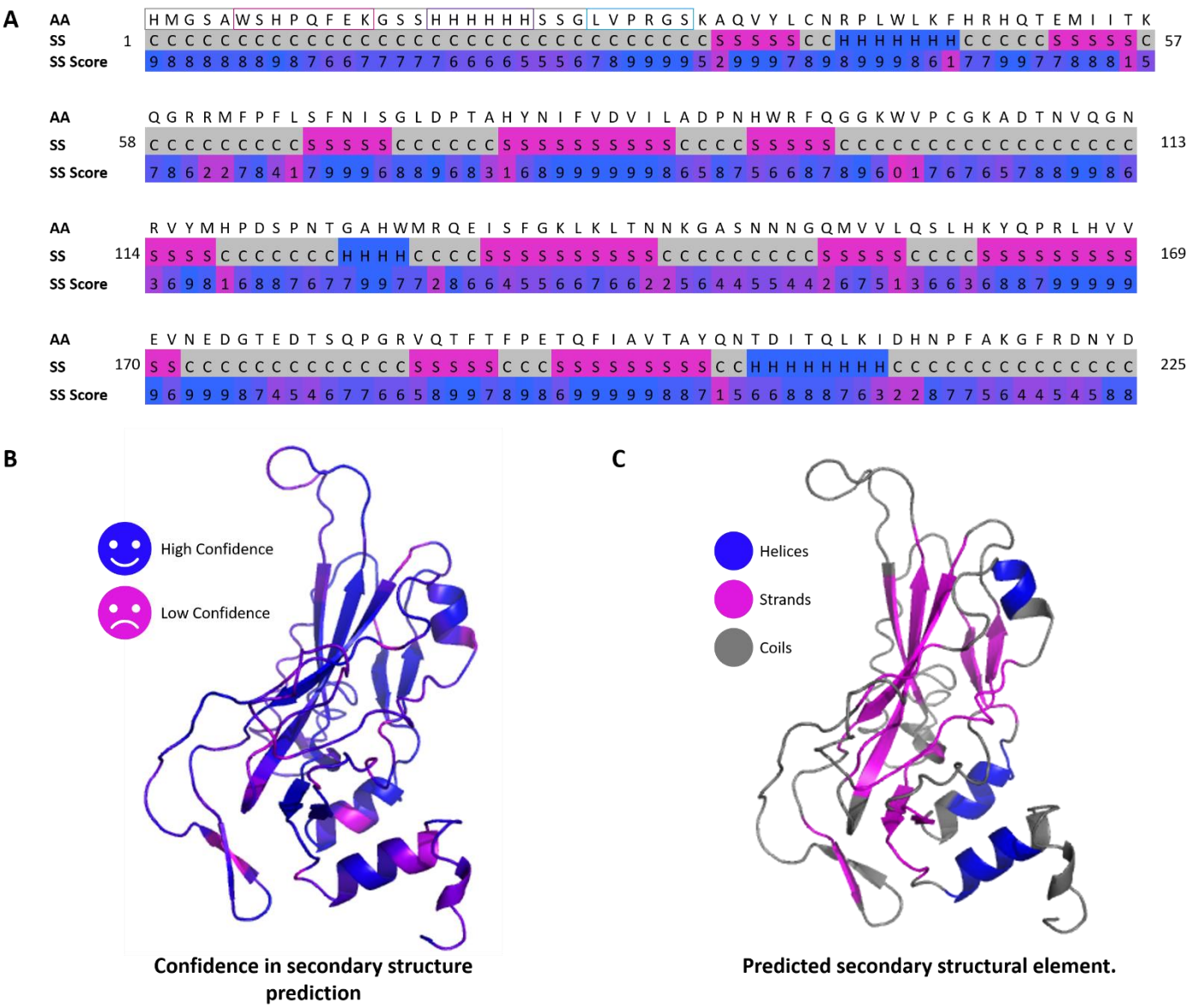


Figure 54: The secondary structure prediction of the TBR1 T–Box using the PSSpred algorithm.

The secondary structure prediction of the TBR1 T–Box domain was performed using the PSSpred algorithm as one of the initial modelling steps used by I–TASSER. **A**: A table showing each residue and its corresponding secondary structural element and confidence score. A higher score (blue) represents a higher confidence than a lower score (pink). **B**: A model of the TBR1 T–Box domain coloured according to the secondary structure prediction confidence score, where blue has a high confidence and pink has a low confidence. **C**: Model 1 coloured by the predicted secondary structure elements (coils– grey, helices– blue, strands– pink). Images rendered with PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.

The I–TASSER server generates multiple conformations using various simulations which are then grouped into clusters. It then reports the best structure from the top 5 clusters. The global accuracy of each model is indicated by the C–score which takes into account the template alignments and the template coverage. The score ranges between –5 and 2 where, the higher the value the more confidence. The top 5 models of this TBR1 T–Box construct (this includes the T–Box domain and our purification tag) generated by I–TASSER have C–scores that range between –0.14 (model 1) and –3.14 (model 5). Model 1 has the highest C–score which suggests that this is the model with the most reliable global structure. I–TASSER also measures the local accuracy of the model which is the deviation between the modelled and native structures. As the native structure is unknown, I–TASSER makes use of various factors such as the coverage of the templates, the predicted secondary structure and predicted solvent accessibility to help identify poorly modelled regions. **Figure 55** displays plots of the local accuracy for each model. From these plots and the alignment in **Figure 56**, it is clear that the only unreliably modelled region in each model is the purification tag. It is likely that the tag is in fact highly mobile and may exist in each of the conformations present in the models. For docking studies, model 1 was chosen as it had the best C–score on I–TASSER and all the regions except for the tag were predicted to be well modelled.

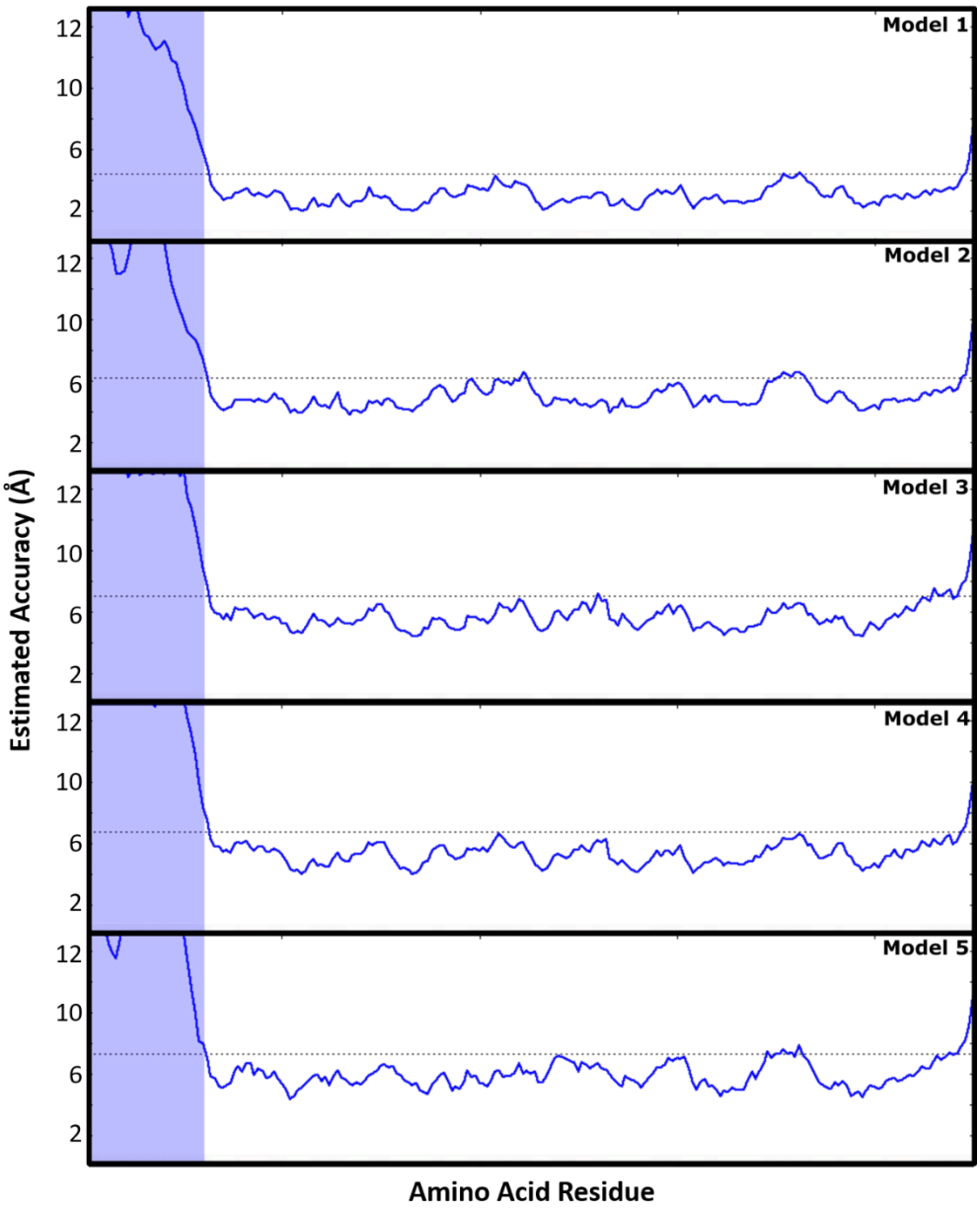


Figure 55: The local accuracy prediction of each T–Box model produced by I–TASSER

The final step of modelling performed by the I–TASSER server is local accuracy prediction. This figure shows the accuracy of the modelling of each residue. The blue blocks highlight the residues that form the purification and cleavage tag. The residues that fall below the dotted line are considered to be modelled accurately. The residues that form part of the purification tag and the last few residues seem to be the only residues that are not modelled accurately. Images generated by I–TASSER and edited using Inkscape 0.92.4.

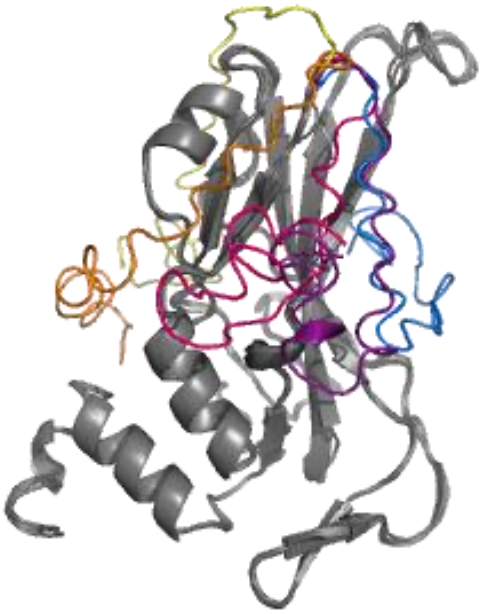


Figure 56: A structural alignment of the 5 models generated by I–TASSER

The 5 models generated by I–TASSER were aligned. The purification tag from each model is coloured (model 1– purple, model 2– blue, model 3– pink, model 4– orange, model 5– yellow). The rest of the structures overlay well (with an RMSD of between 0.36 and 0.52 Å) and were thus all coloured grey. Image rendered with PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.

4.5. Protein Labelling

Obtaining labelled protein was important for performing smFRET and fluorescence anisotropy studies. SmFRET was used to obtain kinetic parameters of the TBR1 T–Box interactions with Pal and SSL DNA, whereas fluorescence anisotropy was used to confirm that the TBR1 T–Box domain interacts with the FOXP2 FHD. For each of these techniques efficiently labelled protein was required to obtain accurate binding parameters. WT TBR1 T–Box was labelled with Alexa Fluor 647 (maleimide reaction) and Atto 550 (NTA via the His–tag) for smFRET and fluorescence anisotropy respectively. Once the labelled protein was separated from the free dye, absorbance spectra were taken **(Figure 57**) and the dye to protein ratios were calculated. A labelling efficiency of 97.4% was achieved for Alexa Fluor 647 while the NTA–Atto 550 only resulted in 45.8% labelling. In addition to using the spectra to calculate protein concentration and labelling efficiency, the spectra were also compared to spectra from the manufacturers to ensure there were no unexpected peaks which could indicate problems with the protein and dyes. These protein stocks were used subsequently for smFRET (Alexa 647) and fluorescence anisotropy (Atto 550). Although, the labelling efficiency was poor for the NTA–Atto 500 the protein was still used for fluorescence anisotropy. Assuming the NTA–Atto 550 does not alter the interaction between T–Box and FHD the poorly labelled protein could still give a good indication of the dissociation constant. The poor labelling

efficiency was accounted for by adjusting the reported concentration of FHD using **Equation 11**.

Equation 11: $[Corrected\ FHD] = T - Box\ labelling\ efficiency/100$

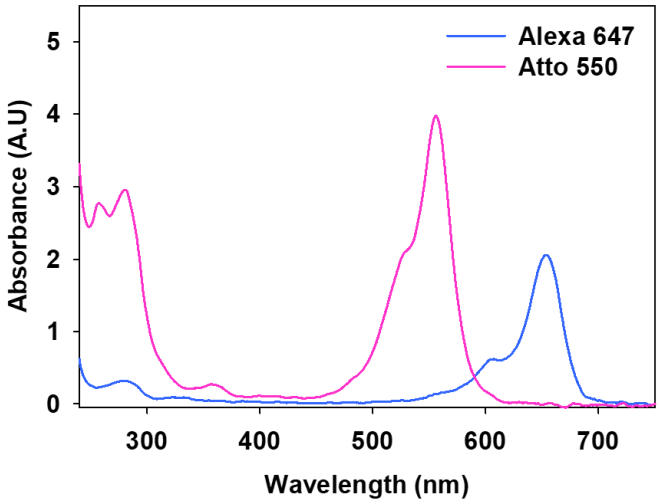


Figure 57: Absorbance spectra of WT T–Box labelled with Alexa 647 and NTA–Atto 550.

Absorbance spectra were taken of WT T–Box labelled with Alexa 647 (blue) and NTA–Atto 550 (pink) using a Thermo Scientific™ NanoDrop™ One Spectrophotometer. The T–Box was labelled with Alexa Fluor 647 and NTA–Atto 550 with labelling efficiencies of 97.4% and 45.8% respectively. The Alexa 647 labelled T–Box (blue) exhibited peaks at 280 nm (absorbance of tryptophan and tyrosine residues) as well as 605 nm and 650 nm (absorbance of Alexa Fluor 647). The Atto 550 labelled T–Box exhibited 3 peaks in addition to the protein absorbance at 280 nm. These were at 255, 525 and 555 nm. These peaks all correspond to a normal Atto 550 absorbance spectrum. Data plotted using SigmaPlot version 12.

4.6. DNA Hybridisation

Forward and reverse strands of Pal and SSL containing Cy3 and biotin were hybridised as previously described in order to contain DNA that was labelled with Cy3 on one end and biotin on the other for smFRET. This was necessary as smFRET require the biotin on one end to immobilise it on the slide and the other end to be labelled with Cy3, a fluorophore that can be excited by the smFRET laser. After the DNA was annealed it was run on a native PAGE (**Figure 58**) and the band corresponding to the double stranded DNA labelled with Cy3 and Biotin was cut out, precipitated, and used in smFRET studies. In **Figure 58**, DNA labelled with Cy3 was visible. The top band in each lane corresponds to double stranded DNA labelled with both Cy3 and biotin while the lower down bands consist of single stranded DNA labelled with Cy3. The bands of single stranded DNA labelled with biotin are not visible. The single stranded Pal DNA

likely runs lower as it may fold in half and interact with itself as it is a palindrome and thus its half sequences are complimentary to one another.

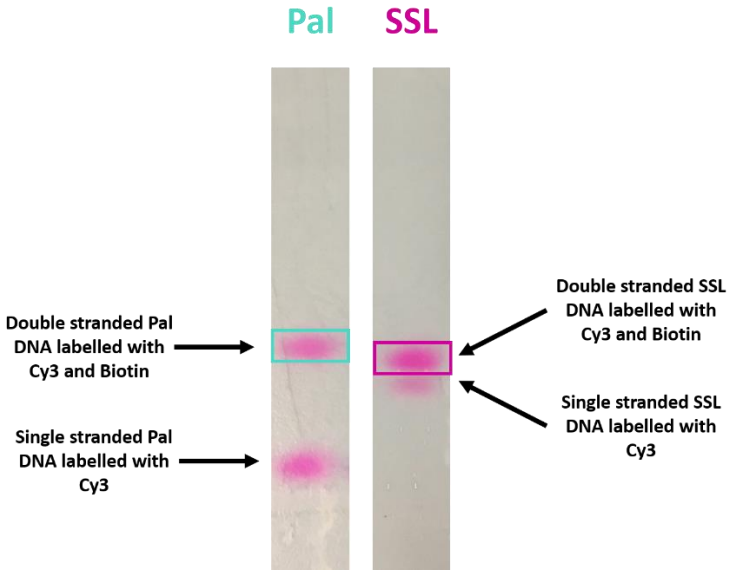


Figure 58: Hybridisation of Cy3 and biotin labelled SSL and Pal DNA.

Single stranded forward and reverse strands labelled with either Cy3 or biotin on the 3' end were annealed to create Pal and SSL DNA labelled with biotin on one end and Cy3 on the other end. The annealed Pal and SSL DNA was resolved on a 10% Native PAGE. The top band in each lane (highlighted in the cyan and pink blocks) contains the correctly annealed DNA while the lower band contains unannealed Cy3 labelled single stranded DNA.

4.7. Binding Studies

4.7.1. Basic Electrophoretic Mobility Shift Assays

The FOXP2 FHD DNA binding has been quite well studied, on the other hand not much is known about the TBR1 T–Box DNA binding. Firstly, an EMSA of the WT FHD, WT T–Box, C209A T–Box and C274S T–Box was performed with each protein's cognate binding sequence so as to ensure that the proteins are correctly folded and active (**Figure 59**). The FHD was shown to be capable of binding the Nelson DNA (Nel) and the WT T–Box, C209A T–Box and C274S T–Box were all three shown to bind DNA containing a single T–Box binding site (SSL). Both the T–Box mutants were run on the EMSA gel alongside the WT T–Box for comparison. All proteins (both wildtype and mutants) show binding to their cognate sequence. The T–Box mutants appear to bind the SSL DNA in the same manner as the WT T–Box.

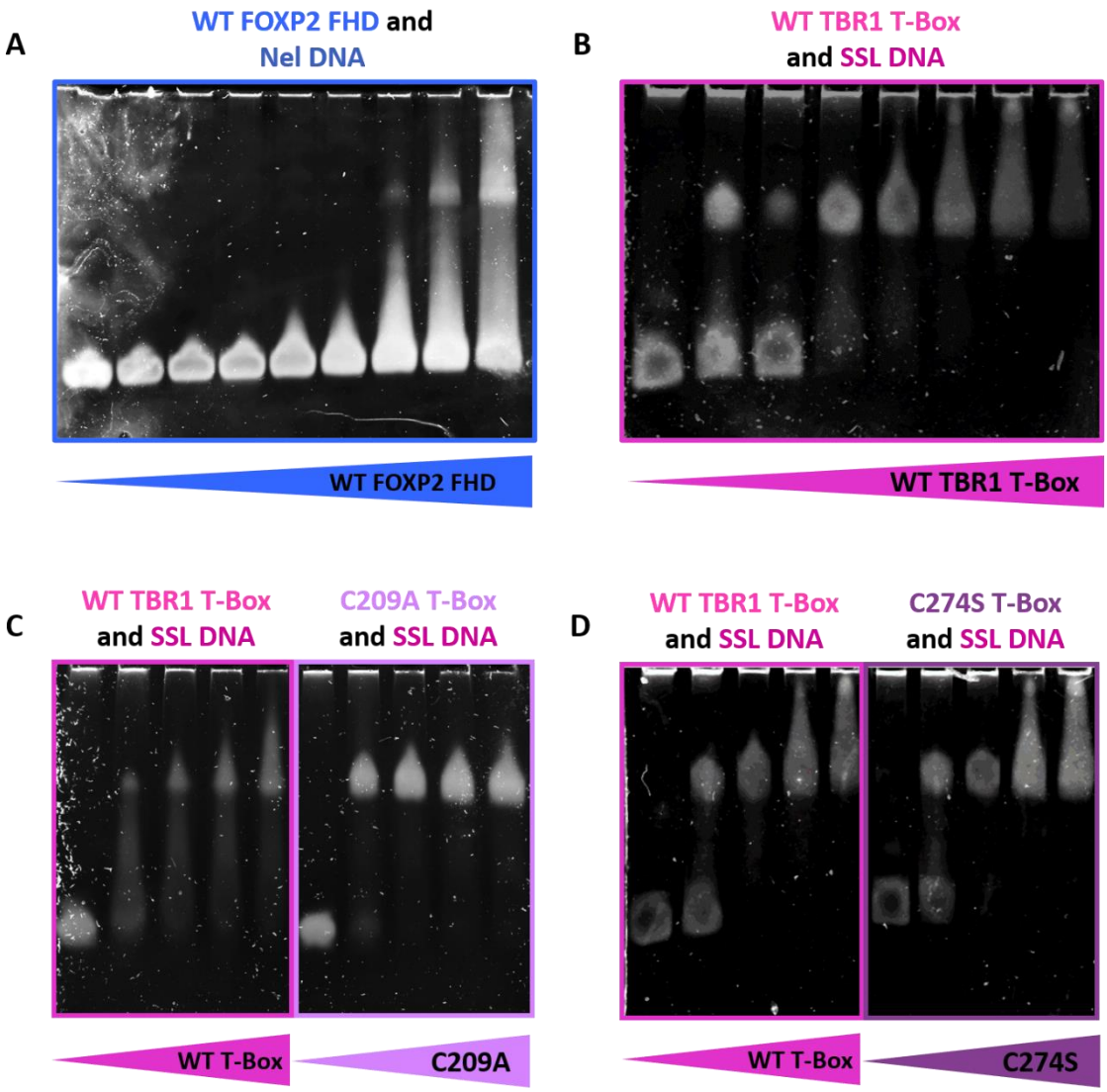
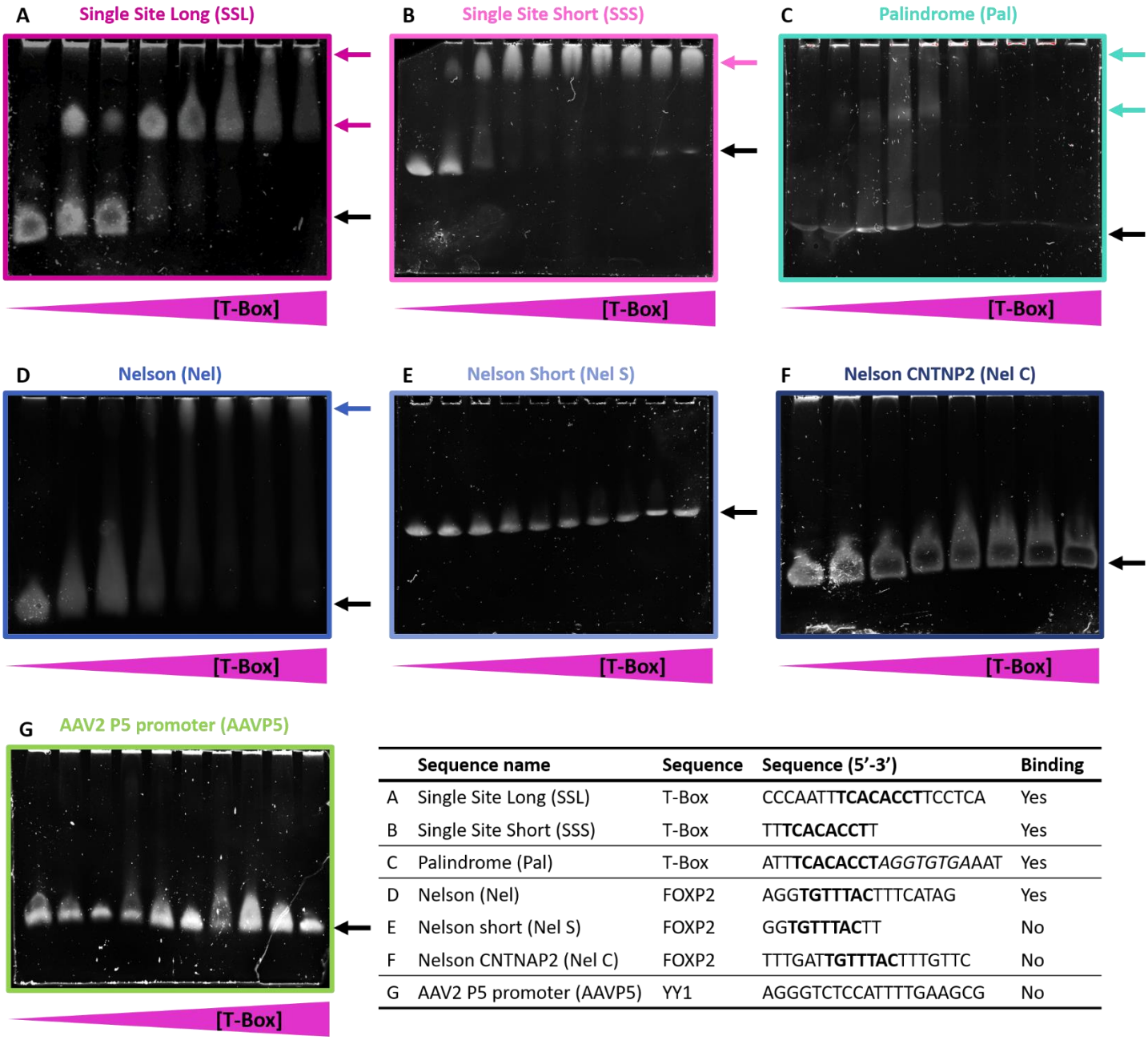


Figure 59: Electrophoretic mobility shift assay showing the WT FHD bound to Nel DNA and the WT T–Box, C209A T–Box and the C274S T–Box bound to SSL DNA.

Electrophoretic mobility shift assays were prepared by running samples with 0.5 µM DNA and increasing concentrations of protein (0 to 20 µM) on a 10% continuous native PAGE. **A:** The FOXP2 FHD was run with its known DNA binding sequence (Nel). **B:** The WT T–Box was run with DNA containing a single T–Box biding site (SSL). In panels **C** and **D** the T–Box mutants were run alongside the WT T–Box with SSL DNA.

The DNA binding specificity of the WT FHD and WT TBR1 T–Box domains was also tested with other DNA sequences to ensure they bind their own DNA sequences specifically and also to ensure that when the technique of fluorescence anisotropy was used, the FHD would not interact with T–Box DNA (SSL, SSS and Pal). In **Figure 60**, the EMSAs of WT T–Box binding to a variety of DNA sequences is displayed. The WT T–Box appears to bind well to SSL, SSS, Pal and Nel sequences. This shows that T–Box is folded and functional. The EMSAs for WT T–Box binding to both the SSL (**Figure 60A**) and Pal (**Figure 60C**) sequences show an additional band. This suggests that in addition to the band

formed by a protein molecule binding to a DNA molecule there is also a larger complex forming. For the Pal sequence, which contains two binding sites, the simplest explanation would be that two protein molecules bind the DNA via both the sites on the palindrome, forming a Protein–DNA–protein complex. However, the SSL DNA only has one binding site, and so the higher band on the gel could be the result of a dimeric T–Box binding DNA forming a protein–protein–DNA complex. Interestingly, T–Box is seen binding the Nel sequence and not the Nel S or Nel C sequences. This means that it is likely that T–Box does not actually bind the Nelson binding site but rather interacts with some of the flanking regions on the Nel DNA.

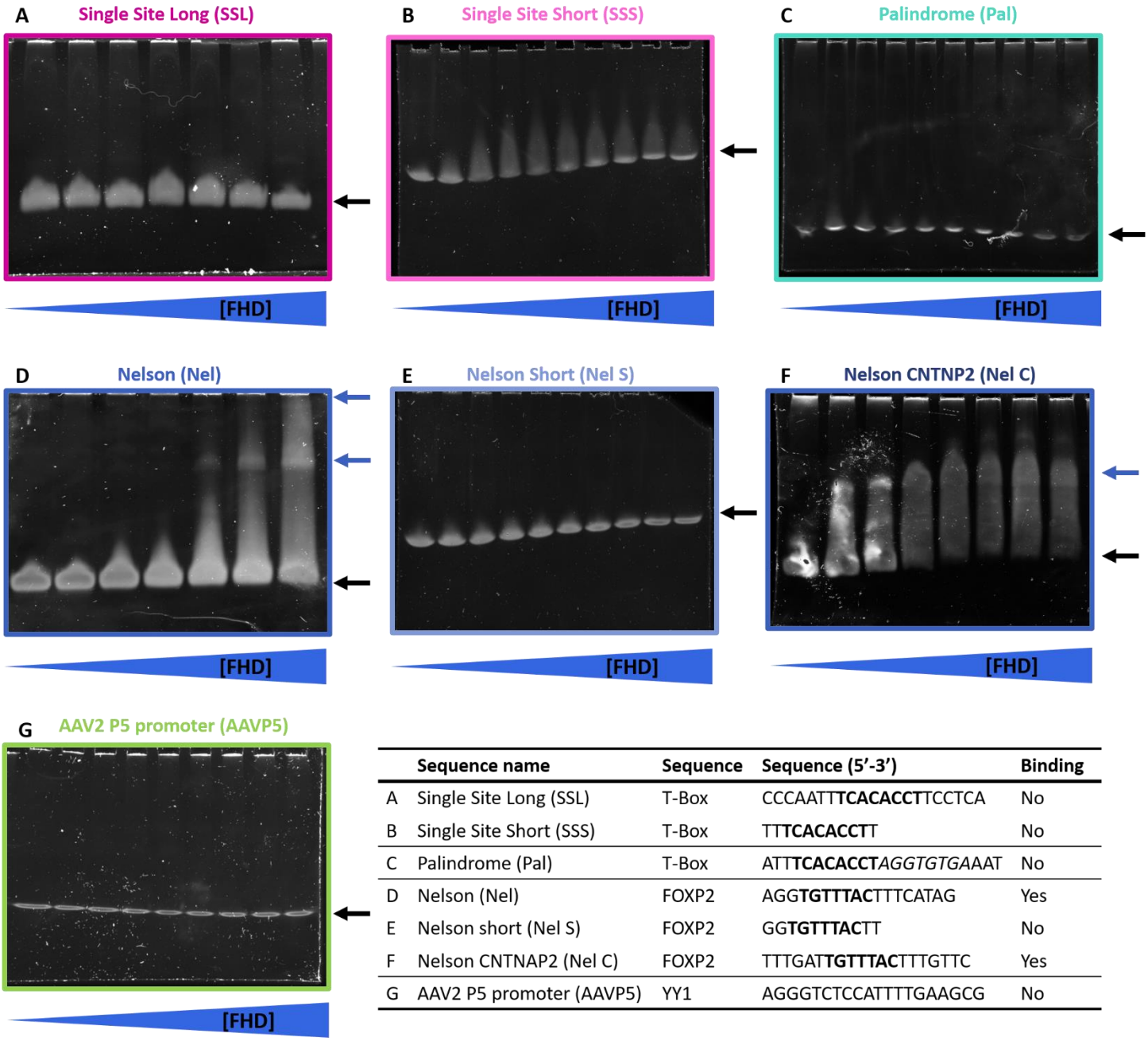


	Sequence name	Sequence	Sequence (5'-3')	Binding
A	Single Site Long (SSL)	T-Box	CCCAATT**TCACACCT**TCCTCA	Yes
B	Single Site Short (SSS)	T-Box	TT**TCACACCT**T	Yes
C	Palindrome (Pal)	T-Box	ATT**TCACACCT***AGGTGTGA*AAT	Yes
D	Nelson (Nel)	FOXP2	AGG**TGTTTAC**TTTCATAG	Yes
E	Nelson short (Nel S)	FOXP2	GG**TGTTTAC**TT	No
F	Nelson CNTNAP2 (Nel C)	FOXP2	TTTGAT**TGTTTAC**TTTGTTC	No
G	AAV2 P5 promoter (AAVP5)	YY1	AGGGTCTCCATTTTGAAGCG	No

Figure 60: Electrophoretic mobility shift assay of WT TBR1 T–Box binding and multiple DNA sequences.

Electrophoretic mobility shift assays of the WT T–Box were done with (**A**) SSL, (**B**) SSS, (**C**) Pal, (**D**) Nel, (**E**) Nel S, (**F**) Nel C and (**G**) AAVP5 DNA sequences on 10% native PAGE and were stained using SyBr Gold. Binding of the WT T–Box to all the T–Box sequences (SSL, SSS and Pal) and the FOXP2 consensus sequence (Nel) was observed. T–Box did not bind to Nel S, Nel C and AAVP5 sequences. The black arrows indicate the free DNA band on each gel and the coloured arrows indicate bands on the gel that contain a protein–DNA complex. In the cases of SSL (**A**) and Pal (**C**) there are two protein–DNA complexes.

In **Figure 61**, the EMSAs of WT FHD binding to a variety of DNA sequences is displayed. The WT FHD appears to bind well only to Nel and Nel C sequences. This shows that FHD is folded and functional. The fact that the FHD does not bind to Nel S, a sequence that contains the same binding site as Nel and Nel C, only with 2 bp as flanking regions suggests that flanking regions play an important role in FHD DNA binding. In the Nel EMSA two bands are apparent in addition to the free DNA band. This suggests that the FOXP2 FHD may be binding DNA in a dimeric form. Additionally, FHD does not bind any of the T–Box DNA or the AAV2 P5 promoter DNA which shows that FHD binds DNA with specificity.



	Sequence name	Sequence	Sequence (5'-3')	Binding
A	Single Site Long (SSL)	T-Box	CCCAATT**TCACACCT**TCCTCA	No
B	Single Site Short (SSS)	T-Box	TT**TCACACCT**T	No
C	Palindrome (Pal)	T-Box	ATT**TCACACCT***AGGTGTGA*AAT	No
D	Nelson (Nel)	FOXP2	AGG**TGTTTAC**TTTCATAG	Yes
E	Nelson short (Nel S)	FOXP2	GG**TGTTTAC**TT	No
F	Nelson CNTNAP2 (Nel C)	FOXP2	TTTGAT**TGTTTAC**TTTGTTC	Yes
G	AAV2 P5 promoter (AAVP5)	YY1	AGGGTCTCCATTTTGAAGCG	No

Figure 61: Electrophoretic mobility shift assay of WT FOXP2 FHD binding and multiple DNA sequences.

Electrophoretic mobility shift assays of the WT FHD were done with (**A**) SSL, (**B**) SSS, (**C**) Pal, (**D**) Nel, (**E**) Nel S, (**F**) Nel C and (**G**) AAVP5 DNA sequences on 10% native PAGE and were stained using SyBr Gold. Binding of the WT FHD to Nel and Nel C was observed. WT FHD did not bind to any of the T–Box sequences, the AAVP5 or Nel S. The black arrows indicate the free DNA band on each gel and the coloured arrows indicate bands on the gel that contain a protein–DNA complex.

4.7.2. Isothermal Titration Calorimetry

Isothermal titration calorimetry was done in order to obtain quantitative binding parameters of the WT T–Box binding to Palindrome (Pal) and Single site (SSL) DNA as DNA binding for the TBR1 T–Box domain has never been characterised before. A thermodynamic profile of TBR1 T–Box domain DNA binding could give important insights into its DNA binding mechanism in solution. Binding isotherms of the TBR1 T–Box binding to both Pal and SSL DNA sequences at two different salt concentrations were obtained in triplicate. An example of an isotherm from each condition is displayed in **Figure 62** below. The isotherms all display pre and post transition regions which means that the change in enthalpy can be reliably obtained. Furthermore, While the isotherms all have pre and post transition regions, these are less clear in the 250 mM NaCl Pal titration. This means that the change in enthalpy for this titration is less reliable than the values obtained from the others. The titrations all fit well to the independent model.

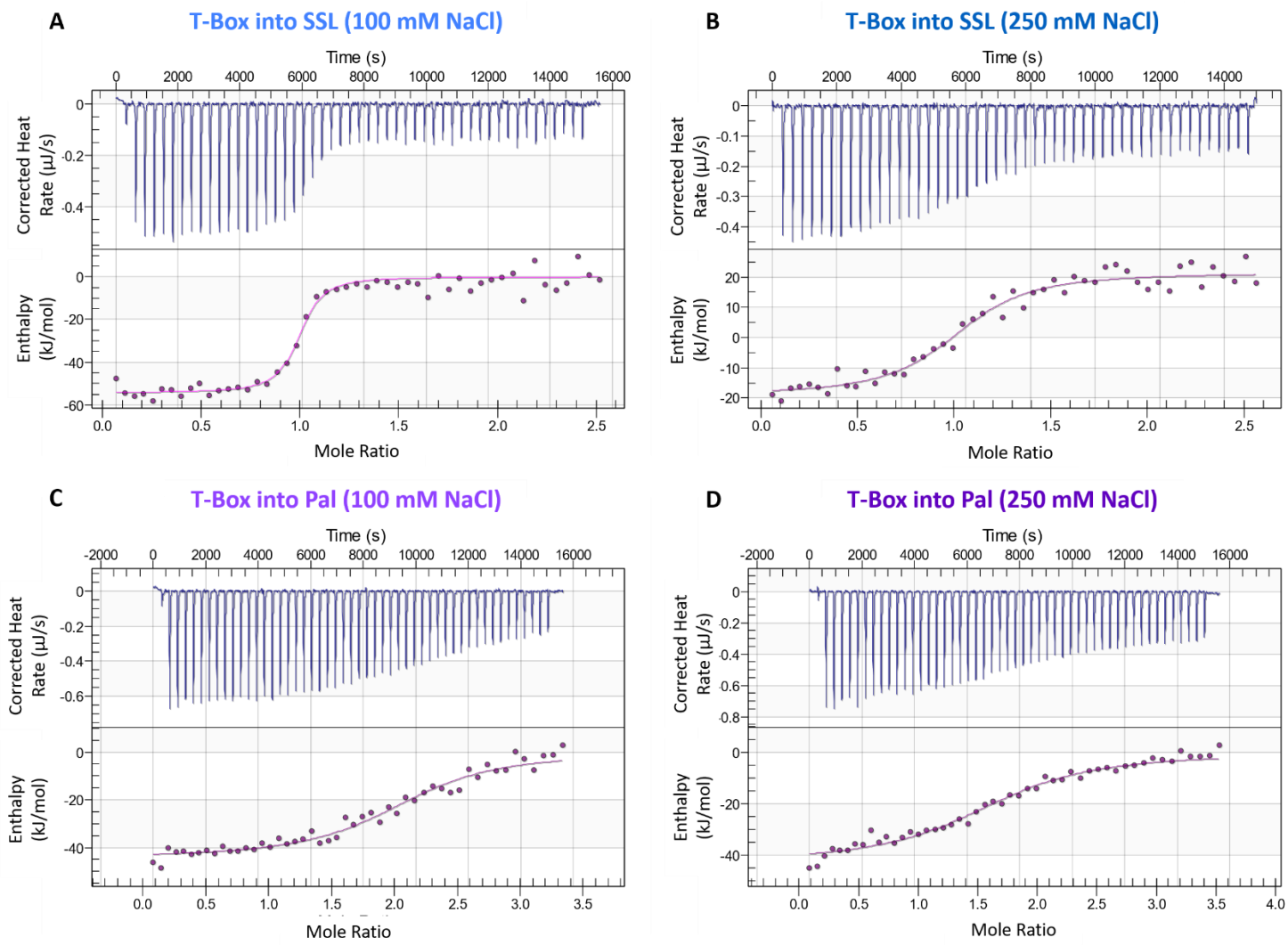


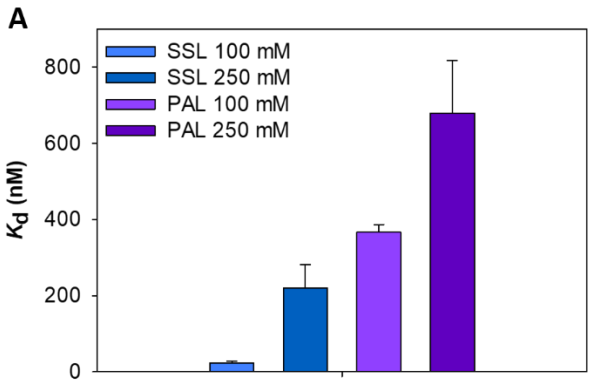
Figure 62: Isotherms of WT T–Box titrated into SSL and Pal with NaCl concentrations of 100 and 250 mM.

The WT TBR1 T–Box was titrated into both Pal and SSL DNA at 2 different NaCl concentrations (100 and 250 mM). The titrations were done on a Nano isothermal titration calorimeter (TA instruments, USA) in triplicate. The buffer consisted of 10 mM sodium phosphate, 3 mM TCEP and either 100 or 250 mM NaCl, pH 7.5. The data was fitted using the NanoAnalyze software to an independent model. One of the replicates for each titration: (**A**) SSL at 100 mM NaCl, (**B**) SSL at 250 mM NaCl, (**C**) Pal at 100 mM NaCl and (**D**) Pal at 250 mM NaCl, is displayed above. While the isotherms all have pre and post transition regions, these are less clear in the 250 mM NaCl Pal titration. This means that the change in enthalpy for this titration is less reliable than the values obtained from the others. The titrations all fit well to the independent model.

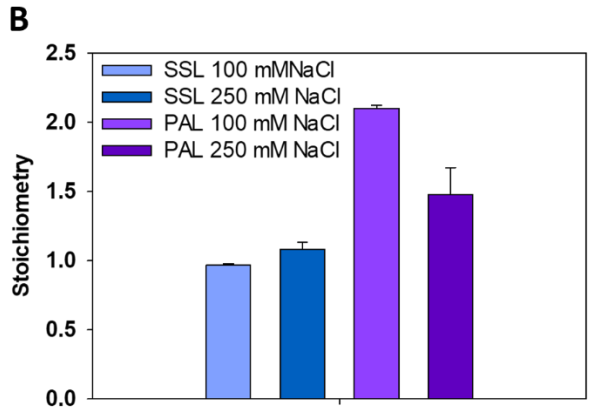
The parameters obtained from titrating WT T–Box into SSL and Pal DNA at 100 and 250 mM NaCl are displayed in **Figure 63**. The dissociation constants (K_d) for T–Box interacting with SSL at 100 and 250 mM NaCl are 23 and 220 nM respectively, showing that binding is dependent on NaCl concentration. The same trend is observed for the Pal DNA where

the K_d at 100 mM NaCl is 367 nM and increased to 679 nM at 250 mM NaCl. T–Box also binds SSL at a much higher affinity than Pal which is interesting because the binding site in SSL is the same as the two binding sites in Pal. This shows that the flanking regions are important, and it is possible that T–Box might not be able to bind Pal as stably as there is no space between the binding sites and only 2 amino acids flanking the other ends of the binding sites. The Stoichiometries (**Figure 63B**) for SSL at both NaCl concentrations are close to 1 which means one T–Box binds to one DNA molecule as expected. The stoichiometries for the Pal sequence is NaCl concentration dependent. At low NaCl concentration the stoichiometry is 2.1 and it decreases to 1.5 at 250 mM NaCl. This means that at low salt concentrations both binding sites on the Pal DNA are occupied whereas, at higher NaCl concentrations, a mixture exists where on some DNA molecules both sites are occupied and on others only one is occupied.

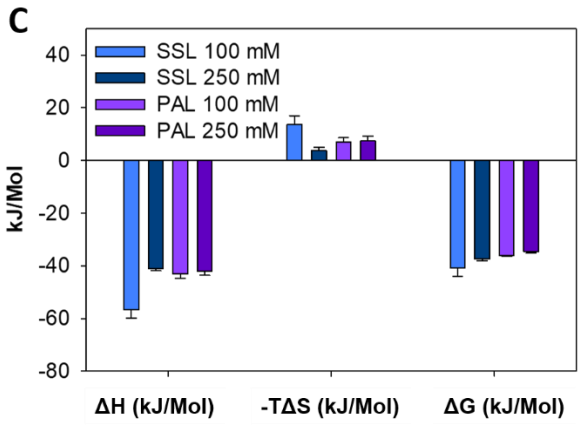
The thermodynamic parameters (ΔG, ΔH and ΔT) for each DNA sequence and salt concentration are very similar meaning that the binding mode seems to be the same irrespective of salt concentration and DNA sequence (**Figure 63C**). The interaction is enthalpically driven as the ΔH values for SSL at 100 mM NaCl, SSL at 250 mM NaCl, Pal at 100 mM NaCl and Pal at 250 mM NaCl are –56.7, –41.1, –43 and –47.1 kJ/mol respectively. The –TΔS terms range between 3.7 kJ/mol (SSL at 250 mM NaCl) to 13.8 kJ/mol (SSL at 100 mM NaCl) meaning the interaction is entropically unfavourable due to a decrease in conformational freedom upon T–Box binding DNA. The ΔG values are all negative (SSL at 100 mM NaCl is –40.8, SSL at 250 mM NaCl is –37.4, Pal at 100 mM NaCl is –36.1 and Pal at 250 mM NaCl is –34.7 kJ/mol) as the interaction is spontaneous.



	K_d (nM)	std deviation
SSL (100 mM)	23	5
SSL (250 mM)	220	62
Pal (100 mM)	367	20
Pal (250 mM)	679	138



	n	std deviation
SSL (100 mM)	0.97	0.01
SSL (250 mM)	1.08	0.05
Pal (100 mM)	2.1	0.02
Pal (250 mM)	1.5	0.2



	ΔH (kJ/mol)	std deviation	-TΔS (kJ/mol)	std deviation	ΔG (KJ/mol)	std deviation
SSL (100 mM)	-57	3	14	3	-41	3
SSL (250 mM)	-41.1	0.8	4	1	-37.4	0.6
Pal (100 mM)	-43	2	7	2	-36.1	0.1
Pal (250 mM)	-42	1	8	2	-34.7	0.5

Figure 63: Thermodynamic parameters obtained from titrating WT TBR1 T–Box into SSL and Pal DNA at 100 and 250 mM NaCl.

Thermodynamic parameters obtained from titrating WT T–Box into SSL and Pal DNA at 100 and 250 mM NaCl. The isotherms were fitted to an independent model and the obtained dissociation constant (**A**), stoichiometry (**B**), ΔH (**C**), –TΔS (**C**) and ΔG (**C**) values are displayed. The titrations were done on a Nano isothermal titration calorimeter (TA instruments, USA) in triplicate. The buffer consisted of 10 mM sodium phosphate, 3 mM TCEP and either 100 or 250 mM NaCl, pH 7.5. In **A,** it is clear that binding affinity to both DNA sequences is salt dependent and that T–Box binds SSL with a greater affinity that it does Pal DNA. In **B** T–Box is seen binding SSL DNA with a stoichiometry of 1 while the stoichiometry when binding to PAL DNA is salt dependent. In **C** the thermodynamic parameters are all similar for T–Box binding

both Pal and SSL at both ionic strengths. The binding is enthalpically favourable and entropically unfavourable. The data was fitted using the NanoAnalyze software to an independent model and plotted using SigmaPlot version 12.

4.7.3. Single Molecule–Förster Resonance Energy Transfer

smFRET studies were done to observe the interaction between immobilised Cy3 labelled DNA and Alexa647 labelled TBR1 T–Box. SmFRET was done in addition to ITC in order to obtain kinetic data to help understand the T–Box DNA binding mechanism. Both Pal and SSL DNA were used, however, unfortunately a full dataset for the SSL sequence was not obtained due to lockdown restrictions for the Covid 19 pandemic and therefore, this section will focus on the TBR1 T–Box interaction with Pal DNA.

4.7.3.1. Palindrome DNA

The channel on the prepared slide was checked for any contaminants before loading 20 pM Pal labelled DNA. **Figure 64** shows the comparison between the empty channel (**A**) and the channel containing immobilised DNA (**B).** There were no spots detected on the empty channel and thus it was deemed clean enough for use. Panel B in **Figure 64** shows the spots obtained when 20 pM Cy3 Pal DNA was immobilised on the slide and washed with imaging buffer. The spots are apparent on the left–hand side of the panel as that is the side which detects donor fluorescence (the DNA labelled with Cy3).

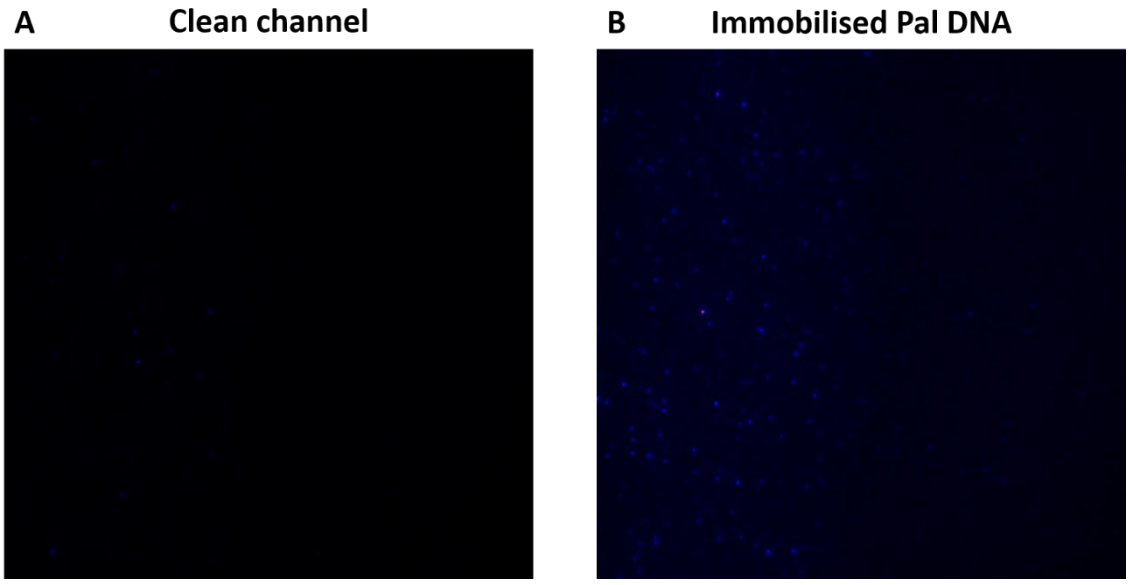


Figure 64: Channel fluorescence with the Cy3 laser before and after immobilisation of 20 pM Pal DNA.

A laser was used to excite the donor fluorophore (Cy3 conjugated to DNA) in the following smFRET experiments. Before immobilisation of the DNA, the channel was excited with the Cy3 laser to ensure it was clean (**A**). After immobilisation of 20 pM Pal DNA, the channel was excited again (**B**) so as to ensure sufficient amounts of DNA were successfully immobilised and well labelled. The Cy3 laser was used to excite the Cy3 labelled DNA. The left half of each image is the Cy3 specific half of the detector. The right–hand side is the Cy5 (or Alexa647) specific half of the detector. In **A** there are almost no spots, showing the channel was clean before use. In **B,** the immobilised Cy3 labelled Pal DNA is visible on the Cy3 side of the detector, while the Cy5 side is clear, indicating the absence of any contaminants causing FRET.

Once the DNA was immobilised on the slide it was washed with increasing concentrations of Alexa 647 labelled T–Box. After the wash of each T–Box concentration, data was collected in order to observe the differences in frequency, duration and FRET intensity due to varying concentrations of T–Box on the T–Box DNA interaction. **Figure 65** displays a representative trace at each concentration of protein that was used. The donor emission decreases as it is absorbed by the acceptor and results in an increase in acceptor emission when there is an interaction between the Cy3 labelled DNA (donor) and Alexa 647–labelled T–Box (acceptor). As the concentration of T–Box increases there is an increase in the number of interactions.

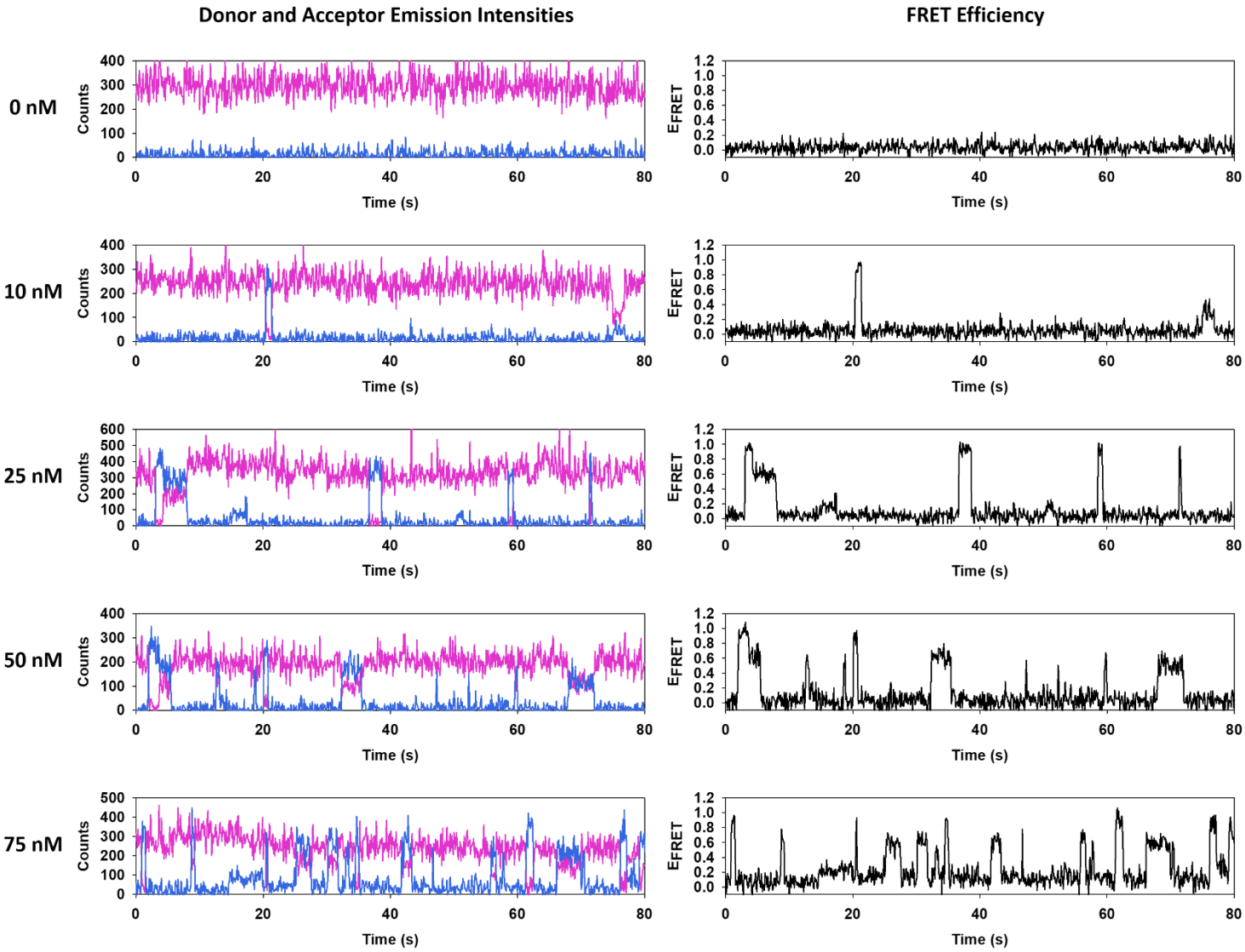


Figure 65: Representative donor and acceptor emission traces and FRET efficiency traces of immobilised Cy3 Pal DNA with increasing concentrations of Alexa 647 labelled TBR1 T–Box.

On the left, the traces of the donor and acceptor emissions intensities are displayed. The pink traces are the donor (Cy3–labelled Pal DNA) emissions while the blue traces are the acceptor (Alexa 647–labelled T–Box) emissions. The FRET efficiency determined from the donor and acceptor emissions intensities is displayed on the right. Each time an interaction occurs between an immobilised DNA molecule and T–Box, the donor emission is absorbed by the acceptor and appears as an anti–correlated drop of the donor emission and increase in the acceptor emission. As the concentration of T–Box is increased, there is a notable increase in the number of interactions that occur. Data was plotted using SigmaPlot version 12.

The interaction between TBR1 T–Box and Pal DNA appears to occur at two different FRET efficiencies, a high FRET state (~ 0.9) and a low FRET state (~ 0.6). FRET efficiency is proportional to the distance between the donor and acceptor fluorophores and since there are two different FRET efficiencies observed they represent two different binding

modes. Since WT T–Box has two Cys residues it is possible that the two FRET states could be a caused by various combinations of labelled Cys residues this, however, is unlikely as only a single FRET state is observed for smFRET studies of WT T–Box binding SSL DNA (**Section 4.7.3.1.**). **Figure 66** depicts 4 traces showing the various modes in which T–Box binds Pal DNA. In **Figure 66A**, T–Box binds Pal DNA at a low FRET state and then transitions, without dissociating, to a high FRET state. The opposite can also happen (**Figure 66B** and **C**), where T–Box binds Pal DNA in the high FRET state and transitions to the lower FRET state. The transition from high to low FRET state appears more often than the transition from low to high. Additionally, T–Box can bind DNA in both the high and low FRET states and dissociate without transition to the other state (**Figure 66D**).

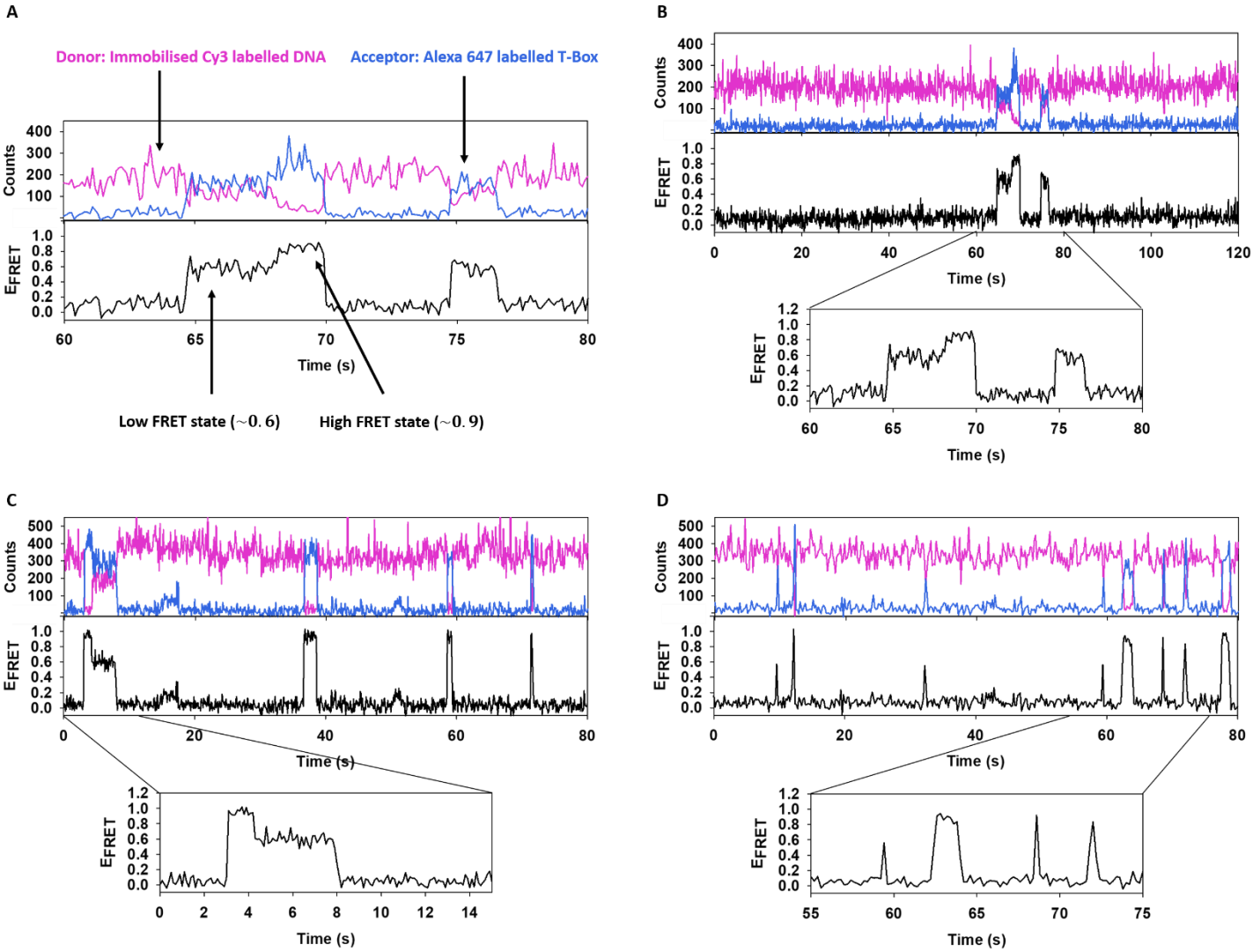


Figure 66: Examples of traces showing the various FRET states in which the TBR1 T–Box binds Pal DNA.

On each trace, the donor and acceptor emission intensities are displayed. The pink traces are the donor emissions while the blue traces are the acceptor emissions. The FRET efficiency determined from the donor and acceptor emission intensities is displayed below each intensity trace. Each trace highlights the various FRET states in which the interaction occurs. **A**: A portion of a trace showing the donor (pink) and the acceptor (blue) as well as the high and low FRET states. **B**: A transition from low FRET efficiency to high FRET efficiency occurs without a dissociation event. **C:** A transition from high FRET efficiency to low FRET efficiency occurs without a dissociation event. **D**: Various binding events that occur at either a high FRET or a low FRET efficiency. Data was plotted using SigmaPlot version 12.

The frequency of binding events at each FRET efficiency is displayed for each TBR1 concentration (**Figure 67**). At each concentration there is a peak with a FRET efficiency between 0.02 and 0.06, this is a background level of FRET that occurs when there is no interaction. The insert in each panel in **Figure 67** displays the peaks that occur at a much lower frequency and with greater FRET efficiency. The two FRET efficiencies at which T–

Box binds Pal DNA are clear. Each peak (blue–low FRET and pink–high FRET) on the histogram was fitted to a Gaussian distribution and the area under the curves was extrapolated. The ratio of the area under the curve of the low FRET peak to the high FRET peak in relation to TBR1 concentration was plotted. The gradient of the linear regression is 0.0359 which suggests that the although the ratio does not change much with an increase in the TBR1 concentration, the low FRET state is slightly more stable than the high FRET state.

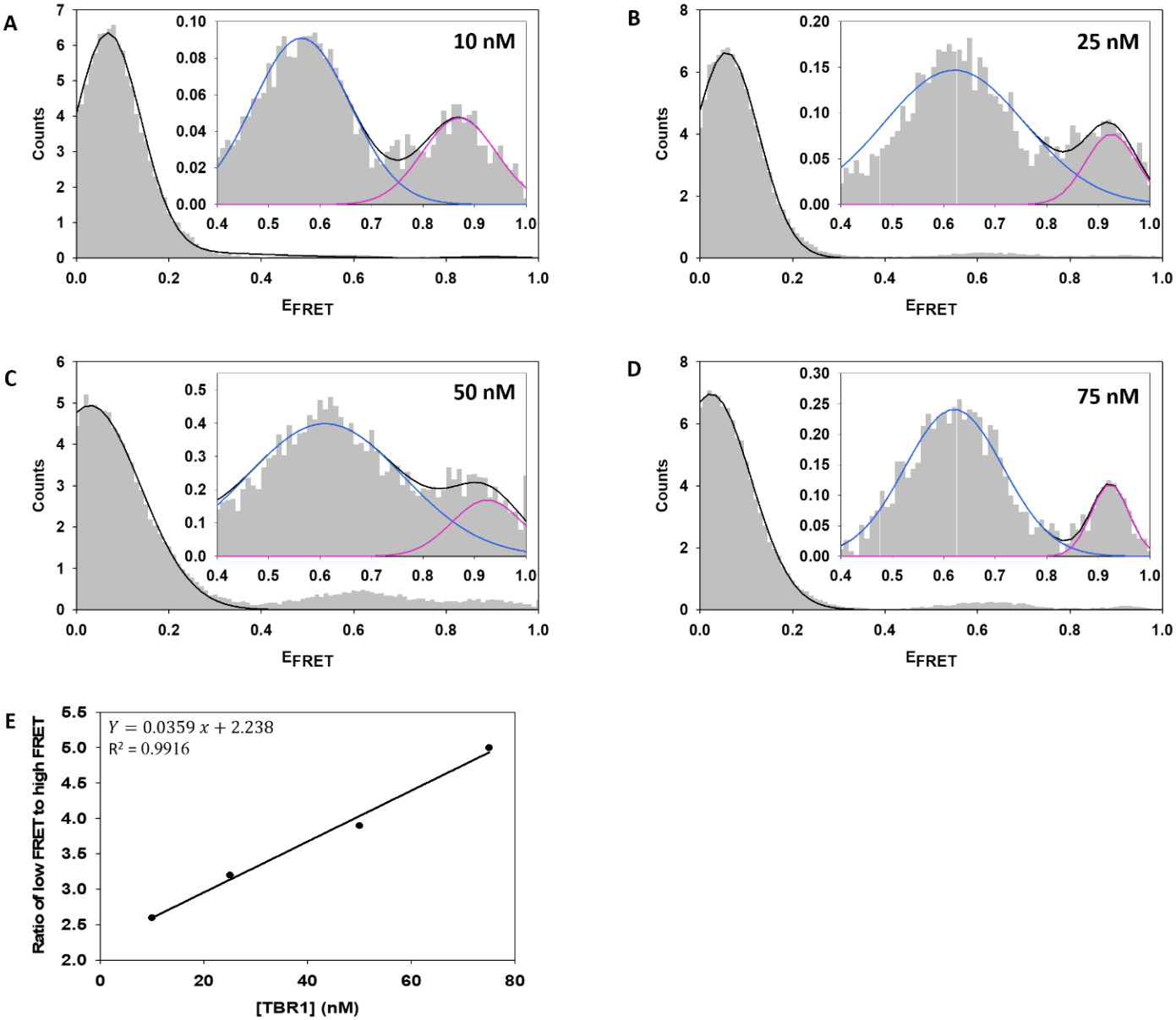


Figure 67: FRET efficiency histograms of the TBR1 T–Box binding to Pal DNA at increasing concentrations.

Histograms of the FRET efficiency counts recorded within the first 20 ns of all the traces. Each peak observed was fitted to a normal distribution and the area under each curve was calculated. The curve in each panel with a FRET efficiency of ~ 0.1 occurs when no FRET is observed (black). The inserted box in each pane depicts the counts of the higher FRET intensities not visible using the original scale. Panels **A**, **B**, **C** and **D** show the interactions that occur at 10, 25, 50 and 75 nM respectively. **E:** The ratio of the area under the low FRET peak (blue) to the high FRET peak (pink) as a function of TBR1 concentration. The low FRET state is slightly more stable than the high FRET state. Data was plotted using SigmaPlot version 12.

Kinetic analysis was done firstly by using scatter plots of the dwell time of Pal DNA in the T–Box bound and unbound states (**Figure 68**). It is clear that the off rates are quite

similar at the different concentrations of T–Box whereas the on rate is concentration dependent. Additionally, at the lower concentrations of T–Box the points are well spread whereas at the higher concentrations the points are more clustered showing that the association and dissociation rates of the interactions are more heterogenous at lower concentrations.

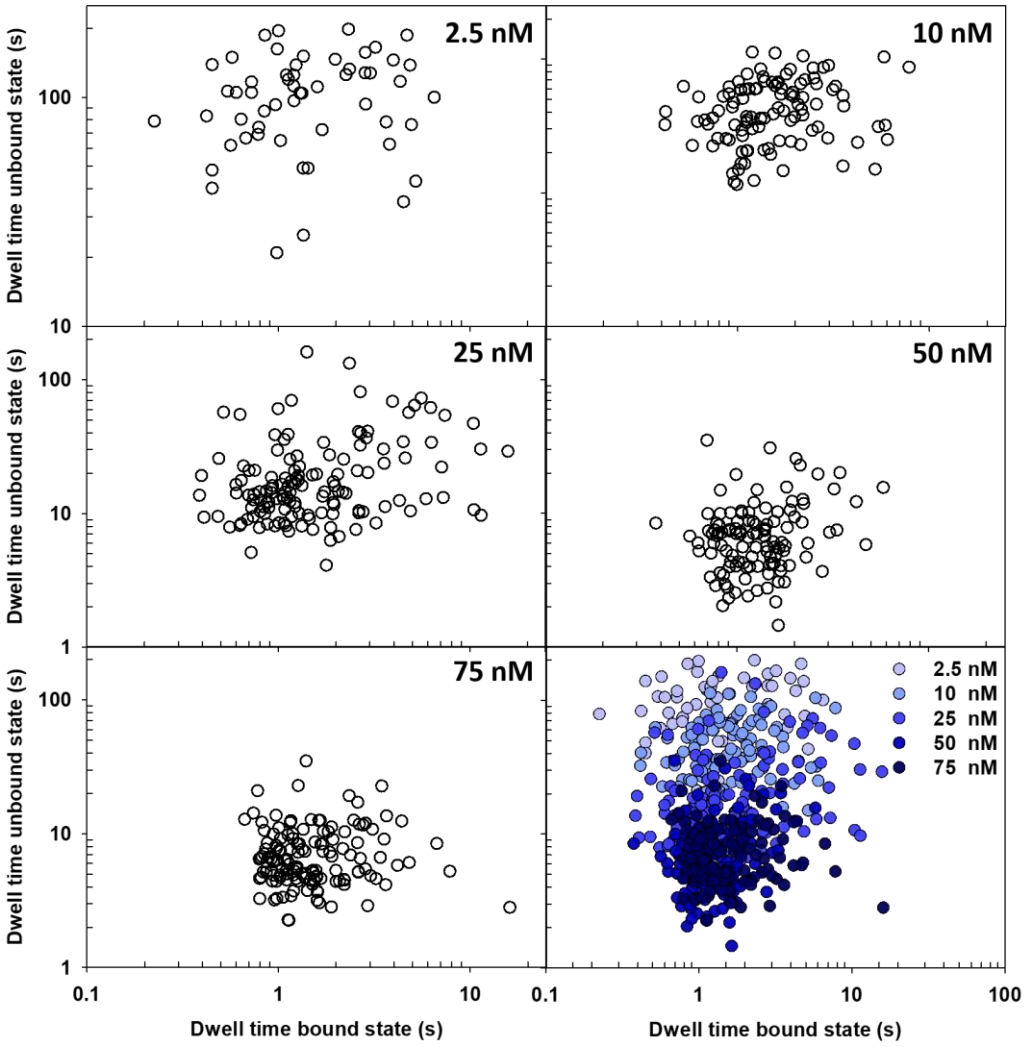


Figure 68: Scatter plots of the dwell time of Pal DNA in the T–Box bound and unbound states.

The dwell time in the bound state was plotted against the dwell time in the unbound state for each interaction. Each dot is representative of an interaction. The on rate is the inverse of the time spent in the unbound state whereas the off rate is an inverse of the time spent in the bound state. Each panel shows a scatter plot for a different T–Box concentration and the last pane is an overlay of them all. The on rate is dependent on T–Box concentrations whereas, the off rate is not. The association and dissociation rates of the interactions are more heterogenous at lower concentrations. Data was plotted using SigmaPlot version 12.

Rate histograms of the T–Box–Pal DNA interaction are displayed in **Figure 69**. These plots were used in addition to the scatter plots to obtain quantitative on and off rates rather than observe heterogeneity. The counts are the normalised number of interactions that remained bound (k_{OFF}) or unbound (k_{ON}) for that particular time frame. The histogram was fitted to an exponential decay function and the k_{ON} and k_{OFF} were derived for each concentration of TBR1 T–Box. Once again, we can see that the k_{ON} values increase significantly with increasing T–Box concentrations while the k_{OFF} values are not concentration dependent. This suggests that the higher the concentration of T–Box, the more likely an interaction is to form but once the stable interaction is formed, its dissociation is not concentration dependent. The observed on and off rates (k_{OBS}) for each concentration were plotted and a linear regression was performed in order to obtain the k_{ON} and K_{OFF} values (**Figure 70**). The on rate and off rate constants for the TBR1 T–Box–Pal DNA interaction were found to be $1.75 \times 10^{-3} nM^{-1} s^{-1}$ and $0.59\ s^{-1}$, respectively. The K_d was determined as 337 nM which is very similar to the K_d obtained from ITC.

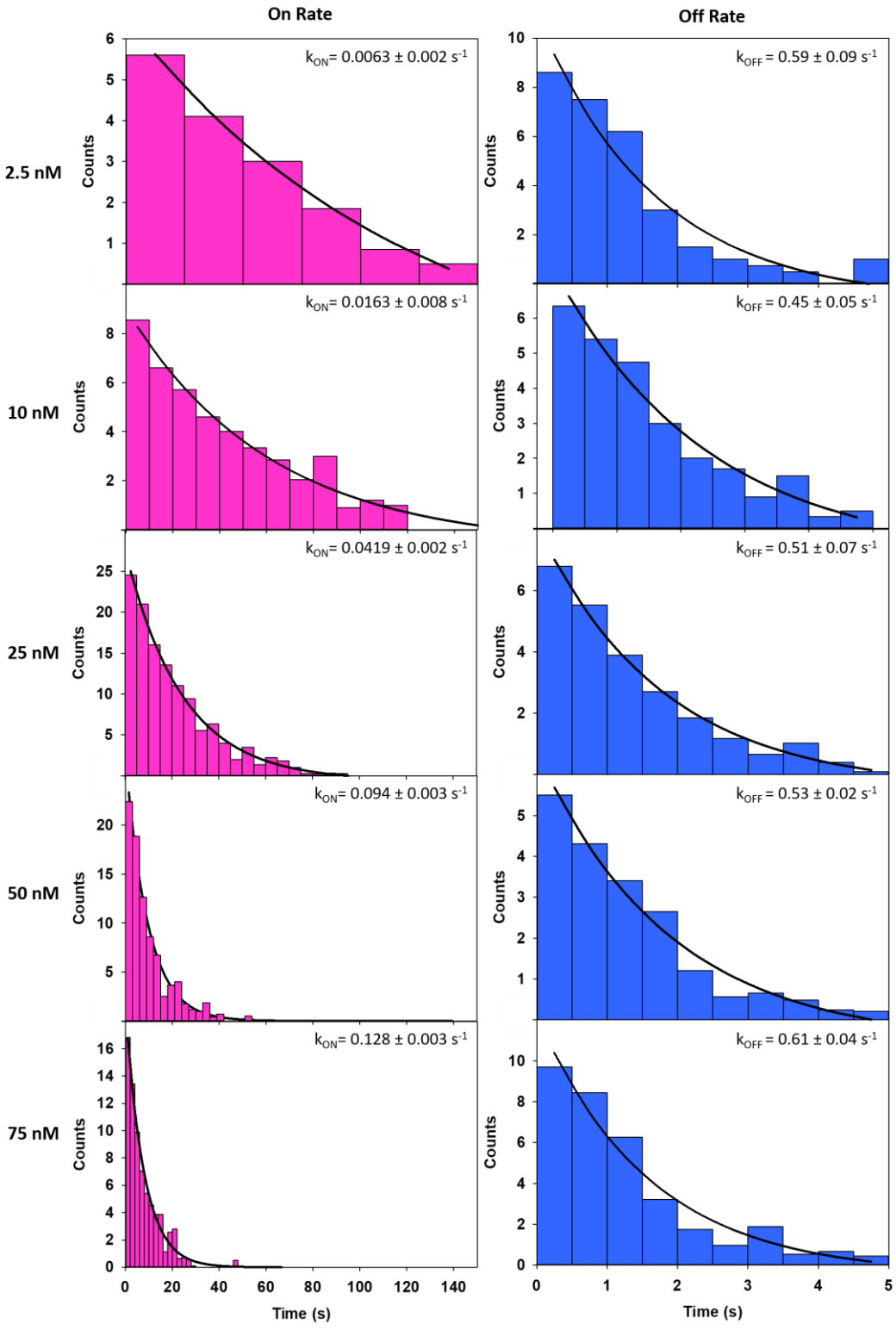


Figure 69: Rate histograms of TBR1 T–Box binding to Pal DNA.

The normalised counts of the time each interaction spends in the unbound state (on rate) and bound (off rate) for each T–Box concentration. Each histogram was fitted to an exponential decay function to obtain a rate constant at each concentration. The on rate plots (pink) are on the left while the off rate plots (blue) are on the right. The observed on rates are T–Box concentration dependent whereas the off rates are not. Data was plotted using SigmaPlot version 12.

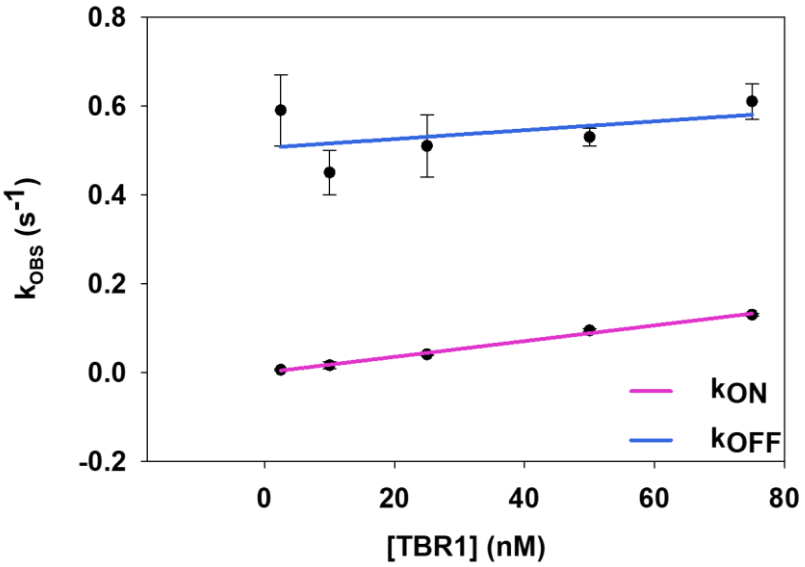


Figure 70: The k_{ON}, k_{OFF} and K_d of the TBR1 T–Box –Pal DNA interaction determined using smFRET

A: The observed K_{ON} (pink) and K_{OFF} (blue) determined for each T–Box concentration was plotted, and a linear regression was fitted. The k_{ON} was determined as the gradient of the linear regression. The k_{OFF} was determined as the average of the observed off rates at each T–Box concentration. The K_{ON} and K_{OFF} for TBR1 T–Box binding to Pal DNA were found to be 1.75x10 $^{-3}$nM^{-1} s^{-1} and 0.54 s^{-1}, respectively. Data was plotted using SigmaPlot version 12.

4.7.3.1. Single Site Long DNA

smFRET studies with SSL DNA were done in the same manner as with Pal DNA, however, there were not a significant number of interactions at concentrations of TBR1 T–Box below 10 nM. A representative traces and FRET efficiency histograms of 25 and 50 nM Alexa 647 labelled T–Box interacting with immobilised Cy3 labelled Pal DNA is displayed in **Figure 71**. The histograms show that T–Box binds SSL DNA at a single FRET efficiency of ~0.6, which is the same efficiency as the low FRET state in which T–Box binds to Pal DNA.

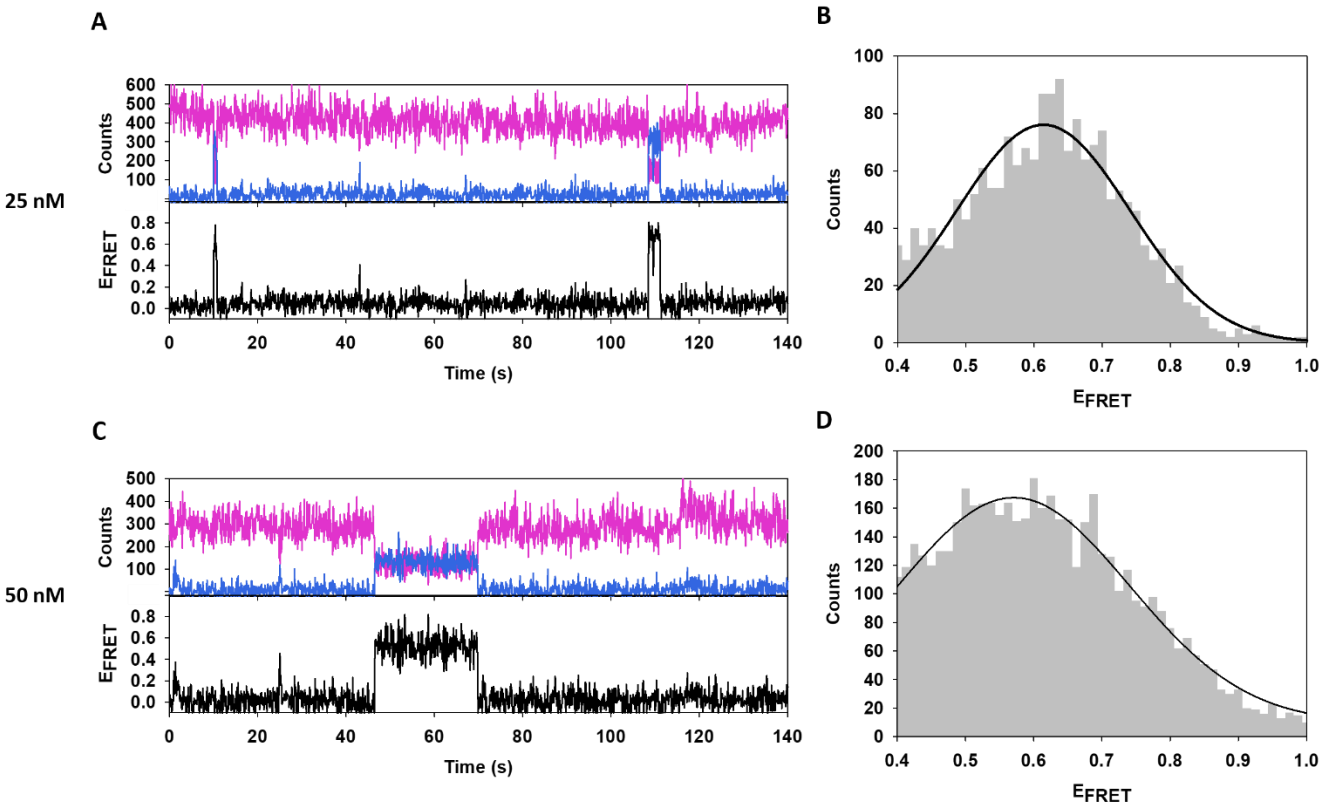


Figure 71: Representative traces and FRET efficiency histograms of immobilised SSL DNA interacting with 25 and 50 nM T–Box.

A representative trace of the donor and acceptor emission intensities SSL DNA interacting with 25 nM (**A**) and 50 nM (**C**) T–Box. The pink traces are the donor emissions while the blue traces are the acceptors emissions. The FRET efficiency determined from the donor and acceptor emission intensities are displayed below each intensity trace (black). **B** and **D**: Histograms of the FRET efficiency counts recorded within the first 20 ns of all the traces. The TBR1 T–Box interacts with SSL with only a single FRET state (~0.6) Data was plotted using SigmaPlot version 12.

4.7.4. Super shift–Electrophoretic Mobility Shift Assay

A super shift–EMSA was performed in order to observe if T–Box is able to interact with T–Box DNA (SSL and Pal) and the FHD at the same time. An EMSA was run with 0.5 µM Pal DNA in each lane and thereafter in each lane either the T–Box or FHD concentration was increased (**Table 12**). From the gels (**Figure 72**) it can be seen that when T–Box is in the presence of DNA but in the absence of FHD, there is only one complex formed. When FHD is added, there is a clear super–shifted band that appears. This band represents a complex containing DNA, FHD and T–Box. From this it can be deduced that T–Box is capable of interacting with both DNA and FHD simultaneously.

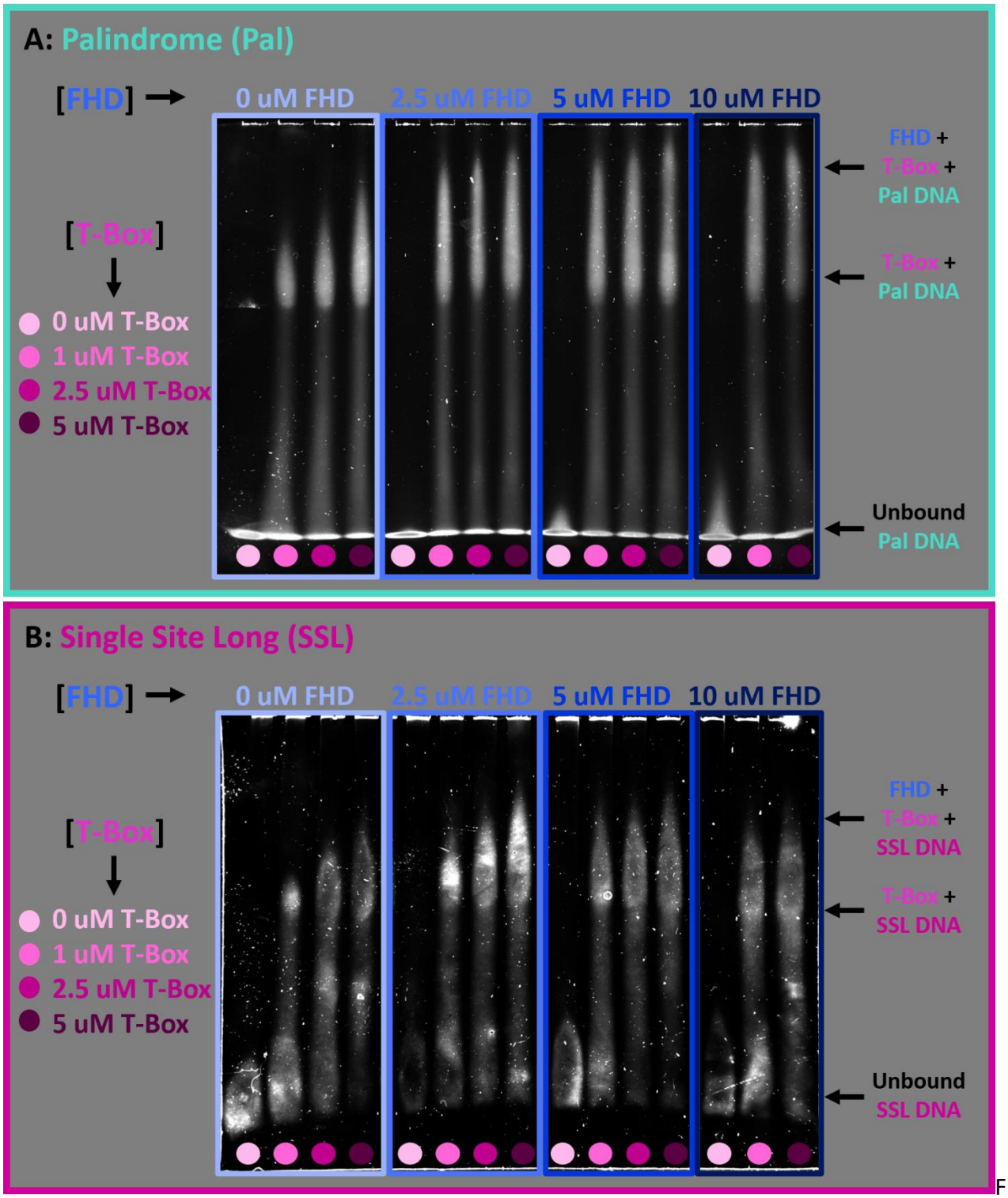


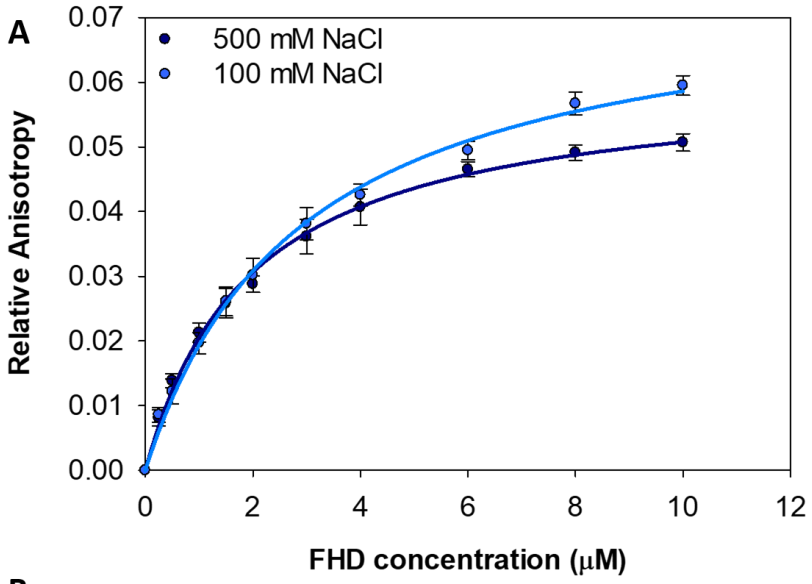
Figure 72: Super shift–electrophoretic mobility shift assay with Pal and SSL DNA, WT T–Box and WT FHD.

A super–shift EMSA performed on a 10% PAGE and run at 160 V for 10 hours. In each lane there was 0.5 µM DNA (**A:** Pal, **B:** SSL) and varying concentrations of T–Box and FHD. In the lanes where there was no FHD only one complex was present in the gel, however, as the FHD concentration was increased there was the formation of a super–shifted band consisting of DNA, T–Box and FHD.

4.7.5. Fluorescence Anisotropy

Fluorescence anisotropy was done to obtain a binding affinity for the interaction between TBR1 T–Box and FOXP2 FHD. This was achieved by labelling T–Box with Atto 550 via its His–tag and titrating the sample with small volumes of a high concentration of unlabelled cleaved FOXP2 FHD (no His–tag). Data was collected at 100 and 500 mM NaCl and the curves were fitted using SigmaPlot 12 to a single site ligand binding model (**Equation 10**) (**Figure 73**). The dissociation constants obtained from the interaction at 500 and 100 mM NaCl were 1.96 and 2.91 µM respectively, which means the interaction is not salt dependent.

A



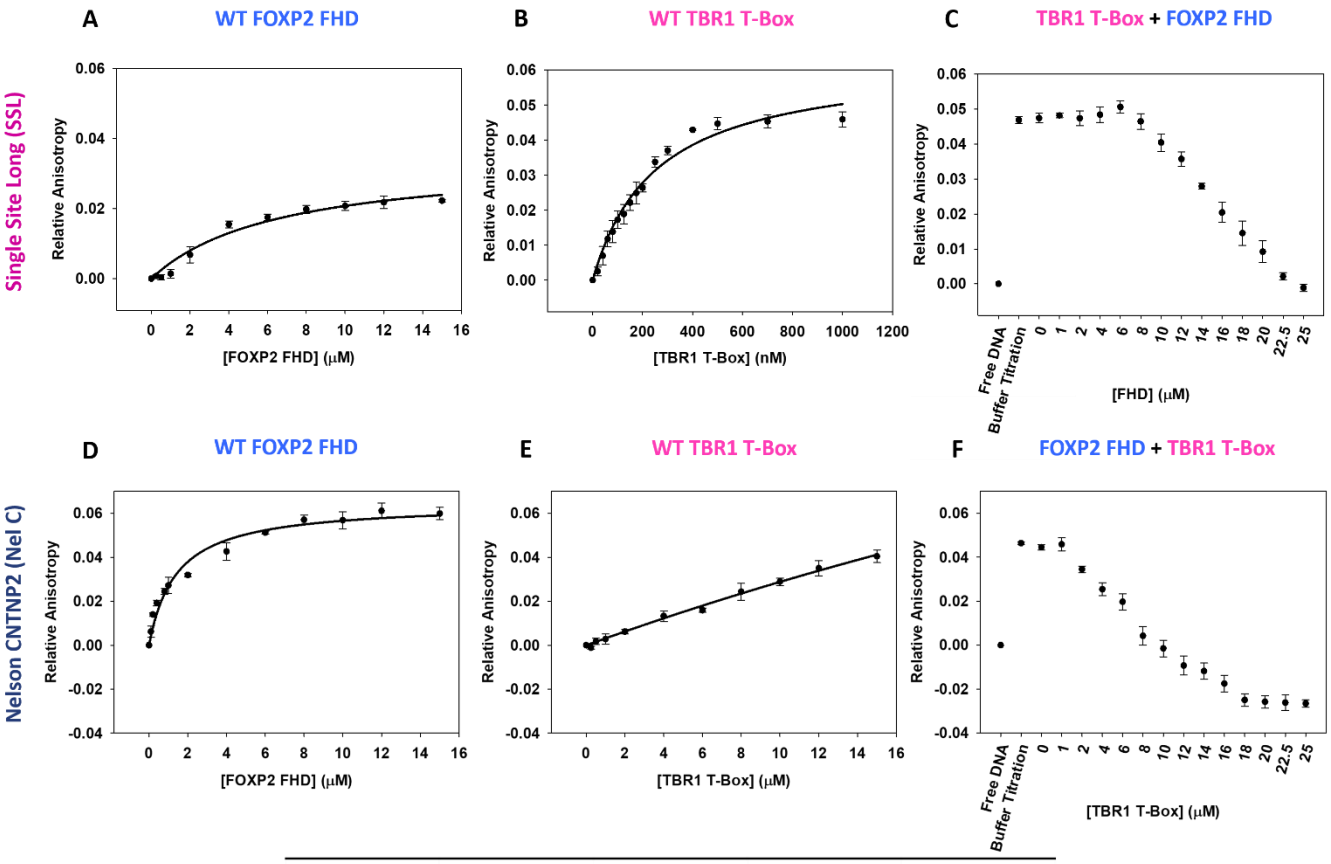
B

[NaCl]	K_d (µM)	B_{max}	R^2
100 mM	2.9 ±0.2	0.076 ± 0.002	0.9902
500 mM	1.9 ± 0.1	0.061 ± 0.001	0.9899

Figure 73: Binding isotherms of the TBR1 T–Box–FOXP2 FHD interaction at 100 and 500 mM NaCl.

A: Isotherms obtained using fluorescence anisotropy of the TBR1 T–Box–FOXP2 FHD interaction at 100 and 500 mM NaCl. **B**: Parameters obtained from fitting the points of the isotherm to a one site ligand binding curve. The data was obtained by titrating a high concentration of FHD into 0.5 µM NTA–Atto labelled T–Box sample. The measurements were acquired on a Perkin Elmer LS50B fluorometer with a buffer of 10 mM sodium phosphate, 3 mM TCEP and either 100 or 500 mM NaCl, pH 7.5. The dissociation constants obtained from the interaction at 500 and 100 mM NaCl were 1.9 and 2.9 µM, respectively. Data was plotted and fitted using SigmaPlot version 12.

Fluorescence anisotropy was used to monitor the binding of TBR1 T–Box and FHD to DNA containing a T–Box binging site (SSL) and as well as to DNA containing a FHD binding site (Nel C). Once the interactions between each protein and each DNA sequence were established, the binding of each protein to its DNA was titrated with the other protein to monitor whether a ternary complex forms or whether the protein dissociated from its DNA when in the presence of the second protein. Firstly, the binding of each protein to both SSL DNA and Nel C DNA was measured (**Figure 74**). The data was fitted to a single site ligand binding model (**Equation 10**) using Sigmaplot 12. Although the fit was not ideal, this single site model was chosen as using models with multiple binding sites resulted in poorer fits and unrealistic dissociation constants. The FOXP2 FHD bound Nel C DNA, its own cognate sequence, with a dissociation constant of 1.4 ± 0.2 μM and the FHD bound the SSL DNA (which contains the T–Box sequence) with a dissociation constant of 6.59 ± 1.3 μM. The low affinity binding to the SSL DNA is likely due to a non–specific electrostatic interaction. T–Box on the other hand had a much greater affinity for SSL DNA (K_d of 0.25 ± 0.021 μM) than Nel C DNA (K_d of 91.02 ± 55.2 μM). There is a discrepancy between the binding affinities obtained from WT T-Box binding SSL using ITC and FA which is expected because ITC is a much more sensitive technique than FA (Morris & Fanucchi 2016). In **panel C of Figure 74** the effect of increasing concentration of FHD on the T–Box–SSL DNA complex is seen. It appears that T–Box binds its SSL DNA and eventually dissociates when the FHD concentration exceeds 8 μM. A similar trend is observed in **panel F of Figure 74**, where FHD appears to dissociate from the Nel C DNA upon the addition of T–Box exceeding a concentration of 2 μM.



G

DNA	Protein	K_d (μM)	B_{max}	R^2
SSL	FHD	7 ± 1	0.035 ± 0.003	0.9595
SSL	T-Box	0.25 ± 0.02	0.063 ± 0.002	0.9665
Nel C	FHD	1.4 ± 0.2	0.065 ± 0.002	0.9645
Nel C	T-Box	91 ± 55	0.3 ± 0.2	0.9789

Figure 74: Fluorescence anisotropy binding studies of TBR1 T–Box, FOXP2 FHD, SSL DNA and Nel C DNA.

A: Binding curve FOXP2 FHD binding to SSL DNA with a K_d of 7 ± 1 μM. **B**: Binding curve TBR1 T–Box binding to SSL DNA with a K_d of 0.25 ± 0.02 μM. **C**: An anisotropy curve obtained by adding increasing concentrations of FHD to the T–BOX–SSL DNA complex. The T–Box control contained SSL DNA alone. The buffer control was obtained by adding the same volume of buffer as FHD was added to the last point. **D**: Binding curve FOXP2 FHD binding to Nel C DNA with a K_d of 1.4 ± 0.2 μM. **E**: Binding curve TBR1 T–Box binding to Nel C DNA with a K_d of 91 ± 55 μM. **F**: An anisotropy curve obtained by adding increasing concentrations of T–Box to the FHD–Nel C DNA complex. The FHD control contained Nel C DNA alone. The buffer control was obtained by adding the same volume of buffer as T–Box was added to the last point. **G**: A table displaying the dissociation constants obtained for the interactions between the two proteins and the two DNA sequences. All curves were plotted and fitted using Sigmaplot 12.

4.7.6. Molecular Docking

The HADDOCK webserver was used to dock the TBR1 T–Box domain model generated by I–TASSER to models of Pal, SSL and Nel DNA to try and see how the DNA binding might differ with the different DNA sequences. Furthermore, these models can be compared to other T–Box crystal structures as well ITC and smFRET data to try and ascertain a binding mechanism. The number of hydrogen bonds, buried surface area and burying of hydrophobic residues in the minor groove were analysed to try and ascertain any potential differences in binding due to DNA sequence (**Figure 75**). In **Figure 76**, the interactions (hydrogen bonds and hydrophobic interactions) between the T–Box model and Pal, SSL and Nel DNA are displayed and the number of hydrogen bonds formed in each model are listed in **Table 15**.

The SSL docked model had 13 hydrogen bonds as opposed to the 9 hydrogen bonds present in the Pal and Nel docked models. It is important to remember that it is not just the number of hydrogen bonds that form but also their geometry that is important when considering their role in affinity and specificity of binding. Additionally, the degree of burial of the hydrophobic residues (Phe387, Phe391 and Tyr395) in the minor groove is displayed in **Figure 75C**. On comparison, it is apparent that for Pal and SSL, the percentage of burying is quite similar whereas for Nel DNA the residues are much less buried. Proteins which bind in the major groove of DNA are typically stabilised by extensive hydrophobic interactions in the minor groove of the DNA (Privalov *et al.* 2007). This suggests that the T–Box binding site might facilitate the formation of these hydrophobic interactions and stabilise the complex. In **Figure 75A**, the buried surface area of the protein and DNA are displayed as well as the interface area. From this it is apparent that the Pal and SSL have greater buried surface areas than that of the Nel model. It is, therefore, likely that T–Box binds the SSL and Pal DNA with a greater affinity and stability than Nel DNA.

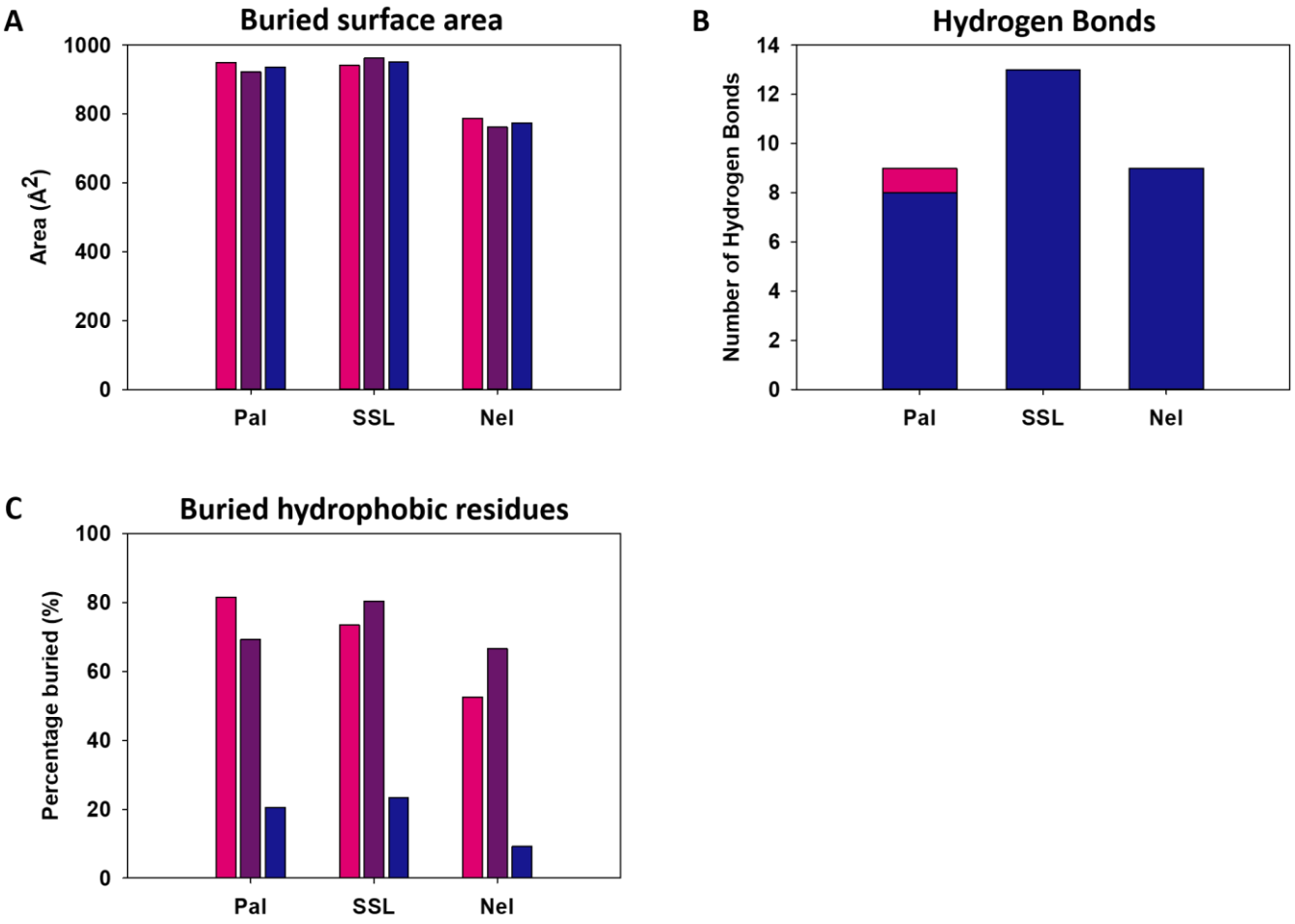


Figure 75: The buried surface areas, number of hydrogen bonds formed, and the percentage burying of hydrophobic residues observed in the docked models of T–Box and Pal, SSL and Nel DNA.

A: The buried protein surface area (pink), the buried DNA surface area (purple) and the area of the interface (blue) of the T–Box docked to Pal, SSL and Nel DNA. The areas were calculated using the PDIviz Pymol plugin (Ribeiro *et al.* 2015). **B:** The number of hydrogen bonds formed between the protein residues and the DNA backbone (blue) and DNA bases (pink). These were identified using Pymol 2.0.7. **C:** The percentage of burying of the 3 large hydrophobic residues (Phe387–pink, Phe391– purple and Tyr395– blue) in the minor groove of the DNA as obtained from the PDIviz Pymol plugin (Ribeiro *et al.* 2015).

Table 15: The hydrogen bonds formed between protein residues and nucleotides in the docked models of TBR1 T–Box and Pal, SSL and Nel DNA.

	Residue	Nucleotide	Total
Pal	**Lys228**	Thy5	9
	Arg231	Gua41	
	Arg232	Gua41	
	Arg265	Thy3***	
	Lys310	Thy40	
	Lys381	Thy42	
	Asn394	Cyt10	
SSL	**Lys228**	Ade30	13
	Gly230	Gua9	
	Arg231	Gua9**	
	Arg232	Gua9	
	Lys310	Gua9	
	Asn313	Cyt29**	
	Thr370	Thy28**	
	Asn394	Thy35**	
	Tyr395	Thy12	
Nel	**Lys228**	Ade11, Ade12	9
	Arg231	Thy22**	
	Arg232	Thy22	
	Arg265	Gua9**	
	Asn394	Thy18**	

The residue and nucleotide where the hydrogen bonds from for each docked model are displayed. *represents a case where more than one hydrogen bond forms between a residue and a nucleotide. The shaded residue indicates that the bond formed was with the nucleotide base rather than the backbone. The amino acid residues in bold are the residues that form interactions in all three models, while the residues in italics form interactions in two of the models.

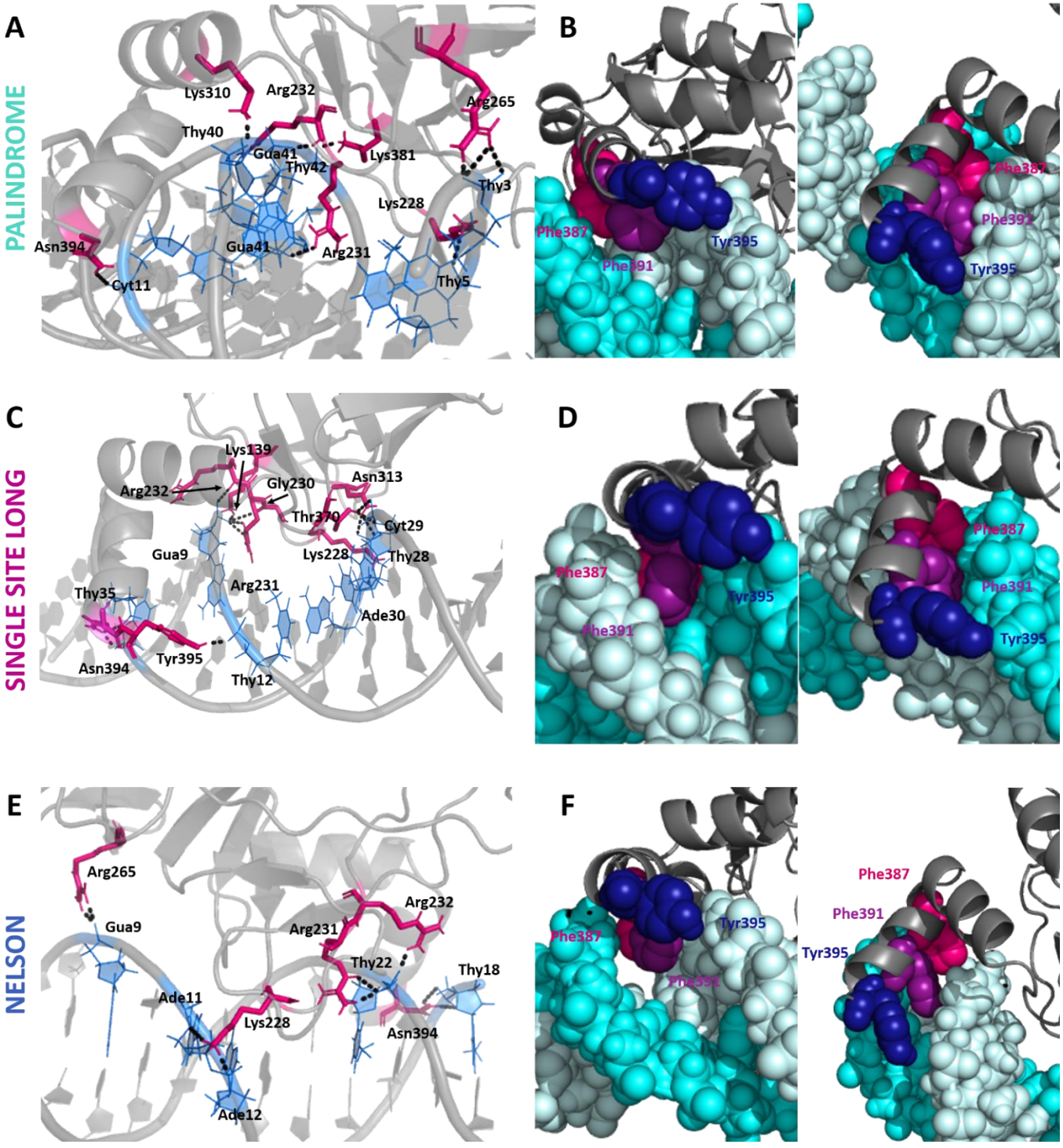


Figure 76: The interactions observed between the T–Box model and various DNA sequences obtained from docking simulations.

The hydrogen bonds formed between the TBR1 T–Box model and Pal (**A**), SSL (**C**) and Nel (**E**) DNA as observed from docking simulations. The protein residues forming part of a hydrogen bond are coloured in pink, whereas the nucleotides are coloured in blue. The hydrogen bonds are represented by the black dashed lines. The surfaces of the large hydrophobic residues (Phe 387: pink, Phe 391: purple and Tyr 395: blue) are displayed in the minor groove of the Pal (**B**), SSL (**D**) and Nel (**F**) DNA. Image rendered with PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.

DISCUSSION

FOXP2 and TBR1 are two transcription factors that are co–expressed in certain areas of the brain. Mutations in both proteins have been associated with speech defects, ASD and intellectual disability (Deriziotis & Fisher 2017). Deriziotis *et al*, (2014) have shown using an *in vivo* BRET assay that not only do TBR1 and FOXP2 interact, but also that their interaction occurs via the T–Box domain of TBR1 and either the polyglutamine rich region or the C terminal end of FOXP2. The aim of this study was, therefore, to determine if the FOXP2 FHD forms a direct interaction with the TBR1 T–Box domain *in vitro*, and what the resulting effects of this interaction are on each protein's DNA binding function. A complete picture of the interactions that FOXP2 and TBR1 form with each other and with DNA is required to understand how conditions such as ASD and ID develop. Only once the network of interactions involved in the relevant developmental pathways is understood, can potential treatments to such conditions be designed. To achieve this, firstly novel structural characterisations and DNA binding studies were performed on the TBR1 T–Box domain. This was important to establish because binding activity relies on structure, and a good understanding of its DNA binding mechanism was necessary in order to properly understand how the protein–protein interaction might influence it. Secondly, the interaction between the FOXP2 FHD and TBR1 T–Box domain, followed by the subsequent effect on DNA binding was characterised. This presented a far more comprehensive understanding of how each transcription factor relies on its interaction with other proteins to either repress or activate transcription due to effects on DNA binding. Furthermore, the influence of mutations could be identified in terms of structural changes and which mechanism important for function they influence in neurological conditions.

5.1. FOXP2 FHD DNA BINDING

FOXP2 FHD DNA binding has previously been extensively characterised using a variety of *in vitro* techniques, such as electrophoretic mobility shift assay, fluorescence anisotropy, crystallography, surface plasmon resonance and isothermal titration calorimetry (Blane *et al.* 2018; Blane & Fanucchi 2015; Morris *et al.* 2018; Morris & Fanucchi 2016; Stroud *et al.* 2006; Webb *et al.* 2017). For this reason, minimal DNA binding studies (EMSA and FA) were conducted to confirm that the FHD was active; that it binds its own cognate DNA sequence specifically; and finally, to study how and if its DNA binding interaction is affected when in the presence of TBR1. The focus of this project with respect to the FHD was, therefore, to determine if it could interact with the T–box domain in isolation and to observe if this interaction would affect the ability of the FHD to interact with its cognate DNA sequence. The aim was to ascertain whether the interaction between these two proteins affects FOXP2s' DNA binding function. If so, this would suggest that the TBR1–FOXP2 interaction is likely to alter how FOXP2 regulates the expression of certain genes and is, therefore, relevant to the developmental processes they are mutually involved in.

Electrophoretic mobility shift assays qualitatively showed that the FHD was active and that it interacted only with DNA sequences containing the Nelson binding site (**Figure 61**). In addition to the EMSAs, fluorescence anisotropy was conducted to obtain dissociation constants for FHD binding to SSL and Nel C DNA, which were 6.6 and 1.4 µM, respectively (**Figure 74**). Previous fluorescence anisotropy studies showed that the FOXP2 FHD binds Nel DNA with a K_d of 0.89 µM (Morris & Fanucchi 2016), which is not significantly different to the dissociation constant obtained in this study for Nel C. According to EMSA (**Figure 61**), the FOXP2 FHD does not bind SSL DNA, however, based on fluorescence anisotropy the K_d was 6.6 µM (**Figure 74**). This is a greater affinity than expected but it is likely a result of a non–specific electrostatic interaction between the negatively charged DNA and the positively charged FHD; as FHD has a predicted p*K*a of 10.2 (Gasteiger *et al.* 2005). In fact, Webb *et al.* (2017) showed using SPR that a sequence that FHD bound to with a 5 times weaker affinity than it bound Nelson DNA (as is seen with FHD binding to SSL DNA), bound with a slower on rate and a much faster off rate. The authors attributed this to non–specific binding which allows the protein to scan DNA in a partially unfolded state in order to find high affinity sites as described by Todeschini *et al.* (2014).

5.2. TBR1 T–Box DNA Binding

5.2.1. Thermodynamics

Unlike the FOXP2 FHD, little is known about TBR1 T–Box domain DNA binding and thus the binding mechanism is unknown. There are, however, various T–Box domains that have been crystalized in their DNA bound forms (XBra: 1XBR, TBXT: 6F58, TBX1: 4A04, TBX3: 1H6F and TBX21: 5T1J). These are seen binding to DNA by inserting helix 4 into the minor groove while helix 3, perpendicular to helix 4, lies over the backbone (**Figure 17**) (Muller & Herrmann 1997). Based on docking studies, (**Figure 75** and **Figure 76**) TBR1 T–Box binds to DNA via the insertion of three conserved, hydrophobic residues (Phe387, Phe391 and Tyr395) located on helix 4, into the minor groove of the DNA without any DNA distortion or bending as is seen in other T–Box crystal structures (Muller & Herrmann 1997; Stirnimann *et al.* 2010). Most minor groove binding proteins exhibit a thermodynamic profile that is entropically driven due to the expulsion of water in the minor groove, which is highly ordered, particularly in AT rich regions, (Privalov *et al.* 2007) as well as distortion of the DNA (Jen–Jacobson *et al.* 2000). The TBR1 T–Box DNA interaction is, however, entropically unfavourable, which is likely caused by the stabilisation of protein and DNA upon binding.

Additionally, in the docked models, (**Figure 75** and **Figure 76**) helix 4 and some amino acids that form loops protruding from the surrounding β–sheets are seen forming several hydrogen bonds with the DNA backbone. The observed thermodynamic profile shows TBR1 T–Box DNA binding to be enthalpic, (**Figure 63**) which suggests it is driven by

hydrogen bond formation. Binding to the major groove is typically enthalpically favourable due to the formation of multiple hydrogen bonds (Privalov *et al.* 2007).

5.2.2. Affinity and Specificity

Transcription factors are required to bind DNA with varying degrees of affinity and specificity in order to allow them to not only bind their target sequence but also scan DNA to find the target sequence (Halford & Marko 2004). Using ITC and FA, dissociation constants were obtained for TBR1 T–Box binding to various DNA sequences under different ionic strength conditions using ITC (**Figure 63**). From this it is immediately obvious that at 250 mM NaCl, the dissociation constants for both SSL and Pal DNA are higher, suggesting that the interaction is at least in part electrostatic in nature. According to Privalov *et al.* (2011) the enthalpic term of a protein–DNA interaction is non–electrostatic whereas the entropic term is comprised of electrostatic and non–electrostatic components. The non–electrostatic component of binding gives rise to binding specificity whereas the electrostatic component largely contributes to the affinity of a typical protein–DNA interaction (Privalov *et al.* 2011). So, although the entropic term is unfavourable, there is an electrostatic component of it which is favourable but is counteracted by an unfavourable non–electrostatic component. This non–electrostatic component of the entropic term is likely the result of stabilisation of the complex upon binding. This demonstrates that TBR1 T–Box DNA binding is not only driven by direct or base readout (the formation of specific hydrogen bonds) but also by indirect or shape readout which is a conformational shift and stabilisation as binding occurs (Rohs *et al.* 2010).

When comparing the K_ds obtained for the different DNA sequences (SSL and Pal), T–Box binds SSL with a significantly greater affinity (**Figure 63**). This is likely a result of interactions with flanking regions. The SSL DNA has its binding site in the centre of the DNA with bases on either side that the protein might interact with. Often transcription factors have been found to interact with flanking regions on DNA. These interactions with flanking regions can stabilise the interaction and can incur DNA specificity (Chen *et al.* 2015; Gordân *et al.* 2013; Leonard *et al.* 1997; Morin *et al.* 2006; Nagaoka *et al.* 2001; Shen *et al.* 2018). The Palindrome DNA, however, has the two binding sites with no flanking bases. This means that T–Box cannot form interactions with flanking regions which may be important. It is also possible that if both sites are occupied there could be steric hindrance which would prevent the stable formation of a protein–DNA complex.

The importance of the flanking regions in DNA binding was also illustrated in EMSAs where T–Box binding to SSS (DNA containing a single site and no flanking bases) and SSL can be compared (**Figure 60**). Clear binding is observed in lane 2 for SSL DNA (**Figure 60A**) at a protein concentration of 0.25 µM, whereas clear binding is only observed for SSS DNA in lane 3 (**Figure 60B**) at a protein concentration of 1 µM. This again shows that flanking regions do contribute to the affinity of T–Box DNA binding.

The SSL DNA sequence contains the typical T–Box binding site with flanking regions found in the AUTS2 promoter region, a gene T–Box is known to regulate. There are many single T–Box binding sites found within the genome with varying flanking regions. This illustrates the importance of these base pairs outside of the binding site in terms of DNA binding affinity. Furthermore, thus far all T–Box domains have been shown to bind the T–Box binding site, so it is possible that the flanking regions might confer specificity to allow preferential binding amongst T–Box proteins.

Based on electrophoretic mobility shift assay (**Figure 60**), T–Box was found capable of not only binding DNA containing the classic T–Box binding site but it was also able to bind the Nelson sequence (but not the other DNA sequences containing the FOXP2 binding site). Docking studies were performed of the TBR1 T–Box model binding to SSL, Pal and Nel DNA (**Figure 75** and **Figure 76**) and in all three docked models T–Box binds the DNA by inserting Helix 4 into the minor groove. The three bulky hydrophobic residues (Phe378, Phe391 and Tyr395) are more buried in the models containing SSL and Pal DNA rather than the Nel docked model. The lesser degree of protein burial in the minor groove in the Nel model might be because the protein is not in the ideal orientation to optimise hydrogen bonding and Van der Waals contacts in other portions of the model. This indicates that while the hydrophobic interaction in the minor groove is a prominent aspect of T–Box DNA binding, it might be other interactions that help orientate T–Box and allow for insertion of Helix 4 into the minor groove.

Association and dissociation rates of T–Box binding to Pal DNA were obtained from smFRET. The rates, k_{ON} (1.75×10^6 M^{-1} s^{-1}) and k_{OFF} (0.59 s^{-1}) were used to calculate the K_d (330 nM). The K_d was similar to the K_d obtained for Pal DNA binding obtained using ITC (366 nM). The small discrepancy could be because firstly, these dissociation constants were obtained using very different techniques and secondly, the NaCl concentration used in smFRET was 50 mM NaCl while it was 100 mM NaCl for ITC (**Figure 77**). The K_ds obtained for Pal DNA binding decrease steadily as NaCl concentration is decreased which is as expected as transcription factor–DNA complexes normally have an electrostatic component that is the main factor in affinity (Privalov *et al.* 2011). The K_d obtained for T–Box binding to SSL DNA using fluorescence anisotropy is quite a lot higher than the value obtained from ITC. This, however, is likely because anisotropy is not as sensitive as ITC.

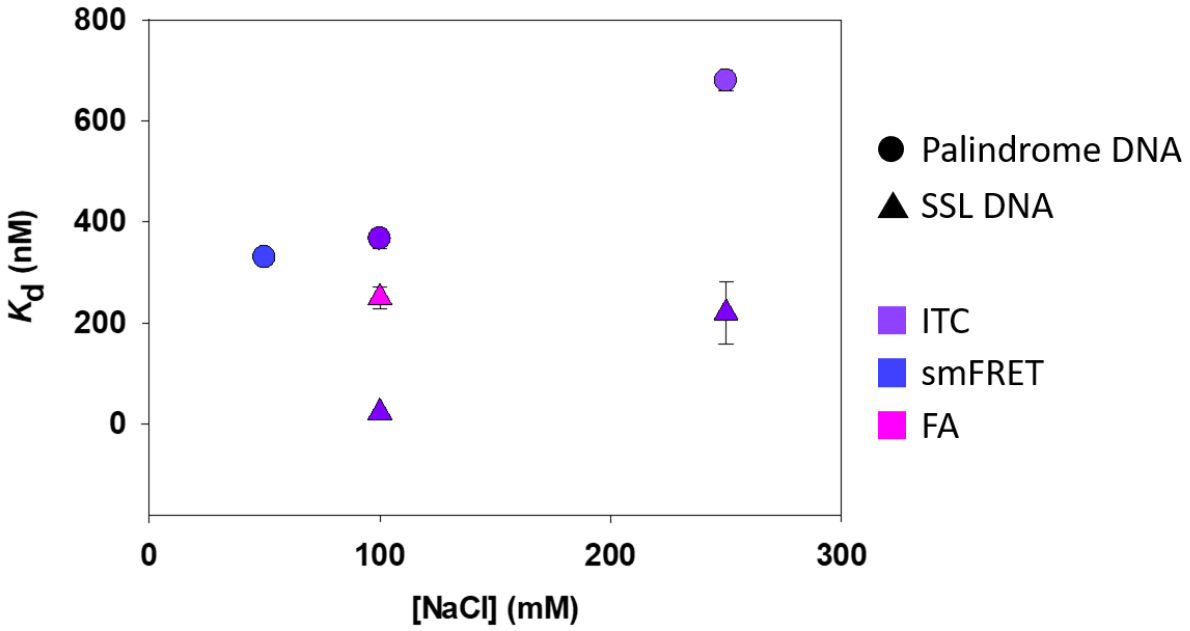


Figure 77: The different dissociation constants obtained for TBR1 T–Box interacting with SSL and Pal DNA under different conditions and using different techniques.

The dissociation constants obtained for T–Box binding Pal DNA (circles) and SSL DNA (triangles) using ITC (purple), smFRET (blue) and FA (pink) at various NaCl concentrations. The dissociation constants of T–Box binding both DNA sequences are NaCl concentration dependent. While ITC and smFRET give similar values for Pal DNA binding the dissociation constant for binding to SSL obtained from FA is much higher than that of ITC. This is because ITC is a much more sensitive technique than FA. Data was plotted using SigmaPlot version 12.

The only other *in vitro* study performed on the TBR1 T–Box domain interaction with DNA used biolayer interferometry to obtain the association and dissociation rates and well as a dissociation constant of T–Box binding to DNA that contained a single binding site and flanking regions from the GRIN2B promoter region (Yook *et al.* 2019). They obtained a k_{ON} of 1.4×10^4 M^{-1} s^{-1}, a k_{OFF} of 0.025 s^{-1} and a K_d of 900 nM. Both the on and off rates are slower than what was obtained from smFRET (T–Box binding to Pal DNA). This could be because different techniques and conditions were used, or it could be due to lack of flanking regions available on the palindrome DNA resulting in a faster off rate. The dissociation constant of 900 nM was significantly higher than that of SSL DNA binding at 100 mM NaCl as obtained from ITC (23 nM). This could again suggest that flanking regions may play an important role in T–Box DNA binding. Further studies using smFRET comparing the kinetics of T–Box binding to DNA sequences with different flanking regions could help identify which bases are important for TBR1 T–Box DNA binding. This could further help identify T–Box sites in the genome which TBR1 T–Box would actually bind, differentiating them from the non–specific binding sites which help in DNA scanning as described by Todeschini *et al.* (2014).

5.2.3. DIMERISATION

The first crystal structure of a T–Box domain (T protein from *Xenopus laevis,* PDB ID: 1XBR) was solved with a T–Box dimer bound to DNA containing two T–Box binding sites arranged in a palindrome (Muller & Herrmann 1997). This sequence was identified in vitro by nucleotide selection from a random pool of oligonucleotides and has not been found in the human genome thus far. To date there are structures of five T–Box domains bound to this palindrome DNA and all (with the exception of TBX21) bind the DNA as dimers but with very small dimer interfaces (**Figure 18**). This suggests that either the dimer interfaces are a result of crystal packing or they only dimerise on the palindrome DNA. TBX21, the only structure of a T–Box domain in the TBR1 family, forms a dimer with an extensive dimer interface and binds only a single site in the palindrome DNA. TBX21 is thus far also the only T–Box domain reported to form a mixture of monomer and dimer in solution (Liu *et al.* 2016). The dimer interface of TBX21 is formed by two extended loops found in TBR1 and TBR2 but not in other T–Box domains.

Analytical size exclusion chromatography (**Figure 53**) revealed that the TBR1 T–Box domain existed in a mixture of monomer and dimer which transitioned to a predominantly monomeric species upon the addition of a reducing agent. The TBR1 T–Box domain has two cysteine residues not conserved in other T–Box domains. Therefore, in order to investigate this further, each cysteine in the T–Box domain was mutated in turn and each mutant (C209A and C274S) was found to be monomeric with and without the addition of DTT. This suggests that the TBR1 T–Box domain can form disulfide linked dimers between Cys 209 in one monomer and Cys 274 in the other monomer. The disulfide linked dimer is unlikely to form the same dimer observed in other T–Box crystal structures due to the location of the Cys residues according to the I–TASSER model (**Figure 78**). Therefore, if a physiological disulfide linked dimer does form it is unlikely to bind palindrome DNA in the same orientation.

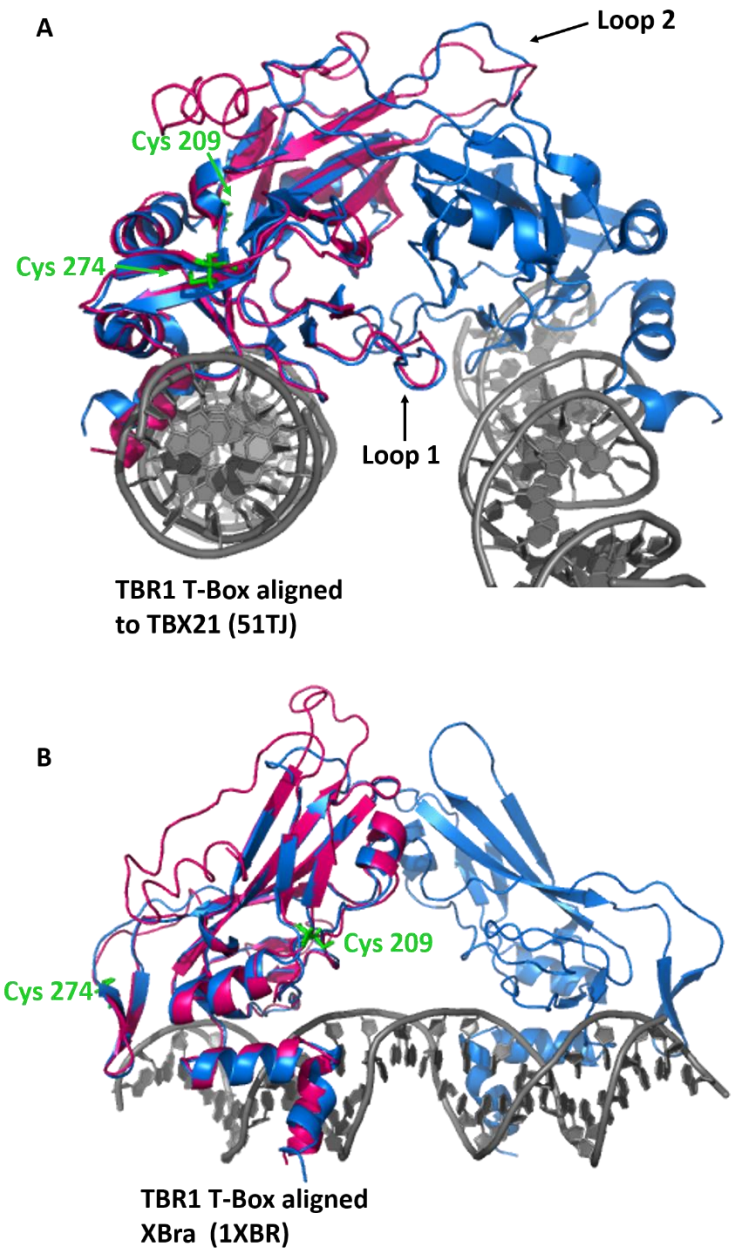


Figure 78: ITASSER model of the TBR1 T–Box aligned with the TBX21 and Xbra crystal structures.

A: ITASSER model of T–Box (pink) aligned to TBX21 (blue) bound to Pal DNA crystal structure (PDB ID: 5T1J)(Liu *et al.* 2016). **B**: ITASSER model of T–Box (blue) aligned to XBra (blue) bound to Pal DNA crystal structure (PDB ID: 1XBR) (Muller & Herrmann 1997). The Cys residues are coloured in green. Image rendered with PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.

ITC data (**Figure 63**) of TBR1 T–Box binding to Pal DNA had stoichiometries of 2.1 and 1.5 at 100 mM and 250 mM NaCl, respectively. This shows that both sites of the Pal DNA are occupied at lower NaCl concentrations when the K_d is lower (366.67 nM). However, at higher NaCl concentrations the stoichiometry decreases which shows that sometimes the second binding site is not occupied. The fact that size exclusion shows that the TBR1 T–

Box is monomeric in solution (**Figure 53**) and the occurrence of partial occupancy of the second binding site in ITC, suggests that T–Box can bind to both sites on the DNA independently, but it does not dimerise in solution and then bind both sites as a stable dimer. It is, however, possible that a small dimer interface does exist between the two monomers when bound to DNA as with the Xbra (PDB: 1XBR), TBXT (PDB: 6F58), TBX1 (PDB: 4A04) and TBX3 (PDB: 1H6F) crystal structures.

smFRET studies showed two FRET states (~0.9 and ~0.6) for T–Box binding to Pal DNA, and that there can be a transition from the one to the other without a dissociation event (**Figure 66** and **Figure 67**). To conclusively assign binding events to each FRET state, further smFRET studies need to be performed. It is, however, notable that although a complete smFRET data set of T–Box binding to SSL was not obtained, only one FRET state was seen. This means that the two FRET states observed when T–Box interacts with Pal DNA are likely a result of the occupancy of the two binding sites or a dimerisation event. There is a likely scenario to explain the low and high FRET states, (**Figure 79**) however, further experimentation needs to be done to confirm this. The low FRET state has one protein bound to the site furthest from the Cy3 (far site) and the higher FRET state could be a combination of both sites being occupied and only the site closest to the Cy3 being occupied. It is possible that a protein bound in the site closest to the DNA could dominate the signal even if a protein molecule is bound to the far site making it difficult to distinguish between the two binding events. This explanation is supported by the ITC data which indicated that both sites can be, but are not always are occupied.

Binding occurred in the low FRET state more frequently and transitions from high FRET to low FRET were more common than the reverse. If the low FRET state does, in fact, represent binding of a single protein to the far site, whereas the high FRET state represents a combination of binding to both sites and binding to only the close site, it would suggest that binding to the far site might be more favourable. This means that the close site could be partially inhibited by the Cy3. A single site occupancy could be more stable than if both sites are bound as there may be some degree of steric hindrance preventing T–Box interacting with bases flanking the binding sites. The more frequent low FRET state and transition from high FRET to low FRET would suggest that while both sites can be occupied, it is not a protein dimerisation event, as a dimerisation event would be more likely to stabilise the complex making the higher FRET state more prevalent.

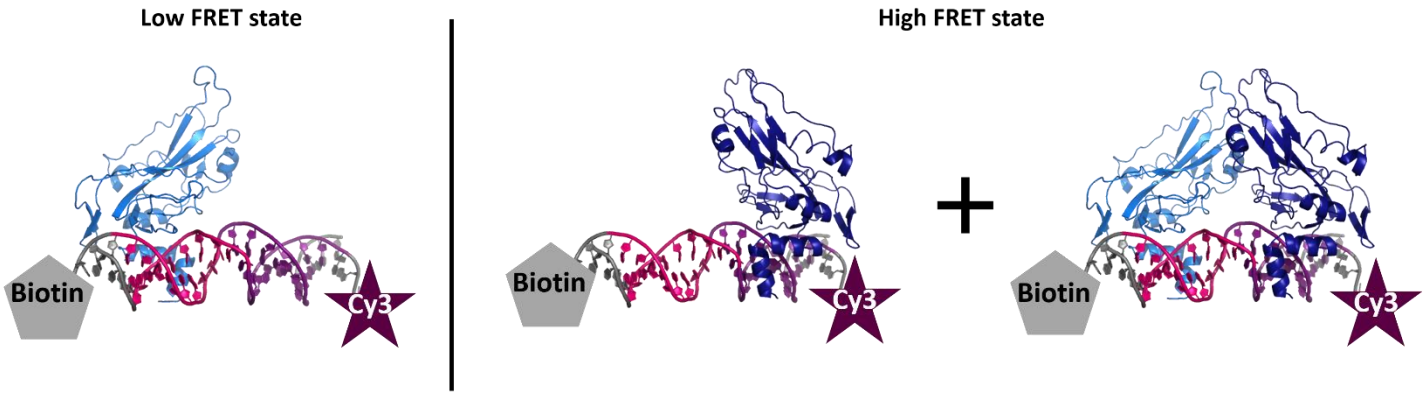


Figure 79: Possible binding events that might occur at the high and Low FRET states observed when TBR1 T–Box interacts with Pal DNA in smFRET.

One of possible scenarios that explain the high and low FRET binding states when T–Box binds to palindrome DNA is that the low FRET state could be one protein molecule binding to the site furthest from the Cy3. The high FRET could be accounted for by two different interactions. The first being a protein molecule binding to the site closest to the Cy3 and the second being both sites occupied by protein. It is possible that if both proteins are bound, the signal is dominated by the protein bound in closest site so it might not be possible to differentiate between a protein molecule bound to the site closest to the Cy3 and both sites being occupied. Image rendered with PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.

5.4. TBR1 T–Box–FOXP2 FHD Interaction

Using bioluminescence resonance energy transfer (BRET) assays, Deriziotis *et al* (2014) found that the TBR1 T–Box domain forms an interaction either with a region within the polyglutamine tract or with the FHD of FOXP2. A BRET assay is an *in vivo* technique where one protein is expressed as a chimera with luciferase. When luciferin (luciferase's substrate) is added, the luciferase catalyses a bioluminescent reaction, and an acceptor fluorophore (attached to the interacting partner) is excited. BRET can occur from a distance anywhere between 10 and 100 Å. This means that the two interacting molecules might not form a direct interaction but may instead interact indirectly through other macromolecules (physical association) (Hermjakob *et al.* 2004). Deriziotis *et al* (2014) conducted the BRET assays with various regions and mutations in both FOXP2 and TBR1 and established the regions of each protein that is likely form the interaction. In this study, fluorescence anisotropy was used to determine if the FOXP2 FHD and the TBR1 T–Box domain form a direct interaction *in vitro* (**Figure 73**). The dissociation constants of this interaction in 100 and 500 mM NaCl were found to be 2.9 and 2 µM, respectively. This confirms that a direct interaction does occur, and that the interaction is not electrostatic in nature as it is not significantly altered by increasing NaCl concentrations.

Protein–protein interactions can be classified as high affinity (K_d < 10^{-10} M), medium affinity (K_d = 10^{-6} to 10^{-10} M), or low affinity (K_d > 10^{-6} M) (Kastritis *et al.* 2011). Such a wide range of affinities exists in order to allow for either stable or transient complex

formation, which facilitates cellular functions ranging from nuclear import to formation of structural elements. For example, many interactions involved in membrane trafficking are low affinity, while some enzyme–inhibitor complexes have very high affinities (Erijman *et al.* 2014). The interaction between FOXP2 FHD and TBR1 T–Box domain is a moderate affinity interaction within in the same range as both proteins' affinity for DNA. This could allow the flexibility to form various multimolecular complexes required for transcription regulation. This is seen with FOXP3, where it acts as either an activator or a repressor depending on whether it forms a complex with RELA, IKZF2 and KAT5 or with RZH2 and YY1 (Kwon *et al.* 2017).

5.5. Ternary Complex Formation

Fluorescence anisotropy as well as super shift electrophoretic mobility shift assays were used to ascertain if a ternary complex consisting of FHD, T–Box and DNA forms or if the protein–protein interaction inhibits DNA binding. In **Figure 72**, super shift–EMSAs show T–Box binding to SSL and Pal DNA in the presence of FHD. In both gels, when FHD is added, the bands are higher up on the gel in lanes containing both proteins showing the formation of a ternary complex. Fluorescence anisotropy showed that FHD and T–Box bind SSL DNA with K_ds of 6.6 and 0.25 μM, respectively. To ascertain whether a ternary complex forms, the SSL DNA was saturated with T–Box and FHD was titrated in. Initially, there was a slight increase in anisotropy value which indicates that there was an increase in the average size of the complex. This shows that a ternary complex forms as seen on the super shift–EMSA. Once the FHD concentration exceeded 6 μM, though, the anisotropy value decreased as a result of T–Box dissociating from the DNA as it interacted with the FHD. The initial increase in anisotropy value suggests that T–Box can, therefore, interact with both its DNA and FHD at the same time but will dissociate from DNA when there is a large enough excess of FHD present. T–Box has an approximately ten times greater affinity for its DNA (K_d: 0.25 μM) than FHD (K_d: 2.9 μM) and thus will likely bind DNA alone or both DNA and FHD before it dissociates from DNA to bind FHD under physiological conditions (**Figure 80**).

The FHD and T–Box bound Nel C DNA (the FHD cognate sequence) with K_ds of 1.4 and 91 μM, respectively. It is important to note that there is a large error for the T–Box K_d as the curve did not reach saturation. The anisotropy curve to determine how FHD, Nel C and T–Box interact was performed in the same manner as with the SSL DNA, except that the Nel C DNA was saturated with FHD and T–Box was titrated in. In this case there was an almost immediate drop in anisotropy value indicating that FHD was preferentially interacting with T–Box and dissociating from the DNA. The K_ds for FHD binding to Nel C and T–Box were 1.4 and 2.9 μM, respectively. These K_ds are very similar and, as a result, FHD might not preferentially bind either the DNA or T–Box. In this case it is unlikely that any ternary complex forms but rather T–Box inhibits DNA binding (**Figure 80**). These results also suggest that the FHD does not bind T–Box on T–Box's DNA binding site, but that T–Box could in fact bind FHD's DNA binding site.

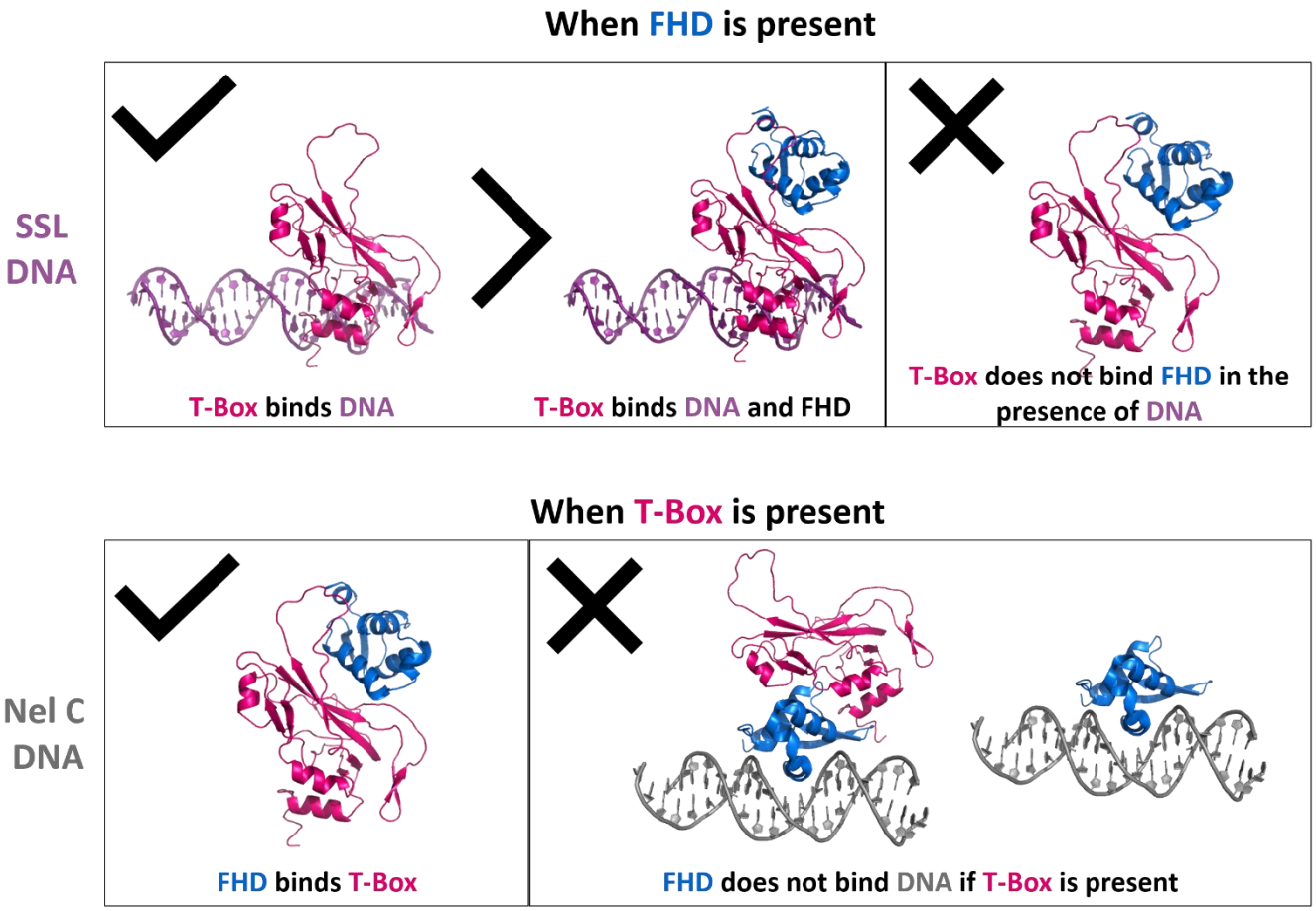


Figure 80: Fluorescence anisotropy indicates that T–Box can interact with SSL DNA and FHD forming a ternary complex, whereas, when FHD binds T–Box it dissociates from Nel C DNA.

Fluorescence anisotropy data suggests that TBR1 (pink) can interact with both FOXP2 (blue) and SSL DNA (purple) simultaneously. The small increase in anisotropy value, however, suggests that the predominant species is likely the TBR1–SSL complex not the ternary complex. The fluorescence anisotropy data indicates that FOXP2 dissociates from Nel C DNA (grey) to interact with TBR1. This suggests that TBR1 might act as an inhibitor for FOXP2 DNA binding. The FOXP2 structure was obtained from PDB (2A07)(Stroud *et al.* 2006). The TBR1 structure was modelled using I–TASSER. Images of proteins rendered with PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.

This helps shed light on a small section of what is likely a complicated mix of regulatory pathways implicated in various neurological disorders. For instance, CASK is co–imported into the nucleus by TBR1. CASK and TBR1 are known to form a complex with CINAP. The absence of CINAP is thought to reduce the regulation activity of TBR1 and CASK (Wang *et al.* 2004). CASK, TBR1 and FOXP2 are also known to form a complex in specific areas of the brain where both TBR1 and FOXP2 are co–expressed (Becker *et al.* 2018). Together, FOXP2 and TBR1 are known to regulate and or interact with multiple genes associated with autism and speech and language development such as AUTS2, GRIN2B and RELN

(Bedogni *et al.* 2010; Hsueh 2006; Wang *et al.* 2004). This study shows that it is likely that the FOXP2–TBR1 interaction inhibits FOXP2 DNA binding but not TBR1 DNA binding. Mutations in either TBR1 or FOXP2 that inhibit the protein–protein interaction could, therefore, allow FOXP2 DNA binding resulting in mis–regulation of its gene targets. TBR1 can interact with FOXP2 and DNA simultaneously and FOXP2 might mediate interactions with other proteins as with the CASK–CINAP–TBR1 complex. Mutations that inhibit FOXP2–TBR1 binding could result in mis–assembly of the regulatory complex and downstream mis–regulation of various gene targets. In healthy individuals the TBR1 might inhibit FOXP2 DNA binding and at the same time the interaction might aid in TBR1 transcriptional activity.

CONCLUSION

The TBR1 T–Box domain is monomeric in solution and can interact with DNA containing either a single or a palindromic T–Box binding site. DNA binding studies show that TBR1 T–Box forms interactions in both the minor and major groove of the DNA and the flanking regions play a role in DNA binding. It is unlikely that TBR1 T–Box binds the palindrome DNA as a dimer, however, further smFRET studies need to be done to confirm this. Nevertheless, this does indicate that the palindrome sites previously used in T–Box DNA binding studies are unlikely to be physiologically relevant or significant functionally or structurally, including previously found quaternary structures with small dimer interfaces. Rather, given the differences in the dissociation constants between the two DNA sequences, this suggests that TBR1 DNA binding is sensitive to flanking regions in DNA, which is likely important in controlling its DNA regulation. The TBR1 T–Box domain also forms a direct interaction with the FOXP2 FHD, strongly suggesting that, because both TBR1 and FOXP2 are co–expressed in the brain and have been shown to be involved in the same developmental pathways, their interaction forms part of the network required for correctly regulated neural development. Additionally, TBR1 T–Box binds the FHD and DNA simultaneously whereas the protein–protein interaction inhibits FHD's DNA binding ability. This suggests that the FOXP2–TBR1 interaction likely inhibits FOXP2 while possibly activating TBR1. This provides further insight into understanding the complex regulatory network involving FOXP2 and TBR1 (**Figure 81**). Overall, both proteins likely interact with each other and with their respective DNA sequences as part of a dynamic system designed to recruit other transcription machinery, inhibit, or activate each other's DNA binding, or even to form bridges between sections of chromatin.

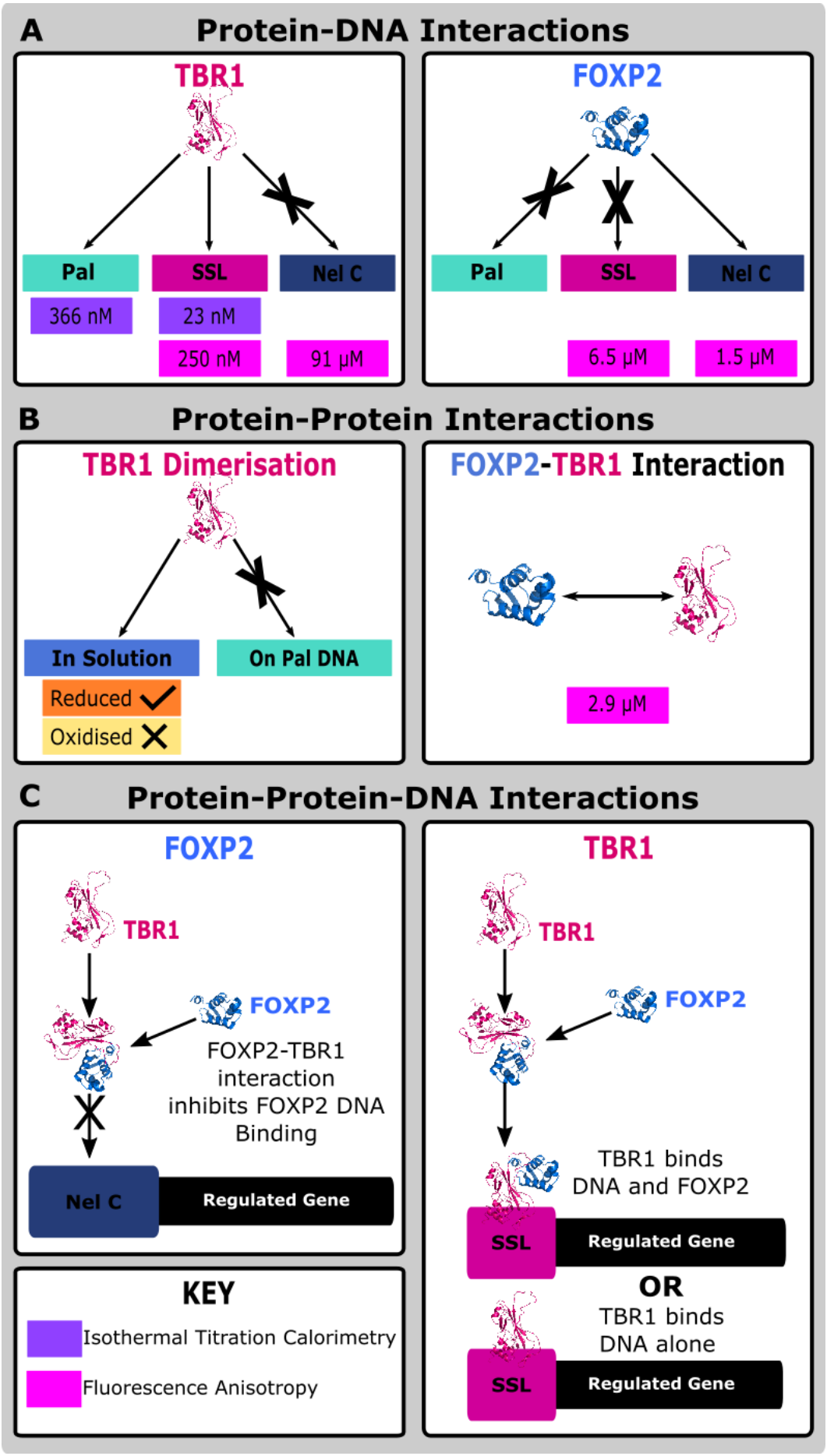

A
Protein-DNA Interactions
TBR1
Pal
SSL
Nel C
366 nM
23 nM
250 nM
91 µM
FOXP2
Pal
SSL
Nel C
6.5 µM
1.5 µM
B
Protein-Protein Interactions
TBR1 Dimerisation
In Solution
On Pal DNA
Reduced
Oxidised
FOXP2-TBR1 Interaction
2.9 µM
C
Protein-Protein-DNA Interactions
FOXP2
TBR1
FOXP2
FOXP2-TBR1 interaction inhibits FOXP2 DNA Binding
Nel C
Regulated Gene
KEY
Isothermal Titration Calorimetry
Fluorescence Anisotropy
TBR1
TBR1
FOXP2
TBR1 binds DNA and FOXP2
SSL
Regulated Gene
OR
TBR1 binds DNA alone
SSL
Regulated Gene

Figure 81: Summary of the various interactions between FOXP2, TBR1 and their DNA sequences.

A: The Protein–DNA interactions formed between both TBR1 and FHD and their DNA sequences (SSL, Pal, and Nel C DNA). The dissociation constants are for each interaction obtained using ITC (purple) and FA (pink) are displayed. **B**: The TBR1 T–Box is monomeric in solution under reducing conditions but may dimerise in non–reducing conditions. The TBR1 T–Box forms a direct interaction with the FHD of FOXP2. The dissociation constant obtained from was 2.9 μM. **C**: The FOXP2–TBR1 interaction inhibits FOXP2 binding to Nel C DNA, whereas the interaction does not inhibit TBR1 DNA binding. In fact, TBR1 can bind SSL DNA and FHD simultaneously. The FOXP2 structure was obtained from the crystal structure PDB (2A07)(Stroud *et al.* 2006) whereas the TBR1 structure was modelled using ITASSER. Images of proteins rendered with PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC. Image created using Inkscape 0.92.4.

REFERENCES

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). *Molecular biology of the cell.*, Fourth E.d, New York: Garland science.

Amoutzias, G. D., Robertson, D. L., Van de Peer, Y., & Oliver, S. G. (2008). Choose your partners: dimerization in eukaryotic transcription factors. *Trends in Biochemical Sciences*, **33**(5), 220–229.

Aravind, L., Anantharaman, V., Balaji, S., Babu, M. M., & Iyer, L. M. (2005). The many faces of the helix-turn-helix domain: Transcription regulation and beyond. *FEMS Microbiology Reviews*, **29**(2), 231–262.

Arnau, J., Lauritzen, C., Petersen, G. E., & Pedersen, J. (2006). Current strategies for the use of affinity tags and tag removal for the purification of recombinant proteins. *Protein Expression and Purification*, **48**, 1–13.

Axelrod, D., Burghardt, T., & Thompson, N. (1984). Total Internal Reflection Fluorescence. *Annual Review of Biophysics and Bioengineering*, **13**, 247–268.

Bandukwala, H. S., Wu, Y., Feuerer, M., … Chen, L. (2011). Structure of a Domain-Swapped FOXP3 Dimer on DNA and Its Function in Regulatory T Cells. *Immunity*, **34**(4), 479–491.

Banham, A. H., Beasley, N., Campo, E., … Cordell, J. L. (2001). The FOXP1 winged helix transcription factor is a novel candidate tumor suppressor gene on chromosome 3p. *Cancer Research*, **61**(24), 8820–8829.

Basei, F. L., Meirelles, G. V., Righetto, G. L., Migueleti, D. L. dos S., Smetana, J. H. C., & Kobarg, J. (2015). New interaction partners for Nek4.1 and Nek4.2 isoforms: From the DNA damage response to RNA splicing. *Proteome Science*, **13**(1), 1–13.

Bedogni, F., Hodge, R. D., Nelson, B. R., Frederick, E. A., Daza, R. A., & Hevner, R. F. (2011). NIH Public Access. *Gene Expression Patterns*, **10**(1), 1–16.

Bellomy, G., & Record, T. (1990). Stable DNA loops in vivo and in vitro: roles in gene regulation at a distance and in biophysical characterization of DNA. *Progress in Nucleic Acid Research and Molecular Biology*, **39**(1), 81–128.

Benayoun, B. A., & Veitia, R. A. (2009). A post-translational modification code for transcription factors: sorting through a sea of signals. *Trends in Cell Biology*, **19**(5), 189–197.

Bennett, C., Christie, J., Ramsdell, F., … Ochs, H. D. (2001). The immune dysregulation, polyendocrinopathy, enteropathy, X-linked syndrome (IPEX) is caused by mutations of FOXP3. *Nature Genetics*, **27**, 20–21.

Berg, O. G. (1978). On diffusion-controlled dissociation. *Chemical Physics*, **31**(1), 47–57.

Berg, O. G., Winter, R. B., & von Hippel, P. H. (1981). Diffusion-Driven Mechanisms of Protein Translocation on Nucleic Acids. 1. Models and Theory. *Biochemistry*, **20**(24), 6929–6948.

Bewley, C. A., Gronenborn, A. M., & Clore, G. M. (1998). Minor groove-binding architectural proteins: Structure, function, and DNA recognition. *Annual Review of Biophysics and Biomolecular Structure*, **27**, 105–131.

Blane, A., Dirr, H. W., & Fanucchi, S. (2018). A Phosphomimetic Study Implicates Ser557 in Regulation of FOXP2 DNA Binding. *Protein Journal*, **37**(4), 311–323.

Blane, A., & Fanucchi, S. (2015). Effect of pH on the Structure and DNA Binding of the FOXP2 Forkhead Domain. *Biochemistry*, **54**(25), 4001–4007.

Blurnenfeld, F., & Gardner, J. (1985). Gel Filtration-An Innovative Separation Technique. *Journal of Chemical Education*, **62**(8), 715–717.

Boldt, K., Van Reeuwijk, J., Lu, Q., … Roepman, R. (2016). An organelle-specific protein landscape identifies novel diseases and molecular mechanisms. *Nature Communications*, **7**(May), 1–13.

Bork, P., Holm, L., & Sander, C. (1994). The Immunoglobulin Fold. *Journal of Molecular Biology*, **242**, 309–320.

Boutchueng-Djidjou, M., Belleau, P., Bilodeau, N., … Faure, R. L. (2018). A type 2 diabetes disease module with a high collective influence for Cdk2 and PTPLAD1 is localized in endosomes. *PLoS ONE*, **13**(10), 1–25.

Braman, J., Papworthf, C., & Greener, A. (1996). Site-directed mutagenesis using double-stranded plasmid DNA templates. In *methods in Molecular Biology*, Vol. 57, pp. 31–44.

Brennan, R. G., Roderick, S. L., Takeda, Y., & Matthews, B. W. (1990). Protein-DNA conformational changes in the crystal structure of a λ Cro-operator complex. *Proceedings of the National Academy of Sciences of the United States of America*, **87**(20), 8165–8169.

Brent, M. M., Anand, R., & Marmorstein, R. (2008). Structural Basis for DNA Recognition by FoxO1 and its Regulation by Post-Translational Modification. *Structure*, **16**(9), 1407–1416.

Buggert, M., Tauriainen, J., Yamamoto, T., … Karlsson, A. C. (2014). T-bet and Eomes Are Differentially Linked to the Exhausted Phenotype of CD8+ T Cells in HIV Infection. *PLoS Pathogens*, **10**(7), 1–15.

Bulfone, A., Smiga, S. M., Shimamura, K., Peterson, A., Puelles, L., & Rubenstein, J. L. R. (1995). T-Brain-1: A homolog of Brachyury whose expression defines molecularly distinct domains within the cerebral cortex. *Neuron*, **15**(1), 63–78.

Buratowski, S., & Chodosh, L. A. (2001). Mobility Shift DNA-Binding Assay Using Gel Electrophoresis. In *Current Protocols in Pharmacology*, Vol. 13, pp. 1–12.

Burley, S. K. (1997). Eukaryotic Transcription Factor-DNA Complexes. *Annual Review of Biomolecular Structure*, **26**, 289–325.

Cao, J., Luo, Z., Cheng, Q., … Song, X. (2015). Three-dimensional regulation of transcription. *Protein and Cell*, **6**(4), 241–253.

Carlsson, P., & Mahlapuu, M. (2002). Forkhead transcription factors: key players in development and metabolism. *Developmental Biology*, **250**(1), 1–23.

Chattopadhyay, A., Jones, N. G., Nietlispach, D., … Carrington, M. (2005). Structure of the C-terminal domain from Trypanosoma brucei variant surface glycoprotein MITat1.2. *Journal of Biological Chemistry*, **280**(8), 7228–7235.

Chen, M. T., Sun, H. F., Li, L. D., … Jin, W. (2018). Downregulation of FOXP2 promotes breast cancer migration and invasion through TGFβ/SMAD signaling pathway. *Oncology Letters*, **15**(6), 8582–8588.

Christov, C., Ianev, D., Shosheva, A., & Atanasov, B. (2004). pH-dependent quenching of the fluorescence of tryptophan residues in class A β-lactamase from E. coli (TEM-1). *Zeitschrift Fur Naturforschung - Section C Journal of Biosciences*, **59**(11–12), 824–827.

Chu, Y. P., Chang, C. H., Shiu, J. H., Chang, Y. T., Chen, C. Y., & Chuang, W. J. (2011). Solution structure and backbone dynamics of the DNA-binding domain of FOXP1: Insight into its domain swapping and DNA binding. *Protein Science*, **20**(5), 908–924.

Chuang, H.-C., Huang, T.-N., & Hsueh, Y.-P. (2014). Neuronal excitation upregulates Tbr1, a high-confidence risk gene of autism, mediating Grin2b expression in the adult brain. *Frontiers in Cellular Neuroscience*, **8**(September), 1–13.

Clark, K. L., Halay, E. D., Lai, E., & Burley, S. K. (1993). Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature*, **364**(6436), 412–420.

Coll, M., Seidman, J. G., & Müller, C. W. (2002). Structure of the DNA-bound T-box domain of human TBX3, a transcription factor responsible for ulnar-mammary syndrome. *Structure*, **10**(3), 343–356.

Conaway, R. C., & Conaway, J. W. (1993). General initiation factors for RNA polymerase II. *Annual Review of*

Biochemistry, **62**, 161–190.

Corominas, R., Yang, X., Lin, G. N., … Iakoucheva, L. M. (2014). Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism. *Nature Communications*, **5**, 1–12.

Coulocheri, S. A., Pigis, D. G., Papavassiliou, K. A., & Papavassiliou, A. G. (2007). Hydrogen bonds in protein-DNA complexes: Where geometry meets plasticity. *Biochimie*, **89**(11), 1291–1303.

Cowell, I. G. (1994). Repression versus activation in the control of gene transcription. *Trends in Biochemical Sciences*, **19**(1), 38–42.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, **227**, 561–562.

Cuiffo, B. G., Campagne, A., Bell, G. W., … Karnoub, A. E. (2014). MSC-regulated microRNAs converge on the transcription factor FOXP2 and promote breast cancer metastasis. *Cell Stem Cell*, **15**(6), 762–774.

De Rubeis, S., He, X., Goldberg, A. P., … Buxbaum, J. D. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, **515**(7526), 209–215.

den Hoed, J., Sollis, E., Venselaar, H., Estruch, S. B., Deriziotis, P., & Fisher, S. E. (2018). Functional characterization of TBR1 variants in neurodevelopmental disorder. *Scientific Reports*, **8**(1), 1–11.

Deriziotis, P., & Fisher, S. E. (2017). Speech and Language: Translating the Genome. *Trends in Genetics*, **33**(9), 642–656.

Deriziotis, P., Graham, S. A., Estruch, S. B., & Fisher, S. E. (2014a). Investigating protein-protein interactions in live cells using bioluminescence resonance energy transfer. *Journal of Visualized Experiments*, (87), 1–15.

Deriziotis, P., O'Roak, B. J., Graham, S. A., … Fisher, S. E. (2014b). De novo TBR1 mutations in sporadic autism disrupt protein functions. *Nature Communications*, **5**, 1–12.

Dias, C., Estruch, S. B., Graham, S. A., … Logan, D. W. (2016). BCL11A Haploinsufficiency Causes an Intellectual Disability Syndrome and Dysregulates Transcription. *American Journal of Human Genetics*, **99**(2), 253–274.

Dominguez, C., Boelens, R., & Bonvin, A. M. J. J. (2003). HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, **125**(7), 1731–1737.

El Omari, K., de Mesmaeker, J., Karia, D., Ginn, H., Bhattacharya, S., & Mancini, E. J. (2012). Structure of the DNA-bound T-box domain of human TBX1, a transcription factor associated with the DiGeorge syndrome. *Proteins: Structure, Function and Bioinformatics*, **80**(2), 655–660.

Estruch, S. B., Graham, S. A., Quevedo, M., … Fisher, S. E. (2018). Proteomic analysis of FOXP proteins reveals interactions between cortical transcription factors associated with neurodevelopmental disorders. *Human Molecular Genetics*, **27**(7), 1212–1227.

Ferland, R. J., Cherry, T. J., Preware, P. O., Morrisey, E. E., & Walsh, C. A. (2003). Characterization of Foxp2 and Foxp1 mRNA and protein in the developing and mature brain. *Journal of Comparative Neurology*, **460**(2), 266–279.

Fisher, S. E., & Scharff, C. (2009). FOXP2 as a molecular window into speech and language. *Trends in Genetics*, **25**(4), 166–177.

Fox, A. H., Kowalski, K., King, G. F., Mackay, J. P., & Crossley, M. (1998). Key residues characteristic of GATA N-fingers are recognized by FOG. *Journal of Biological Chemistry*, **273**(50), 33595–33603.

Fox, A. H., Liew, C., Holmes, M., Kowalski, K., Mackay, J., & Crossley, M. (1999). Transcriptional cofactors of the FOG family interact with GATA proteins by means of multiple zinc fingers. *EMBO Journal*, **18**(10), 2812–2822.

Fraczkiewicz, R., & Braun, W. (1998). Exact and efficient analytical calculation of the accessible surface areas and

their gradients for macromolecules. *Journal of Computational Chemistry*, **19**(3), 319–333.

Friedman, J. R., & Kaestner, K. H. (2006). The Foxa family of transcription factors in development and metabolism. *Cellular and Molecular Life Sciences*, **63**(19–20), 2317–2328.

Garner, M. M., & Revzin, A. (1981). A gel electrophoresis mothod for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic Acids Research*, **9**(13), 3047–3060.

Gasteiger, E., Hoogland, C., Gattiker, A., … Bairoch, A. (2005). Protein Analysis Tools on the ExPASy Server. In *The Proteomics Protocols Handbook*, pp. 571–607.

Gerland, U., Moroz, J. D., & Hwa, T. (2002). Physical constraints and functional characteristics of transcription factor-DNA interaction. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(19), 12015–12020.

Gill, G. (2004). SUMO and ubiquitin in the nucleus: Different functions, similar mechanisms? *Genes and Development*, **18**(17), 2046–2059.

Gong, X., Jia, M., Ruan, Y., … Zhang, D. (2004). Association between theFOXP2 gene and autistic disorder in Chinese population. *American Journal of Medical Genetics*, **127B**(1), 113–116.

Gowers, D. M., & Halford, S. E. (2003). Protein motion from non-specific to specific DNA by three-dimensional routes aided by supercoiling. *EMBO Journal*, **22**(6), 1410–1418.

Grünberg, S., & Hahn, S. (2013). Structural insights into transcription initiation by RNA polymerase II. *Trends in Biochemical Sciences*, **38**(12), 603–611.

Harrison, S., & Aggarwal, A. (1990). DNA Recognition By Proteins With The Helix-Turn-Helix Motif. *Annual Review of Biochemistry*, **59**(1), 933–969.

Häußermann, K., Young, G., Kukura, P., & Dietz, H. (2019). Dissecting FOXP2 Oligomerization and DNA Binding. *Angewandte Chemie - International Edition*, **58**(23), 7662–7667.

Herrero, M. J., & Gitton, Y. (2018). The untold stories of the speech gene, the FOXP2 cancer gene. *Genes and Cancer*, **9**(1–2), 818–845.

Herrmann, B. G. (1995). The mouse Brachyury (T) gene. *Seminars in Developmental Biologyy*, **6**, 385–394.

Hiroi, Y., Kudoh, S., Monzen, K., … Komuro, I. (2001). Tbx5 associates with Nkx2-5 and synergistically promotes cardiomyocyte differentiation. *Nature Genetics*, **28**(3), 276–280.

Holden, S. J., Uphoff, S., Hohlbein, J., … Kapanidis, A. N. (2010). Defining the Limits of Single-Molecule FRET Resolution in TIRF Microscopy. *Biophysj*, **99**(9), 3102–3111.

Hong, P., Koza, S., & Bouvier, E. S. P. (2012). A review size-exclusion chromatography for the analysis of protein biotherapeutics and their aggregates. *Journal of Liquid Chromatography and Related Technologies*, **35**(20), 2923–2950.

Hsueh, Y., Wang, T., Yang, F., & Sheng, M. (2000). Nuclear translocation and guanylate kinase CASK / LIN-2. *Nature*, **404**(March).

Hu, H., Wang, B., Borde, M., … Rao, A. (2006). Foxp1 is an essential transcriptional regulator of B cell development. *Nature Immunology*, **7**(8), 819–826.

Huang, T. N., & Hsueh, Y. P. (2015). Brain-specific transcriptional regulator T-brain-1 controls brain wiring and neuronal activity in autism spectrum disorders. *Frontiers in Neuroscience*, **9**(NOV), 1–13.

Huffman, J. L., & Brennan, R. G. (2002). Prokaryotic transcription regulators: More than just the helix-turn-helix motif. *Current Opinion in Structural Biology*, **12**(1), 98–106.

Huttlin, E. L., Bruckner, R. J., Paulo, J. A., … Wade, J. (2017). Architecture of the human interactome defines protein communities and disease networks. *Nature*, **545**(7655), 505–509.

Jia, W. Z., Yu, T., An, Q., … Xiao, G. (2016). MicroRNA-190 regulates FOXP2 genes in human gastric cancer. *OncoTargets and Therapy*, **9**, 3643–3651.

Jiang, M., Sun, L., Isupov, M. N., … Wu, Y. (2019). Structural basis for the Target DNA recognition and binding by the MYB domain of phosphate starvation response 1. *FEBS Journal*, **286**(14), 2809–2821.

Joo, C., Balci, H., Ishitsuka, Y., Buranachai, C., & Ha, T. (2008). Advance in Single- Molecule Fluorescence Methods for Molecular Biology. *Annual Review of Biochemistry*, **77**, 51–76.

Kim, J. L., Nikolov, D. B., & Burley, S. K. (1993a). Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature*, **365**(6446), 520–527.

Kim, Y., Geiger, J. H., Hahn, S., & Sigler, P. B. (1993b). Crystal structure of a yeast TBP/TATA-box complex. *Nature*, **365**(6446), 512–520.

Kimura, N., Nakashima, K., Ueno, M., Kiyama, H., & Taga, T. (1999). A novel mammalian T-box-containing gene, Tbr2, expressed in mouse developing brain. *Developmental Brain Research*, **115**(2), 183–193.

Kloks, C. P., Spronk, C. A., Lasonder, E., … Hilbers, C. W. (2002). The solution structure and DNA-binding properties of the cold-shock domain of the human Y-box protein YB-1. *Journal of Molecular Biology*, **316**, 317–326.

Koh, K. P., Sundrud, M. S., & Rao, A. (2009). Domain requirements and sequence specificity of DNA binding for the forkhead transcription factor FOXP3. *PLoS ONE*, **4**(12), 1–9.

Konopka, G., Bomar, J. M., Winden, K., … Geschwind, D. H. (2009). Human-specific transcriptional regulation of CNS development genes by FOXP2. *Nature*, **462**(7270), 213–217.

Koudelka, G. B., Mauro, S. A., & Ciubotaru, M. (2006). Indirect Readout of DNA Sequence by Proteins: The Roles of DNA Sequence-Dependent Intrinsic and Extrinsic Forces. *Progress in Nucleic Acid Research and Molecular Biology*, **81**(06), 143–177.

Krumm, N., O'Roak, B. J., Shendure, J., & Eichler, E. E. (2014). A de novo convergence of autism genetics and molecular neuroscience. *Trends in Neurosciences*, **37**(2), 95–105.

Ladbury, J. E., & Chowdhry, B. Z. (1996). Sensing the heat: the application of isothermal titration calorimetry to thermodynamic studies of biomolecular interactions. *Chemistry & Biology*, **3**(10), 791–801.

Laemmli, U. (1970). Cleavage of Structural Proteins during the Assembly of the Head of Bacteriophage T4. *Nature*, **227**, 680–685.

Lakowicz, J. R. (2006). *Principles of fluorescence spectroscopy*. *Principles of Fluorescence Spectroscopy*. doi:10.1007/978-0-387-46312-4

Larkin, M. A., Blackshields, G., Brown, N. P., … Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**(21), 2947–2948.

Latchman, D. S. (1997). Transcription factors: An overview. *International Journal of Biochemistry and Cell Biology*, **29**(12), 1305–1312.

Latchman, D. S. (2004). DNA Sequences, Transcription Factors and Chromatin Structure. *Eukaryotic Transcription Factors*, 1–22.

Latchman, D. S. (2005). *Gene regulation.*, Fifth E.d, New York: Taylor & Francis Group.

Lathe, G., & Ruthven, R. (1956). The Separation of Substances and Estimation of their Relative Molecular Sizes by the use of Columns of Starch in Water. *Biochemical Journal*, **62**, 665–674.

Lee, C.-H., Lu, W., Michel, J., … Gouaux, E. (2014). NMDA receptor structures reveal subunit arrangement and pore architecture. *Nature*, **511**(7508), 191–197.

Lee, T. I., & Young, R. A. (2013). Transcriptional Regulation and its Misregulation in Disease Tong. *Cell*, **152**(6), 1237–1251.

Li, J., Zhang, W., Yang, H., … Coba, M. P. (2017). Spatio-temporal profile of postynaptic interactomes integrates components of complex brain disorders. *Nature Neuroscience*, **20**(8), 1150–1161.

Li, S., Weidenfeld, J., & Morrisey, E. E. (2004). Transcriptional and DNA binding activity of the Foxp1/2/4 family is modulated by heterotypic and homotypic protein interactions. *Molecular and Cellular Biology*, **24**(2), 809–22.

Li, X., Wang, W., Wang, J., … Chen, J. (2015). Proteomic analyses reveal distinct chromatin-associated and soluble transcription factor complexes. *Molecular Systems Biology*, **11**(1), 775.

LiCata, V. J., & Wowor, A. J. (2008). Applications of Fluorescence Anisotropy to the Study of Protein-DNA Interactions. *Methods in Cell Biology*, **84**(07), 243–262.

Liu, C. F., Brandt, G. S., Hoang, Q. Q., … Petsko, G. A. (2016). Crystal structure of the DNA binding domain of the transcription factor T-bet suggests simultaneous recognition of distant genome sites. *Proceedings of the National Academy of Sciences of the United States of America*, **113**(43), E6572–E6581.

Liu, X., Salokas, K., Tamene, F., … Varjosalo, M. (2018). An AP-MS- and BioID-compatible MAC-tag enables comprehensive mapping of protein interactions and subcellular localizations. *Nature Communications*, **9**(1), 1–16.

Long, Q. Z., Du, Y. F., Ding, X. Y., … Liu, X. G. (2012). Replication and fine mapping for association of the C2orf43, FOXP4, GPRC6A and RFX6 genes with prostate cancer in the Chinese population. *PLoS ONE*, **7**(5), 5–10.

Lord, C., Cook, E. H., Leventhal, B. L., & Amaral, D. G. (2000). Autism spectrum disorders. *Neuron*, **28**, 355–363.

Luisi, B. F., Xu, W. X., Otwinowski, Z., Freedman, L. P., Yamamoto, K. R., & Sigler, P. B. (1991). Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. *Nature*, pp. 497–505.

Luscombe, N., Laskowski, R., & Thronton, J. (2001). Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. *Nucleic Acids Research*, **29**(13), 2860–2874.

Maio, N., Kim, K. S., Singh, A., & Rouault, T. A. (2017). A Single Adaptable Cochaperone-Scaffold Complex Delivers Nascent Iron-Sulfur Clusters to Mammalian Respiratory Chain Complexes I–III. *Cell Metabolism*, **25**(4), 945-953.e6.

Marshall, J. J. (1970). Comments on the use of Blue Dextran in gel chromatography. *Journal of Chromatography A*, **53**(2), 379–380.

Masse, J. E., Wong, B., Yen, Y. M., Allain, F. H. T., Johnson, R. C., & Feigon, J. (2002). The S. cerevisiae architectural HMGB protein NHP6A complexed with DNA: DNA and protein conformational changes upon binding. *Journal of Molecular Biology*, **323**(2), 263–284.

Mendoza, E., & Scharff, C. (2017). Protein-Protein Interaction Among the FoxP Family Members and their Regulation of Two Target Genes, VLDLR and CNTNAP2 in the Zebra Finch Song System. *Frontiers in Molecular Neuroscience*, **10**(May), 1–15.

Miller, S. A., Huang, A. C., Miazgowicz, M. M., Brassil, M. M., & Weinmann, A. S. (2008). Coordinated but physically separable interaction with H3K27-demethylase and H3K4-methyltransferase activities are required for T-box protein-mediated activation of developmental gene expression. *Genes and Development*, **22**(21), 2980–2993.

Mitsui, S., Igarashi, K. M., Mori, K., & Yoshihara, Y. (2011). Genetic visualization of the secondary olfactory pathway in Tbx21 transgenic mice. *Neural Systems & Circuits*, **1**(1), 5.

Morris, G., & Fanucchi, S. (2016). A Key Evolutionary Mutation Enhances DNA Binding of the FOXP2 Forkhead Domain. *Biochemistry*, **55**(13), 1959–1967.

Morris, G., Stoychev, S., Naicker, P., Dirr, H. W., & Fanucchi, S. (2018). The forkhead domain hinge-loop plays a pivotal role in DNA binding and transcriptional activity of FOXP2. *Biological Chemistry*, **399**(8), 881–893.

Mukame, Z., Konopka, G., Wexler, E., … Geschwind, D. H. (2011). Regulation of MET by FOXP2, genes implicated in higher cognitive dysfunction and autism risk. *Journal of Neuroscience*, **31**(32), 11437–11442.

Muller, C., & Herrmann, B. (1997). Crystallographic structure of the T domain – DNA complex of the Brachyury transcription factor. *Nature*, **389**, 884–888.

Naiche, L. A., Harrelson, Z., Kelly, R. G., & Papaioannou, V. E. (2005). T-Box Genes in Vertebrate Development. *Annual Review of Genetics*, **39**, 219–239.

Nelson, C. S., Fuller, C. K., Fordyce, P. M., Greninger, A. L., Li, H., & DeRisi, J. L. (2013). Microfluidic affinity and ChIP-seq analyses converge on a conserved FOXP2-binding motif in chimp and human, which enables the detection of evolutionarily novel targets. *Nucleic Acids Research*, **41**(12), 5991–6004.

Nikolov, D. B., Chen, H., Halay, E. D., Hoffmann, A., Roeder, R. G., & Burley, S. K. (1996). Crystal structure of a human TATA box-binding protein/TATA element complex. *Proceedings of the National Academy of Sciences of the United States of America*, **93**(10), 4862–4867.

Oestreich, K. J., & Weinmann, A. S. (2012). T-bet employs diverse regulatory mechanisms to repress transcription. *Trends in Immunology*, **33**(2), 78–83.

Olins, D. E., & Olins, A. L. (2003). Chromatin history: Our view from the bridge. *Nature Reviews Molecular Cell Biology*, **4**(10), 809–814.

Orchard, S., Kerrien, S., Abbani, S., … Hermjakob, H. (2012). Protein interaction data curation: The International Molecular Exchange (IMEx) consortium. *Nature Methods*, **9**(4), 345–350.

Otwinowski, Z., Schevitz, R. W., Zhang, R. G., … Sigler, P. B. (1988). Crystal structure of trp represser/operator complex at atomic resolution. *Nature*, pp. 321–329.

Pan, G., Li, J., Zhou, Y., Zheng, H., & Pei, D. (2006). A negative feedback loop of transcription factors that controls stem cell pluripotency and self-renewal. *The FASEB Journal*, **20**(10), 1730–1732.

Pares- Matos, E. . (2014). Electrophoretic Mobility- shift and Super-Shift Assays for studies and Charactreisastion of Protein- DNA Complexes. In *Gene Regulation: Methods and Protocols, Methods in Molecular Biology*, Vol. 977, pp. 159–167.

Patikoglou, G. A., Kim, J. L., Sun, L., Yang, S. H., Kodadek, T., & Burley, S. K. (1999). TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. *Genes and Development*, **13**(24), 3217–3230.

Pedersen, A. G., Baldi, P., Chauvin, Y., & Brunak, S. (1999). The biology of eukaryotic promoter prediction--a review. *Computers & Chemistry*, **23**, 191–207.

Peng, Y., & Jahroudi, N. (2002). The NFY transcription factor functions as a repressor and activator of the von Willebrand factor promoter. *Blood*, **99**(7), 2408–2417.

Perumal, K., Dirr, H. W., & Fanucchi, S. (2015). A Single Amino Acid in the Hinge Loop Region of the FOXP Forkhead Domain is Significant for Dimerisation. *Protein Journal*, **34**(2), 111–121.

Pesce, A., Dewilde, S., Kiger, L., … Bolognesi, M. (2001). Very high resolution structure of a trematode hemoglobin displaying a TyrB10-TyrE7 heme distal residue pair and high oxygen affinity. *Journal of Molecular Biology*, **309**(5), 1153–1164.

Pettersen, E. F., Goddard, T. D., Huang, C. C., … Ferrin, T. E. (2004). UCSF Chimera - A visualization system for

exploratory research and analysis. *Journal of Computational Chemistry*, **25**(13), 1605–1612.

Pradhan, L., Gopal, S., Li, S., … Nam, H. J. (2016). Intermolecular Interactions of Cardiac Transcription Factors NKX2.5 and TBX5. *Biochemistry*, **55**(12), 1702–1710.

Privalov, P. L., Dragan, A. I., Crane-Robinson, C., Breslauer, K. J., Remeta, D. P., & Minetti, C. A. S. A. (2007). What Drives Proteins into the Major or Minor Grooves of DNA? *Journal of Molecular Biology*, **365**(1), 1–9.

Ribeiro, J., Melo, F., & Schüller, A. (2015). PDIviz: Analysis and visualization of protein-DNA binding interfaces. *Bioinformatics*, **31**(16), 2751–2753.

Roberts, S. G. E. (2000). Mechanisms of action of transcription activation and repression domains. *Cellular and Molecular Life Sciences*, **57**(8–9), 1149–1160.

Roeder, R. . (1991). The complexities of eukaryotic transcription initiation: Regulation of preinitiation complex assembly. *Science*, **16**, 402–408.

Rohs, R., Jin, X., West, S. M., Joshi, R., Honig, B., & Mann, R. S. (2010). Origins of Specificity in Protein-DNA Recognition. *Annual Review of Biochemistry*, **79**(1), 233–269.

Rolland, T., Taşan, M., Charloteaux, B., … Vidal, M. (2014). A proteome-scale map of the human interactome network. *Cell*, **159**(5), 1212–1226.

Roy, A., Kucukural, A., & Zhang, Y. (2010). I-TASSER: A unified platform for automated protein structure and function prediction. *Nature Protocols*, **5**(4), 725–738.

Roy, R., Hohng, S., & Ha, T. (2008). A practical guide to single molecule FRET. *Nature Methods*, **5**(6), 507–516.

Sainsbury, S., Bernecky, C., & Cramer, P. (2015). Structural basis of transcription initiation by RNA polymerase II. *Nature Reviews Molecular Cell Biology*, **16**(3), 129–143.

Sanjuán, J., Tolosa, A., González, J. C., … Frutos, R. De. (2006). Association between FOXP2 polymorphisms and schizophrenia with auditory hallucinations. *Psychiatric Genetics*, **16**(2), 67–72.

Schägger, H. (2006). Tricine-SDS-PAGE. *Nature Protocols*, **1**(1), 16–22.

Schägger, H., & Von Jagow, G. (1987). Tricine- Sodium Dodecyl Sulfate- Polyacrylamide Gel Electrophoresis for the Separation of Protein in the Range from 1 to 100 kDa. *Analytical Biochemistry*, **166**, 368–379.

Schuler, B. (2013). Single-molecule FRET of protein structure and dynamics - a primer. *Journal of Nanobiotechnology*, **11**, 1–17.

Seeman, N. C., Rosenberg, J. M., & Rich, A. (1976). Sequence specific recognition of double helical nucleic acids by proteins. *Proceedings of the National Academy of Sciences of the United States of America*, **73**(3), 804–808.

Sharma, S., Yang, S., Zhu, L., Reckamp, K., & Gardner, B. (2005). Tumor Cyclooxygenase-2 / Prostaglandin E 2 − Dependent Promotion of FOXP3 Expression and CD4 + CD25 + T Regulatory Cell Activities in Lung Cancer. *Cancer Research*, **65**(12), 5211–5220.

Sharrocks, A. D. (2001). The ETS-domain transcription factor family. *Nature Reviews Molecular Cell Biology*, **2**(11), 827–837.

Shrivastava, A., & Calame, K. (1994). An analysis of genes regulated by the multi-functional transcription factor Yin Yang-1. *Nucleic Acids Research*, **22**(24), 5151–5155.

Shu, W., Lu, M. M., Zhang, Y., Tucker, P. W., Zhou, D., & Morrisey, E. E. (2007). Foxp2 and Foxp1 cooperatively regulate lung and esophagus development. *Development*, **134**(10), 1991–2000.

Shu, W., Yang, H., Zhang, L., Lu, M. M., & Morrisey, E. E. (2001). Characterization of a New Subfamily of Winged-

helix/Forkhead (Fox) Genes That Are Expressed in the Lung and Act as Transcriptional Repressors. *Journal of Biological Chemistry*, **276**(29), 27488–27497.

Sidorova, N. Y., Hung, S., & Rau, D. C. (2010). Stabilizing labile DNA-protein complexes in polyacrylamide gels. *Electrophoresis*, **31**, 648–653.

Sin, C., Li, H., & Crawford, D. A. (2014). Transcriptional Regulation by FOXP1, FOXP2, and FOXP4 Dimerization. *Journal of Molecular Neuroscience*, **55**(2), 437–448.

Sollis, E., Deriziotis, P., Saitsu, H., … Fisher, S. E. (2017). Equivalent missense variant in the FOXP2 and FOXP1 transcription factors causes distinct neurodevelopmental disorders. *Human Mutation*, **38**(11), 1542–1554.

Sollis, E., Graham, S. A., Vino, A., … Fisher, S. E. (2016). Identification and functional characterization of de novo FOXP1 variants provides novel insights into the etiology of neurodevelopmental disorder. *Human Molecular Genetics*, **25**(3), 546–557.

Španiel, F., Horáček, J., Tintěra, J., … Höschl, C. (2011). Genetic variation in FOXP2 alters grey matter concentrations in schizophrenia patients. *Neuroscience Letters*, **493**(3), 131–135.

Stellato, C., Nassa, G., Tarallo, R., … Ambrosino, C. (2015). Identification of cytoplasmic proteins interacting with unliganded estrogen receptor α and β in human breast cancer cells. *Proteomics*, **15**(11), 1801–1807.

Stirnimann, C. U., Ptchelkine, D., Grimm, C., & Müller, C. W. (2010). Structural basis of TBX5-DNA recognition: The T-box domain in Its DNA-bound and -unbound form. *Journal of Molecular Biology*, **400**(1), 71–81.

Stroud, J. C., Wu, Y., Bates, D. L., … Chen, L. (2006). Structure of the forkhead domain of FOXP2 bound to DNA. *Structure*, **14**(1), 159–166.

Stryer, L., & Haughland, R. (1967). Energy transfer: A spectroscopic ruler. *Proceedings of the National Academy of Sciences of the United States of America*, **58**, 719–726.

Terpe, K. (2003). Overview of tag protein fusions: From molecular and biochemical fundamentals to commercial systems. *Applied Microbiology and Biotechnology*, **60**(5), 523–533.

Teufel, A., Wong, E. A., Mukhopadhyay, M., Malik, N., & Westphal, H. (2003). FoxP4, a novel forkhead transcription factor. *Biochimica et Biophysica Acta - Gene Structure and Expression*, **1627**(2–3), 147–152.

Thompson, J. F., & Landy, A. (1988). Empirical estimation of protein-induced DNA bending angles: Applications to λ site-specific recombination complexes. *Nucleic Acids Research*, **16**(20), 9687–9705.

Tolosa, A., Sanjuán, J., Dagnall, A. M., Moltó, M. D., Herrero, N., & de Frutos, R. (2010). FOXP2 gene and language impairment in schizophrenia: Association and epigenetic studies. *BMC Medical Genetics*, **11**(1), 1–8.

Tootle, T. L., & Rebay, I. (2005). Post-translational modifications influence transcription factor activity: A view from the ETS superfamily. *BioEssays*, **27**(3), 285–298.

Van Dyke, M. W., Roeder, R. G., & Sawadogo, M. (1988). Physical analysis of transcription preinitiation complex assembly on a class II gene promoter. *Science*, **241**(1986), 1335–1338.

Vance, K. W., Shaw, H. M., Rodriguez, M., Ott, S., & Goding, C. R. (2010). Mutant SOD1 and Attenuates Mutant SOD1-induced Reactive Oxygen Species Generation. *Molecular Biology of the Cell*, **21**, 2770–2779.

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., & Luscombe, N. M. (2009). A census of human transcription factors: Function, expression and evolution. *Nature Reviews Genetics*, **10**(4), 252–263.

Venter, C., Adams, M., Myers, E., … Zhu, X. (2001). The sequence of the human geonome. *Science*, **291**, 1304–1352.

Verger, A., Perdomo, J., & Crossley, M. (2003). Modification with SUMO. A role in transcriptional regulation. *EMBO Reports*, **4**(2), 137–142.

Von Hippel, P. H., & Berg, O. G. (1989). Facilitated target location in biological systems. *Journal of Biological Chemistry*, **264**(2), 675–678.

Walter, J., Steigemann, W., Singh, T., Bartunik, H., Bode, W., & Huber, R. (1982). On the disordered activation domain in trypsinogen: chemical labelling and low-temperature crystallography. *Acta Crystallographica Section B*, **B38**, 1462–1472.

Wang, B., Lin, D., Li, C., & Tucker, P. (2003). Multiple Domains Define the Expression and Regulatory Properties of Foxp1 Forkhead Transcriptional Repressors. *Journal of Biological Chemistry*, **278**(27), 24259–24268.

Warren, A. (2002). Eukaryotic transcription factors. *Current Opinion in Structural Biology*, **12**, 107–114.

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., & Barton, G. J. (2009). Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**(9), 1189–1191.

Webb, H., Steeb, O., Blane, A., … Fanucchi, S. (2017). The FOXP2 forkhead domain binds to a variety of DNA sequences with different rates and affinities. *Journal of Biochemistry*, **162**(1), 45–54.

Weill, L., Shestakova, E., & Bonnefoy, E. (2003). Transcription Factor YY1 Binds to the Murine Beta Interferon Promoter and Regulates Its Transcriptional Capacity with a Dual Activator/Repressor Role. *Journal of Virology*, **77**(5), 2903–2914.

Westermarck, J., Ivaska, J., & Corthals, G. L. (2013). Identification of protein interactions involved in cellular signaling. *Molecular and Cellular Proteomics*, **12**(7), 1752–1763.

White, A. (1959). Effect of pH on fluorescence of tryosine, tryptophan and related compounds. *The Biochemical Journal*, **71**(2), 217–220.

Whiteside, S. T., & Goodbourn, S. (1993). Signal transduction and nuclear targeting: Regulation of transcription factor activity by subcellular localisation. *Journal of Cell Science*, **104**(4), 949–955.

Wilcke, A., Ligges, C., Burkhardt, J., … Kirsten, H. (2012). Imaging genetics of FOXP2 in dyslexia. *European Journal of Human Genetics*, **20**(2), 224–229.

Wilson, V., & Conlon, F. L. (2002). The T-box family. *Genome Biology*, **3**(6), 1–7.

Wingender, E. (2013). Criteria for an updated classification of human transcription factor dna-binding domains. *Journal of Bioinformatics and Computational Biology*, **11**(1), 1–21.

Wingender, E., Schoeps, T., Haubrock, M., & Dönitz, J. (2015). TFClass: A classification of human transcription factors and their rodent orthologs. *Nucleic Acids Research*, **43**(D1), D97–D102.

Winter, R. B., Berg, O. G., & von Hippel, P. H. (1981). Diffusion-Driven Mechanisms of Protein Translocation on Nucleic Acids. 3. The Escherichia coli lac Repressor-Operator Interaction: Kinetic Measurements and Conclusions. *Biochemistry*, **20**(24), 6961–6977.

Wodak, S. ., & Janin, J. (2004). Structural Basis of Protein-Protein Interactions. In *Protein- protein interactions, Methods and applications, Methods in molecular biology,261.*, New Jersey: Humana Press, pp. 3–12.

Xu, D., & Zhang, Y. (2011). Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophysical Journal*, **101**(10), 2525–2534.

Yan, R., Xu, D., Yang, J., Walker, S., & Zhang, Y. (2013). A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. *Scientific Reports*, **3**, 1–9.

Yan, X., Zhou, H., Zhang, T., … Zhang, T. (2015). Downregulation of FOXP2 promoter human hepatocellular carcinoma cell invasion. *Tumor Biology*, **36**(12), 9611–9619.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2014). The I-TASSER suite: Protein structure and function prediction. *Nature Methods*, **12**(1), 7–8.

Yang, J., & Zhang, Y. (2015). I-TASSER server: New development for protein structure and function predictions. *Nucleic Acids Research*, **43**(W1), W174–W181.

Yang, Y., Shi, Y., Lee, J.-S., & Galvin, K. M. (1997). Everything you have ever wanted to know about. *Biochimica et Biophysica Acta*, **1332**, 49–66.

Young, C. L., Britton, Z. T., & Robinson, A. S. (2012). Recombinant protein expression and purification : A comprehensive review of affinity tags and microbial applications. *Biotechnology Journal*, **7**, 620–634.

Yue, Y., Liu, J., Cui, X., … Liu, J. (2018). VIRMA mediates preferential m6A mRNA methylation in 3'UTR and near stop codon and associates with alternative polyadenylation. *Cell Discovery*, **4**(1), 1–17.

PLAGIARISM REPORT

940523:turni_it_in.pdf

ORIGINALITY REPORT

15%	12%	6%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	hdl.handle.net Internet Source	5%
2	Submitted to University of Witwatersrand Student Paper	3%
3	Ashleigh Blane, Heini W. Dirr, Sylvia Fanucchi. "A Phosphomimetic Study Implicates Ser557 in Regulation of FOXP2 DNA Binding", The Protein Journal, 2018 Publication	<1%
4	link.springer.com Internet Source	<1%
5	epdf.pub Internet Source	<1%
6	www.nature.com Internet Source	<1%
7	epdf.tips Internet Source	<1%
8	www.hitpredict.org Internet Source	<1%