

https://dspace.org/

Why, YY1? The effect of dynamic multimerisation of YY1 and FOXP2 in DNA binding regulation

Donald, Heather

2022
https://hdl.handle.net/10539/37097

Downloaded from a DSpace repository. https://dspace.org/



Why, YY1? The effect of dynamic multimerisation of YY1 and FOXP2 in DNA binding regulation

by

Heather Donald

(848182)

Thesis

Submitted in fulfilment of the requirements for the degree

Philosophiae Doctor

in

Molecular and Cell Biology

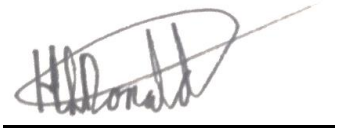
in the Faculty of Science, University of the Witwatersrand, Johannesburg, South Africa

Supervisor: Dr Sylvia Fanucchi

August 2022

DECLARATION

I declare that this thesis is my own, unaided work. It is being submitted for the Degree of Doctor of Philosophy at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.



(Signature of candidate)

24th day of August 2022 at Johannesburg

ABSTRACT

Yin Yang 1 (YY1) is a ubiquitously expressed transcription factor involved in neural development. Depending on the conditions of its own expression, YY1 is capable of either activating or repressing transcription. The oligomeric state of YY1 as well as its association with other regulatory proteins is likely to alter its DNA binding ability and transcriptional activity. Some evidence has suggested that one of YY1's binding partners may be FOXP2, another transcription factor which is also involved in neural development. The nature of this interaction, however, has not been investigated. The aim of this project was to study the potential binding mechanisms of YY1 by exploring the structure and dynamics of YY1, as well as the interactions it makes with itself, FOXP2 forkhead domain (FHD) and DNA. To do this, the structural integrity of purified recombinant YY1 and FOXP2 FHD was first confirmed using circular dichroism spectroscopy (CD) (secondary structure) and intrinsic tryptophan fluorescence spectroscopy (tertiary structure). In order to determine how YY1 oligomerisation acts as a mechanism of transcriptional control, the quaternary structure of YY1 was observed using both fluorescence anisotropy and size exclusion chromatography (SEC). The specificity of DNA binding to both proteins was studied using electrophoretic mobility shift assays (EMSAs), and the YY1-DNA interaction was further investigated using fluorescence anisotropy. The dynamics of apo and DNA-bound YY1 was determined using hydrogen deuterium exchange mass spectrometry (HDXMS), so as to determine how the YY1 structural integrity changes upon DNA binding, as this will give insight into the binding mechanism. Lastly, the YY1-FHD interaction was confirmed and assessed using fluorescence anisotropy, while supershift EMSAs were used to observe the effect of this interaction on FHD and YY1's respective interactions with DNA. In this work, YY1 was confirmed to be an intrinsically disordered protein, with large regions of disorder. YY1 was also seen to form heterogeneous mixtures of multimers, which were affected by ionic strength and by protein concentration. These oligomers could bind to DNA with decreased specificity compared to the monomer, but the oligomers could also be disrupted by saturating concentrations of DNA. The general DNA binding mechanism of YY1 involves an increase in disorder in the N-terminal region of the protein upon DNA binding. In addition to binding DNA, it was proved in this study that YY1 is capable of binding to the FOXP2 FHD. Furthermore, YY1 can simultaneously bind to both the FOXP2 FHD and DNA, although the FHD competes with YY1 for FHD specific DNA.

Overall, these data indicate that YY1 is a highly dynamic protein with behaviour typical of intrinsic disorder. Its heterogeneous, dynamic multimerisation is likely important for its ability to interact with many other transcription factors, including FOXP2. The interactions that YY1 forms with itself, the FHD and DNA form part of an intricate mechanism of transcriptional regulation by YY1, which is likely dependent on its concentration, localisation and available binding partners. This association is likely to form part of how YY1 and also FOXP2 regulate neural development, where mis-regulated expression of either protein can have detrimental effects resulting in intellectual disorders.

RESEARCH OUTPUTS

Conference Oral and Poster presentations

10th Cross-Faculty Postgraduate Symposium

- Poster presentation
- The structural and functional characterisation of YY1
- Heather Donald and Sylvia Fanucchi
- University of the Witwatersrand, Johannesburg, South Africa
- 3-10 September 2019

Molecular Biosciences Research Thrust Research Day 2019

- Poster presentation
- The structural and functional characterisation of YY1
- Heather Donald and Sylvia Fanucchi
- University of the Witwatersrand, Johannesburg, South Africa
- 28 November 2019

11th Cross-Faculty Postgraduate Symposium

- Poster presentation (Runner up prize in the Faculty of Science)
- Observing the interactions between YY1, FOXP2 and DNA
- Heather Donald and Sylvia Fanucchi
- University of the Witwatersrand, Johannesburg, South Africa
- 12-16 October 2020

Molecular Biosciences Research Thrust Research Day 2021

- Oral presentation
- Determining the Role of the Oligomerisation and Dynamics of YY1 in DNA Binding

- Heather Donald, Sindisiwe Buthelezi, Previn Naicker, Stoyan Stoychev and Sylvia Fanucchi
- University of the Witwatersrand, Johannesburg, South Africa
- 9 December 2021

South African Society of Biochemistry and Molecular Biology Congress 2022

- Electronic Poster presentation
- Determining the Role of the Oligomerisation and Dynamics of YY1 in DNA Binding
- Heather Donald, Sindisiwe Buthelezi, Previn Naicker, Stoyan Stoychev and Sylvia Fanucchi
- South African Society of Biochemistry and Molecular Biology, Virtual platform
- 23-26 January 2022

Publications

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.

“Life is simple when you have three pairs of boots.”

- *Towers of Midnight*, by Robert Jordan and Brandon Sanderson -

To my dad, my mom and my sister;

Who have always had my back.

ACKNOWLEDGEMENTS

To start with, I must thank my supervisor Dr Fanucchi. I would not have gotten here without her support, the spontaneous hour long chats in her office, the progress meetings which changed into gossip sessions, and her willingness to listen to my ideas. Without those, I would not even have started on this in the first place.

Then, I must also thank Prof Sayed and Dr Ike from the PSFRU. Both were always willing and enthusiastic with advice and help when asked. Dr Stoychev, Dr Naicker and Dr Buthelezi are also people without whom this work would look very different. Their willingness to collaborate, along with their patience in helping me with the aftermath of data acquisition was invaluable.

My colleagues from the PSFRU have made lab work what it was to me. In the midst of long hours and after lab disasters there was always someone willing to exchange ideas or just to have a chat. This stands especially true for the French Toast group – without that french toast, life would have been far less pleasant. Their input and support as well as their friendship sometimes what should have been far more difficult quite easy. My lab partner Ash (or Dr Germ now) must get special thanks: she took me out for walks when needed, made sure I fed myself and had enough coffee, and sometimes even made sure I had something without caffeine to drink.

The incredible support of my parents and sister amazes me. I know first-hand what it is like to live with someone doing their PhD (especially towards the end), and it could not have been easy for them. My parents especially deserve acknowledgment, because had they not egged me on, I might have never have taken this particular leap. I do hope someone will one day tell my dad that this was not a ‘cheap’ PhD. We shall see.

Besides family, my friends have given me a lot of understanding and support over the last few years, including endless patience with my cancellations or delays because of over-optimism about lab work going smoothly.

Lastly – and certainly not least – I would like to thank the NRF for funding me. Without their contribution, this thesis would not exist.

TABLE OF CONTENTS

LIST OF FIGURES xi
LIST OF TABLES xiii
LIST OF EQUATIONS xiv
LIST OF ABBREVIATIONS i
INTRODUCTION 1
1.1 TRANSCRIPTION 1
1.1.1. THE BASIC TRANSCRIPTION MODEL 2
1.1.2. TRANSCRIPTION FACTOR STRUCTURAL FEATURES 3
1.1.3. MEMBRANE-LESS ORGANELLES 5
1.2. FOX FAMILY 6
1.2.1. FOXP SUBFAMILY 7
1.3. FOXP2 7
1.3.3. FUNCTION 7
1.3.4. STRUCTURE 8
1.3.5. FORKHEAD DOMAIN 10
1.3.6. DYNAMICS 13
1.3.7. DNA BINDING 14
1.3.8. PROTEIN-PROTEIN INTERACTIONS 15
1.3.9. HETEROLOGOUS INTERACTIONS 16
1.4. YY1 18
1.4.1. YY1 FAMILY 18
1.4.2. YY1 FUNCTION 18
1.4.3. YY1 REPRESSION 20
1.4.4. YY1 ACTIVATION 22
1.4.5. YY1 INDUCTION 24
1.4.7. YY1 STRUCTURE 24
1.4.6. DNA STRUCTURING AND NUCLEIC ACID BINDING FUNCTIONS 28
1.4.8. YY1 DNA BINDING 29
1.4.9. YY1 PROTEIN-PROTEIN INTERACTIONS 32
RATIONALE 36
AIM AND OBJECTIVES 39
3.1. AIM 39
3.2. OBJECTIVES 39
METHODS AND MATERIALS 40

4.1. PROTEIN CONSTRUCTS 40
4.2. TRANSFORMATION 41
4.3. YY1 INDUCTION TRIALS 43
4.4. EXPRESSION 43
4.5. SDS PAGE 44
4.6. PROTEIN PURIFICATION 48
4.6.1 PREPARATION OF CELL LYSATE 48
4.6.2. IMAC 48
4.8.2. EMSA 51
4.6.3. THROMBIN CLEAVAGE 53
4.6.4. SIZE EXCLUSION 55
4.7. PROTEIN AND DNA CONCENTRATION DETERMINATION 55
4.8. DNA AND PROTEIN BINDING STUDIES OF YY1 AND FOXP2 58
4.8.1. DNA OLIGONUCLEOTIDES 58
4.8.2. SUPERSHIFT ASSAY 59
4.8.3. PROTEIN LABELLING 60
4.8.4. FLUORESCENCE ANISOTROPY 64
4.9. STRUCTURAL ANALYSIS 66
4.9.1. CIRCULAR DICHROISM 66
4.9.2. INTRINSIC TRYPTOPHAN FLUORESCENCE 67
4.9.3. SIZE EXCLUSION CHROMATOGRAPHY 69
4.10. YY1 DYNAMICS: HDXMS 71
4.10.1. SAMPLE ANALYSIS 75
4.10.2. DATA ANALYSIS 76
RESULTS 78
5.1. PROTEIN EXPRESSION 78
5.1.1. YY1 78
5.2. PROTEIN CONCENTRATION AND DNA CONTAMINATION 84
5.3. DNA BINDING STUDIES 86
5.3.1. YY1 DNA BINDING 86
5.3.2. FHD-DNA BINDING 92
5.4. STRUCTURAL CHARACTERISATION 93
5.4.1. SECONDARY STRUCTURE 93
5.4.2. YY1 TERTIARY STRUCTURE 95
5.4.3. FHD TERTIARY STRUCTURE 98
5.5. QUATERNARY STRUCTURAL STUDIES 99

5.6. YY1 STRUCTURAL DYNAMICS 106
5.6.1. ACID RICH REGION 1 107
5.6.2. THE POLYHISTIDINE REGION 109
5.6.3. ACIDIC REGION 2 110
5.6.4. GA AND GK RICH REGION. 113
5.6.5. SPACER REGION 115
5.6.6. DNA BINDING DOMAIN 119
5.6.7. OVERALL DYNAMICS 121
5.7. YY1-FOXP2 FHD INTERACTION 121
5.8. YY1-FHD INTERACTION INFLUENCE ON DNA BINDING 125
DISCUSSION 129
6.1. YY1 STRUCTURE AND INTRINSIC DISORDER 130
6.2. EFFECT OF DNA BINDING ON YY1 STRUCTURE 132
6.3. YY1 DNA BINDING 133
6.4. YY1 OLIGOMERISATION IN CONJUNCTION WITH DNA BINDING 134
6.5. YY1 AND THE FOXP2 FHD INTERACTION 141
CONCLUSION 146
REFERENCES 148

LIST OF FIGURES

Figure 1: Domain composition of FOXP subfamily members. 9
Figure 2: The structure of the FOXP2 FHD when bound to DNA. 12
Figure 3: The models describing YY1 translational repression. 21
Figure 4: Sections of YY1's sequence and their associated functions. 22
Figure 5: Transcriptional activation models for YY1. 23
Figure 6: REPO domain of YY1 bound to the 4MBT domain of MBTD1. 26
Figure 7: The crystal structure of the zinc finger domain of YY1 bound to DNA. 31
Figure 8: Diagram of a pET11-a plasmid with sites of interest. 42
Figure 9: Diagram of the processes that occur during IMAC. 49
Figure 10: Diagram of an EMSA 52
Figure 11: Separation of cleaved FOXP2 FHD from uncleaved FHD, His tags and thrombin 54
Figure 12: ATTO 550 structure. 61
Figure 13: A schematic of AsCy3 labelling the TCP tag to label TCP-YY1. 63
Figure 14: HDXMS workflow summary. 72
Figure 15: Kinetics of hydrogen-deuterium Exchange. 74
Figure 16: Expression Trials of YY1 at 20 and 30 °C. 80
Figure 17: Purification profiles and corresponding SDS PAGE gels for YY1. 81
Figure 18: FHD purification elution profiles and corresponding gels. 83
Figure 19: UV absorbance spectra of YY1 (purple) and the FOXP2 FHD (blue). 84
Figure 20: DNA gels contain YY1 to check for DNA contamination. 85
Figure 21: EMSAs of YY1 with several different DNA sequences. 87
Figure 22: A comparison of YY1 binding for oligomers, single stranded DNA, Freeze thawed and TCP-tagged YY1. 89
Figure 23: YY1 binding to DNA as seen with fluorescence anisotropy. 91
Figure 24: DNA binding assays of FHD to various DNA sequences. 92
Figure 25: CD spectra of YY1 in various conditions. 94
Figure 26: Far UV CD Spectrum of FOXP2 FHD. 95
Figure 27: Fluorescence emission spectra of YY1 under different conditions to confirm tertiary structural integrity. 97
Figure 28: Fluorescence emission spectra of YY1 in the presence and absence of DNA. 98
Figure 29: Comparison of the current and previously published FOXP2 FHD fluorescence emission spectra when excited at 280 nm. 99
Figure 30: SE-FPLC elution profile of YY1 at 25 and 50 µM, in reducing and non-reducing conditions. 101
Figure 31: SE-HPLC profiles for YY1 under high (500mM) and low (300 mM) salt conditions. 103
Figure 32: Anisotropy binding curves showing YY1-YY1 interaction at different salt concentrations. 104
Figure 33: Heat map and difference map of the first acidic region of YY1 in its apo and DNA bound forms. 108

Figure 34: HDX heat maps and difference map of the second acidic region of YY1 when bound and unbound to DNA. 111
Figure 35: HDX heat and difference maps of GA and GK rich regions of YY1 in apo form and bound to DNA. 114
Figure 36: HDXMS heat and difference maps of apo and DNA-bound YY1 in its spacer region. 117
Figure 37: Heat and difference maps of the zinc finger region of apo and DNA-bound YY1. 120
Figure 38: FHD-YY1 binding curves using fluorescence anisotropy. 123
Figure 39: Supershift assay to determine the effects of the FOXP2 FHD on the YY1-AAV P5 DNA interaction. 126
Figure 40: Supershift assay showing the influence of YY1 on the FHD-CNTNAP2 DNA interaction. 128
Figure 41: Summary of YY1 oligomerisation and DNA binding behaviour. 138
Figure 42: Model of how YY1 might oligomerise and how the presence of DNA may make it dissociate. 140
Figure 43: Model of how the FOXP2 FHD binds to YY1 in the presence and absence of DNA. 143

LIST OF TABLES

Table 1: The YY1 and FOXP2 constructs used in this study 41
Table 2: SDS PAGE composition for glycine gels and tricine gels. 46
Table 3: Buffers used during glycine and tricine methods of SDS PAGE. 47
Table 4: Buffers used during preparative chromatography 50
Table 5: EMSA gel composition table. 53
Table 6: DNA oligonucleotide sequences used and the proteins to which they are specific. 57
Table 7: Concentrations of proteins and DNA used in a supershift gel to observe how FHD affects the specific YY1-DNA interaction. 59
Table 8: Concentrations of proteins and DNA used in a supershift gel to observe how YY1 affects the specific FHD-DNA interaction. 60
Table 9: Samples of YY1 used in SE-HPLC for elution profiles at 300 and 500 mM NaCl 71
Table 10: Parameters used during Mass spectroscopic analysis of apo and DNA-bound YY1 76

LIST OF EQUATIONS

Equation 1: $A = \varepsilon cl$ 55
Equation 2: $\varepsilon 280 = (NTyr * 1490) + (NTrp * 5500)$ 56
Equation 3: $Aprotein = A280 - (A554/10)$ 61
Equation 4: $r = I \parallel -GI \perp I \parallel +2GI \perp$ 64
Equation 5: $MRE\ [\theta] = \theta/10Cnl$ 67
Equation 6: $kHX = kop$ 73
Equation 7: $kHX = kopkcl \times kch = Kop \times kch$ 73
Equation 8: $D = Mt - M0\%DM100\%D - M0\%D \times N$ 77

LIST OF ABBREVIATIONS

4MBT domain	Four Malignant Brain Tumour Domain
A280	Absorbance at 280 nm
AAV	Adeno-Associated virus
Adenovirus E1A	Adenovirus Early Region 1A
ADHD	Attention-Deficit/Hyperactivity Disorder
ASD	Autism Spectrum Disorder
ATPase	Adenosine Triphosphatase
BRET	Bioluminescence Resonance Energy Transfer
C2H2	Cys2-His2
CBP	CREB-Binding Protein
CD	Circular Dichroism
CF1	Common Factor 1
DBD	DNA Binding Domain
DMSO	Dimethyl Sulfoxide
DNA	Deoxyribonucleic Acid
dsDNA	Double Stranded Deoxyribonucleic Acid
DTT	Dithiothreitol
E1A	Adenovirus Early Region 1A
EMSA	Electrophoretic Mobility Shift Assay
Far UV CD	Far Ultraviolet Visible Circular Dichroism
FEN1	Flap Endonuclease 1
FHD	Forkhead Domain
FOX	Forkhead Box
FOXP	Forkhead Box P
FOXP1	Forkhead Box P1
FOXP2	Forkhead Box P2
FOXP3	Forkhead Box P3

FOXP4	Forkhead Box P4
FPLC	Fast Performance Liquid Chromatography
G Factor	Grating Factor
HCFC1	Host Cell Factor C1
HDAC	Histone Deacetylase
HDAC2	Histone Deacetylase 2
Hdm2	Double Minute 2 Protein
HDXMS	Hydrogen Deuterium Exchange Mass Spectrometry
HEPES	4-(2-Hydroxyethyl)-1-Piperazineethanesulfonic Acid
His tag	Hexahistidine Tag
HIV-1	Human Immunodeficiency Virus 1
HMW	High Molecular Weight
Hox	Homeobox
HPLC	High Performance Liquid Chromatography
IMAC	Immobilised Metal Affinity Chromatography
INO80	Inositol-Requiring Protein 80
IPTG	Isopropyl β–D–1–Thiogalactopyranoside
ITC	Isothermal Titration Calorimetry
*k*ch	Rate of Exchange Between a Peptide Backbone Hydrogen and Deuterium
*k*cl	Rate of Protein Conformational Closing
K_D	Dissociation Constant
*k*HX	Overall Rate of Deuterium Exchange
k_{op}	Rate of Protein Conformational Opening
LacI	Beta-Galactosidase Large Subunit
LB	Lysogeny Broth
MBTD1	Malignant Brain Tumour Domain Containing Protein 1
MBTD1	Mbt Domain Containing 1
mRNA	Messenger Ribonucleic Acid
MRE	Mean Residue Ellipticity
MW	Molecular Weight
NTA	Nitrilotriacetic Acid

NuRD	Nucleosome Remodeling and Deacetylase
OD	Optical Density
OD_{600}	Optical Density at 600 nm
Peg3	Paternally Expressed 3
Pho	Protein Pleiohomeotic
Phol	Pleiohomeotic Like
PIAS	Protein Inhibitor of Activated STAT
PRC1	Protein Regulator of Cytokinesis 1
PRC2	Protein Regulator of Cytokinesis 2
PTM	Post Translational Modification
REPO	Recruitment of Polycomb
RNA	Ribonucleic Acid
rpm	Rotations Per Minute
RuvBL1	RuvB-Like 1
RuvBL2	RuvB-Like 2
SATB1	Special AT-Rich Sequence-Binding Protein 1
SDS PAGE	Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis
SDS	Sodium Dodecyl Sulfate
SE	Size Exclusion
SEC	Size Exclusion Chromatography
SOX2	SRY-Box Transcription Factor 2
SPR	Surface Plasmon Resonance
ssDNA	Single Stranded Deoxyribonucleic Acid
Ssn6	General Transcriptional Corepressor ssn6
SUMO	Small Ubiquitin-like Modifier
TBE buffer	Tris-Borate-EDTA Buffer
T-Box Pal	T-Box Palindrome
TBP	TATA-Box Binding Protein
TBR1	T-Box Brain Transcription Factor 1
TCEP	Tris (2–carboxyethyl) Phosphine
TCP	Tetracysteine Peptide
TCP	Tetracysteine Peptide

TEG	Triethylene Glycol
TEMED	N,N,N′,N′–Tetramethyl Ethylenediamine
TFIIB	Transcription Initiation Factor IIB
TUP1	General Transcriptional Corepressor TUP1
UCRBP	UCR-Motif DNA-Binding Protein
WT	Wild Type
WT-YY1	Wild Type Ying Yang 1
Xist	X-Inactive Specific Transcript
Yaf2	YY1 Associated Factor 2
YT	Yeast Tryptone Broth
YY1	Yin Yang 1
YY2	Yin Yang 2

INTRODUCTION

1.1 TRANSCRIPTION

Both development and homeostasis are complex processes which demand precise regulation of gene expression in order for a specific phenotype to be achieved. As transcription factors are responsible for activating, repressing and inducing specific genes, they are responsible for controlling gene expression. A comprehensive understanding of how this is achieved is vital for understanding processes in the body, such as development and homeostasis, which both require specific genes to be expressed for the correct phenotype Shi *et al.* (1991) . There are many mechanisms by which transcription factors can act or be regulated, the combination of which allows for detailed control of expression both spatially and temporally, which is especially important during development and cell differentiation (Lalmansingh *et al.*, 2012).

Many transcription factors function by binding DNA, either to induce transcription directly, or to activate or repress transcription (Li *et al.*, 2004). These functions may be achieved either by the transcription factor preventing another factor from binding to DNA, or by recruiting other transcription factors to form complexes (Belak and Ovsenek, 2007; Lalmansingh *et al.*, 2012). When bound to DNA, a transcriptional repressor may be blocking a binding site for a transcriptional activator, preventing activation by preventing the activator's binding (Gordon *et al.*, 2006). A transcriptional activator bound to DNA may also exclude a repressor from binding, and may be able to activate transcription directly (Gordon *et al.*, 2006).

Frequently, therefore, the ability to either activate or repress a gene is not reliant on the interaction between only one transcription factor and its specific DNA (Li *et al.*, 2004). In both transcriptional activation and repression, other transcription factors are often recruited either after DNA binding has occurred, or independently of DNA binding (Estruch *et al.*, 2018). Such interactions can involve varying numbers of proteins. In some cases, only a few transcription factors may be required to initiate transcription while in other cases, large complexes are needed (Alfieri *et al.*, 2013; Estruch *et al.*, 2018). These complexes can affect transcription through several mechanisms, including recruitment of general transcription factors to the promoter of a gene; or restructuring of chromatin to make it either more or less accessible to

other transcription factors (Wilkinson *et al.*, 2006). Alternatively, the interactions between the components in a complex can result in the activation of one of the member's transcriptional activity (Gordon *et al.*, 2006). This can occur through post translational modifications (PTMs) or through structural changes caused by the interaction. Recruited transcription factors may also be directly responsible for either repression or activation of transcription (Gordon *et al.*, 2006).

Besides their DNA binding abilities, therefore, the protein-protein interactions between transcription factors can be important for their ability to correctly regulate gene expression. While these protein-protein interactions can serve the purpose of recruitment to DNA, they can also be independent of DNA binding – this includes proteins involved in chromatin remodelling (Yue *et al.*, 2009). The combination of protein-protein interactions and protein-DNA interactions results in an intricate system whose purpose is to control gene expression (Alfieri *et al.*, 2013). Therefore, when attempting to understand how a transcription factor regulates a pathway, it is important to understand not only how it interacts with DNA, but also how it interacts with other proteins in the nucleus.

1.1.1. THE BASIC TRANSCRIPTION MODEL

While the activity of a transcription factor is tightly controlled by numerous potential interactions, certain requirements generally exist for the overall process. Firstly, the chromatin containing the gene sequence to be transcribed must be available for the transcription machinery to bind (Li *et al.*, 2007). This is not possible when it is in the form of heterochromatin, where DNA is tightly packaged around histone proteins, creating nucleosomes (Knezetic and Luse, 1986). Histones are proteins that DNA is looped around, which are usually assembled into octamers (Luger *et al.*, 1997). Additional linker histones also help to hold the DNA in place at the beginnings and ends of the loops (Allan *et al.*, 1986). This facilitates large amounts of DNA being packed into a small space.

For the DNA to become accessible to transcription factors, it needs to be converted into euchromatin, which is more loosely structured DNA. This is accomplished by modifying the histone complexes around which DNA is wound with PTMs (Li *et al.*, 2007). For example, acetylation is often used to alter the charge of the positive lysine residues, thus decreasing the

affinity between the protein and DNA (Dion *et al.*, 2005). This process alone is complex, with many different PTMs available to induce different effects (Li *et al.*, 2007).

Once the DNA is physically accessible to the transcription machinery, certain other events need to take place. Genes are associated with a promoter sequence upstream of them, which facilitates the binding of RNA polymerase and other general transcription factors necessary for transcription (Luse, 2014). These elements generally include the downstream promoter element, transcription initiation site, the TATA box site about thirty bases upstream of this, and the proximal promoter region which is usually within several hundred bases of the start site (Haberle and Stark, 2018). Additionally, distal regulatory elements such as enhancers or silencers can be found much further away from the start site (up to tens of kilobases) (Haberle and Stark, 2018).

The regulatory elements - both distal and proximal - are the most relevant sites of DNA for control of gene expression in a spacio-temporal manner. The transcription factors which bind to these are responsible for activation and repression of expression (Luse, 2014). When binding to a distal enhancer, transcription factor complexes loop the DNA around so that the distal elements are next to the promoter (Weintraub *et al.*, 2017). This facilitates binding of the necessary complexes to initiate transcription. Alternatively, repression can occur via a silencer site, or by means of a repressor transcription factor blocking activators from binding to the necessary elements (Luse, 2014).

Besides protein-DNA and protein-protein interactions causing activation or repression of transcription, the messenger ribonucleic acid (mRNA) produced has been suggested to play a role in this dynamic (Weintraub *et al.*, 2017). Several transcription factors can bind either DNA or RNA, either independently or simultaneously (Sigova *et al.*, 2015). This can facilitate the appropriate localisation of transcription factors to the promoter region, and has also in some cases been found to stabilise transcription factor interactions, such as homo-oligomerisation (Weintraub *et al.*, 2017).

1.1.2. TRANSCRIPTION FACTOR STRUCTURAL FEATURES

Because of the complex network of interactions involved in transcription, the molecules involved need to be structured in a way which can facilitate this. Not all proteins related to

transcriptional regulation have the ability to bind to DNA. Most transcription factors, however, contain a distinct DNA binding domain, which is frequently well conserved within families. Some of these include the helix turn helix motif, the zinc finger motif (of which there are several), beta sheets and leucine zippers (Alberts *et al.*, 2002). These motifs are often further specialised by the flanking sequences, forming structurally defined DNA binding motifs, many of which can also be involved in oligomerisation, RNA binding, and heterologous protein-protein interactions (Stroud *et al.*, 2006; Wai *et al.*, 2016; Estruch *et al.*, 2018). These defined domains are often conserved within transcription factor families.

Besides the DNA binding domains, many transcription factors contain additional, very long sequences involved in other interactions. These can form defined domains or motifs and can fold into a globular structure. However, many other transcription factors contain either short or very long regions of intrinsic disorder (Uversky and Dunker, 2010; Górecki *et al.*, 2015).

Intrinsically disordered proteins, or regions, are defined as having very little secondary or tertiary structure (Uversky and Dunker, 2010). They can occur as molten globules or pre-molten globules, which retain some secondary structure; or they can form pure coils, which are frequently charged (Uversky and Dunker, 2010). This occurs in protein sequences which often contain many charged amino acids and few hydrophobic amino acids, making it difficult for any hydrophobic interior to be formed (Uversky *et al.*, 2000). Proteins containing these regions are frequently involved in the regulation of processes such as transcription, and rely on their relative disorder to accomplish this (Oldfield and Dunker, 2014).

Within intrinsically disordered regions of proteins, certain sections of the sequences have been found to specifically interact with proteins and, to a lesser extent, RNA (Holmes, 1983). These sites are referred to as molecular recognition features, and upon binding to a macromolecular partner have been observed to gain structure where they previously lacked it (Mohan *et al.*, 2006). One molecular recognition feature might recognise more than one protein, which might induce a different structural formation in the intrinsically disordered region (Uversky and Dunker, 2010). These molecular recognition feature structures have been observed for regions of proteins which are usually disordered in isolation when their peptides (often around 25 amino acids) have been crystallised in complex with an entire protein or domain (Uversky and Dunker, 2010).

This ability to bind to other proteins with induced structural changes leads to relatively high specificity and relatively low affinity interactions (Oldfield and Dunker, 2014). These have

been postulated to be indispensable for tightly regulated pathways – a lower affinity might help to stop a process quickly, allowing for fine-tuning (Dunker *et al.*, 2005). They may also help to assist in localisation of specific factors (Oldfield and Dunker, 2014).

An intrinsically disordered protein usually has at least several molecular recognition features, which can be for the same or for different partners. As such, they can act as recruitment centres for large complexes, which are often necessary for transcription, where multiple factors can be recruited at once by one intrinsically disordered protein (Dunker *et al.*, 2005). The relative lack of disorder also introduces alternative splicing as a cell specific control mechanism, where certain molecular recognition features can be left out of a sequence, which changes the interacting partners of the transcription factor (Colak *et al.*, 2013). Lastly, intrinsically disordered proteins are strongly associated with PTMs, which add another dimension to their flexibility of function (Hsu *et al.*, 2013). These features of disordered regions of many transcription factors, therefore, play a very important role in regulating transcription, by facilitating the necessary macromolecular interactions.

1.1.3. MEMBRANE-LESS ORGANELLES

The nucleus does not consist of a homogeneous mixture of molecules. Within it there are particular regions of high concentrations of macromolecules which consist mostly of proteins, as well as nucleic acids. These regions are often referred to as proteinaceous membrane-less organelles, have their own specific functions, and contain specific members (Uversky, 2019). Their formation occurs via liquid-liquid phase transitions, resulting in distinct bodies of liquid-like heterogeneous proteins, RNA and sometimes DNA. Some examples of these include nucleoi (Shav-Tal *et al.*), which are well studied, chromatin, nuclear speckles and polycomb bodies (Uversky, 2019).

Nuclear speckles are proteinaceous membrane-less organelles which have been associated with localisation of certain transcription factors for the recruitment of RNA processing (Lamond and Spector, 2003). Polycomb bodies are associated with the polycomb group of proteins and are involved in transcriptional repression, including silencing of Homeobox (*Hox)* genes and X-inactivation (Pirrotta and Li, 2012; Brockdorff, 2017). Other proteinaceous membrane-less organelles are associated with different functions.

The formation of proteinaceous membrane-less organelles is a dynamic and also reversible process, where these functional bodies change and can even dissipate over time (Lamond and Spector, 2003). This is because liquid-liquid phase transitions are also reversible, and are sensitive to many conditions. Liquid-liquid phase transitions are driven by factors in the environment and also by the properties of their protein members. Changes in salinity, temperature, osmolarity, PTMs, the concentrations of the molecules involved and the binding partners present all have an effect on liquid-liquid phase transitions (Uversky, 2019). Additionally, the intrinsic properties of the proteins involved are key factors for the formation of proteinaceous membrane-less organelles (Uversky, 2019).

Most of the proteins which help to drive liquid-liquid phase transitions are either fully or partially intrinsically disordered (Uversky, 2019). This is because liquid-liquid phase transitions require the unique properties belonging to this group of proteins, which provide high levels of heterogeneity and are far more dynamic than the more rigid behaviour of globular proteins (Banani *et al.*, 2017). More specifically, many of these proteins are capable of binding to RNA, contain low complexity domains, have repeating charged and polar amino acids, and form low affinity multivalent interactions which are relatively non-specific (Uversky, 2019). The dynamic nature resulting from many of these properties allows these highly concentrated organelles to exist in liquid form, and, therefore, for them to remain functional and easily regulated.

Many proteinaceous membrane-less organelles are present in the nucleus and are vital for distinct transcriptional regulatory processes. The formation and dissociation of these organelles drives many processes in transcriptional regulation, including transcriptional repression and RNA assembly. As such, protein members responsible for driving the formation of these organelles are especially important to understand.

1.2. FOX FAMILY

The members of the forkhead box (FOX) family of transcription factors are characterised by a conserved forkhead box domain (FHD) (Wang *et al.*, 2003). This domain is a DNA binding domain of approximately 100 amino acids (Wang *et al.*, 2003). It is typically characterised by three alpha helices, two loop regions and three beta strands (Stroud *et al.*, 2006). This family

of transcription factors has a wide range of regulatory properties, which include developmental processes such as organogenesis, longevity, immunity, and DNA repair (Katoh *et al.*, 2013).

1.2.1. FOXP SUBFAMILY

Within the FOX family is the FOXP subfamily, which is unique within the family in that the forkhead domain (FHD) is located at the C-terminal end of the protein (Shu *et al.*, 2001). Furthermore, members of this family contain additional leucine zipper and zinc finger domains, giving them the ability to dimerise specifically (Wang *et al.*, 2003). The FOXP subfamily is involved in neural development, embryonic development, immunity, and homeostasis. There are four members of this family, namely FOXP1, FOXP2, FOXP3 and FOXP4 (Lalmansingh *et al.*, 2012). FOXP3 shares the least amount of homology within this group, and its expression in tissues overlaps less than the other three do.

1.3. FOXP2

1.3.3. FUNCTION

FOXP2 is a widely expressed member of the FOXP subfamily. As such, it is involved in a number of processes, including development of the lung, pancreas, thyroid, smooth muscle tissue and brain (Bowers and Konopka, 2012; Sin *et al.*, 2014; Li *et al.*, 2015). Given that FOXP2 is strongly associated with language development, where dysfunctional variants have been found to cause verbal dyspraxia (Katoh *et al.*, 2013), its role in neural development is of particular interest. Additionally, there are some implications of FOXP2 variants being associated with autism spectrum disorder (ASD), although studies concerning this have given conflicting results (Bowers and Konopka, 2012). Besides its medical relevance, however, the involvement of FOXP2 in language is by itself of importance. While language is one of the characteristics that differentiates humans from other animals, its development is a poorly understood process (Scharff and Haesler, 2005). Therefore, understanding how FOXP2 influences this process in the brain could give insight into this.

Specifically, FOXP2 functions by both binding to DNA and to other proteins (Stroud *et al.*, 2006; Estruch *et al.*, 2018). It acts as either a transcriptional repressor or activator when bound to DNA, and its binding affinity and repressive activity are regulated by its interaction with other proteins (Lalmansingh *et al.*, 2012). These interactions include those with FOXP2 itself, where it can homodimerise, as well as make heterologous interactions, including those with FOXP1 and 4, or alternatively with other transcription factors (Li *et al.*, 2004). Depending on which other transcription factors are present in the cell, therefore, FOXP2's function can be altered, making protein-protein interactions a vital part of its capabilities.

1.3.4. STRUCTURE

The ability of FOXP2 to bind either to DNA or to other proteins to regulate transcription is dependent on its structure. FOXP2 is a relatively large protein, consisting of 715 amino acids. At the N-terminal end it contains a polyglutamine region, followed further downstream by a zinc finger motif adjacent to a leucine zipper region (***Figure 1***) (Lalmansingh *et al.*, 2012). The forkhead domain is at the C-terminal end (***Figure 1***). The C-terminal region has been implicated in DNA binding, while the N-terminal motifs are associated with protein-protein interactions as well as transcriptional repression (Li *et al.*, 2004).

The function of the polyglutamine region has not been confirmed. However, it is suspected to be involved in the regulation of transcriptional activity (Shu *et al.*, 2001). This repeat is relatively common in several transcription factors, and while relatively little is known about its role and how it affects protein function, it has been found to be involved in protein-protein interactions (Gemayel *et al.*, 2015). For example, variations in the length of the polyglutamine region of Ssn6, another transcription factor, can result in a lack of interaction with TUP1, thereby disrupting its ability to regulate transcription (Gemayel *et al.*, 2015). While the actual mechanism by which the polyglutamine region is involved in this protein-protein interaction is unknown, it has been suggested that variations in the length of this region could cause changes in protein solubility (Gemayel *et al.*, 2015). It is possible that the polyglutamine region in FOXP2 has a similar function and also achieves this by mediating protein-protein interactions. Evidence to support this possibility is that FOXP2 interacts with PIAS proteins between residues 1 and 256 of FOXP2, which lie within the polyglutamine region (Estruch *et al.*, 2016). In other proteins, polyglutamine regions have also been associated with intrinsic disorder

(Crick *et al.*, 2006; Vitalis *et al.*, 2007). This would fit with the protein-protein interactions associated with it.

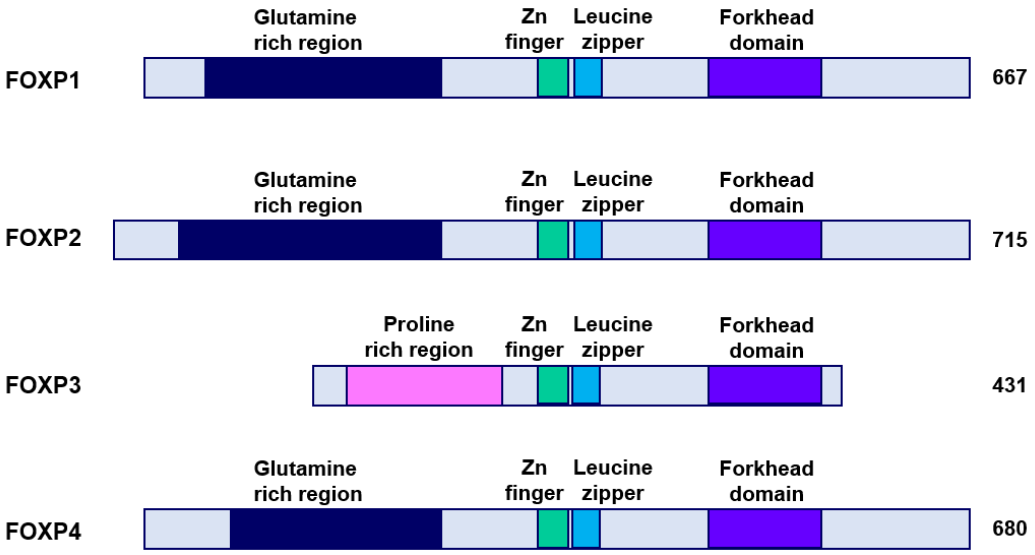


Figure 1: Domain composition of FOXP subfamily members.
Regions are indicated into boxes. Those not labelled are predicted disordered regions. The light unlabelled regions are the spacer regions, which are relatively unstructured, flexible regions of FOXP2. The navy region comprises the polyglutamine domain, or the glutamine rich region; the zinc finger is shown in green while next to it is the leucine zipper in a light blue. Pink represents a proline rich region. Dark purple represents the FHD. All four members of the FOXP subfamily contain the forkhead domain, zinc finger and leucine zipper regions. The numbers shown to the right indicate the total number of amino acids in each protein's sequence.

While they are not directly involved in DNA binding, both the leucine zipper and, to a lesser extent, the zinc finger motif are important for transcriptional repression through their ability to form protein-protein interactions (Li *et al.*, 2004). When the leucine zipper and zinc finger motifs are both removed from FOXP2, its capacity for transcriptional regulation is affected. This is thought to be a result of the elimination of the dimerisation for which both motifs are responsible (Li *et al.*, 2004).

In general, both types of motifs can be involved in either DNA binding or in dimerisation. An early understanding of Zinc finger motifs classified them as typically containing a repeating sequence of approximately 30 residues, which repeats nine times (Struhl, 1989). This folds into a structure resembling a finger, which can be capable of binding zinc or other forms of metals. Alternatively, zinc finger motifs have been found to form salt bridges (Laity *et al.*, 2001).

However, a much wider range of zinc finger types have been discovered, with different looking structures and different sequences (Cassandri *et al.*, 2017). FOXP2 contains a CCHC type zinc finger, which is associated more strongly with protein dimerisation than with DNA binding (Li *et al.*, 2004). As transcriptional repression is relatively unaffected by deletion of the zinc finger, it has been suggested that this motif may not be necessary for dimersiation to occur, but may rather be involved in the specificity of dimerisation through the leucine zipper (Wang *et al.*, 2003). Therefore, the zinc finger may be necessary for FOXP2 to bind the correct partners, but its actual binding ability is not dependent on this motif.

Like zinc fingers, leucine zippers consist of a characteristic repeat of residues, where leucine is typically the fourth residue in a seven amino acid repeat. When dimerising, the C-terminal of leucine zipper motifs forms a coiled coil (Ellenberger, 1994). The FOXP2 leucine zipper has been found to be important for dimerisation – more so than the zinc finger motif directly upstream (Li *et al.*, 2004). Additionally, the leucine zipper has been found to be important in transcriptional repression (Li *et al.*, 2004). It is likely that this function is conferred by the interaction between the leucine zipper and other proteins, where the interacting partner is required for transcriptional regulation to effectively occur. This domain of FOXP proteins appears to not only be important for self-association but also for heterologous interactions (Wang *et al.*, 2003; Estruch *et al.*, 2018).

1.3.5. FORKHEAD DOMAIN

The forkhead domain (FHD) of FOXP2 is its most well studied domain, with a crystal structure available of it bound to DNA (***Figure 2***) (Stroud *et al.*, 2006). The FHD is responsible for DNA binding, but has also been found to be involved in protein-protein interactions (Estruch *et al.*, 2018). Evidence for these protein-protein interactions includes both studies of the isolated FHD (Wu *et al.*, 2006) and studies which investigated the full length FOXP2 (Estruch *et al.*, 2018).

A typical FHD of the FOX family is characterised by its three alpha helices, three antiparallel beta strands and two loop regions, which are referred to as wings. These central components of the domain are conserved across all FOX family members. Helix three inserts into the major groove of DNA, interacting with the bases via hydrogen bonds and van der Waals forces

(Stroud *et al.*, 2006). Furthermore, two wings are present in the FHD – wing 1 and wing 2 – which interact with DNA extensively. Because of these components, the FHD is also known as the winged helix domain (Stroud *et al.*, 2006). The FOXP2 FHD also folds into three α helices, referred to as H1, H2, and H3, which are followed by the three stranded (S1, S2 and S3) antiparallel β sheet (***Figure* 2**). It does, however, deviate from the basic blueprint of a FOX FHD where a 3_{10} type helix (H4) occurs in the turn between H2 and H3 (Stroud *et al.*, 2006). Furthermore, both of the wings differ from most FOX protein FHDs in that they are truncated, with wing 1 becoming a simple turn joining strands S2 and S3, while wing 2 forms a helix (H5) (Stroud *et al.*, 2006). Given the relative variability between wings in FOX proteins, these regions may have differing functions in each subfamily. In the case of FOXP2, the truncated wings result in lessened DNA contacts, which may decrease the affinity of the FOXP proteins for DNA compared to other family members (Stroud *et al.*, 2006; Webb *et al.*, 2017).

While the quaternary structure of the full length FOXP2 has not been fully investigated, the isolated FHD has been studied and it is capable of existing both as a monomer and as a dimer in solution (Stroud *et al.*, 2006). When bound to DNA, it can occur in monomeric, dimeric and higher order forms, depending on the concentration of FOXP2 (Morris and Fanucchi, 2016). The crystal structure for FOXP2 shows it bound to DNA both as a monomer and as a domain swapped dimer, where the two monomers exchange H3, S2 and S3 (***Figure* 2**) (Stroud *et al.*, 2006). This ability of the FHD to dimerise is unique to FOXP proteins, and is likely due to a conserved proline having changed to an alanine in the FOXP FHD at residue 539, which lies in the loop between H2 and H3 (Stroud *et al.*, 2006). This allows for H3 to be extended into a more flexible hinge loop, which facilitates this structural change, allowing H3 to move (Medina *et al.*, 2016). As with the crystal structure, the FOXP2 FHD in solution has been shown to occur as both a monomer and a dimer, where increasing concentrations – from about 300 μM – result in the increasing formation of dimer (Morris and Fanucchi, 2016).

It has been suggested that the purpose of the domain swapped dimer might be for the dimer of the protein to bind two sections of DNA at once (Stroud *et al.*, 2006). In this way, it could potentially loop the DNA, or alternatively it might allow for interchromosomal association if each subunit binds to a sequence belonging to two different chromosomes (Morris and Fanucchi, 2016). However, whether the FHD of FOXP2 has the propensity to form domain swapped dimers in the full length protein rather than simply as an isolated domain or whether or not dimerisation is even feasible at physiological concentrations has not been investigated. While this is possible, it does need to be confirmed, especially as the leucine zipper has been

identified as the most important interface for dimerisation when studying the full length protein (Li *et al.*, 2004).

Indeed, in longer constructs of FOXP2 (from the leucine zipper onwards), the leucine zipper seems to be the most important FOXP2 element to initiate oligomerisation (Thulo *et al.*, 2021). Furthermore, the acid rich tail of the protein also seems to have an influence on oligomerisation, possibly through an interaction with the FHD which could influence domain-swapped dimer formation (Thulo *et al.*, 2021). These longer constructs were observed to form a variety of oligomeric species, further indicating that FOXP2 domain swapped-dimers are at least not the only relevant oligomeric species, if they do indeed form physiologically.

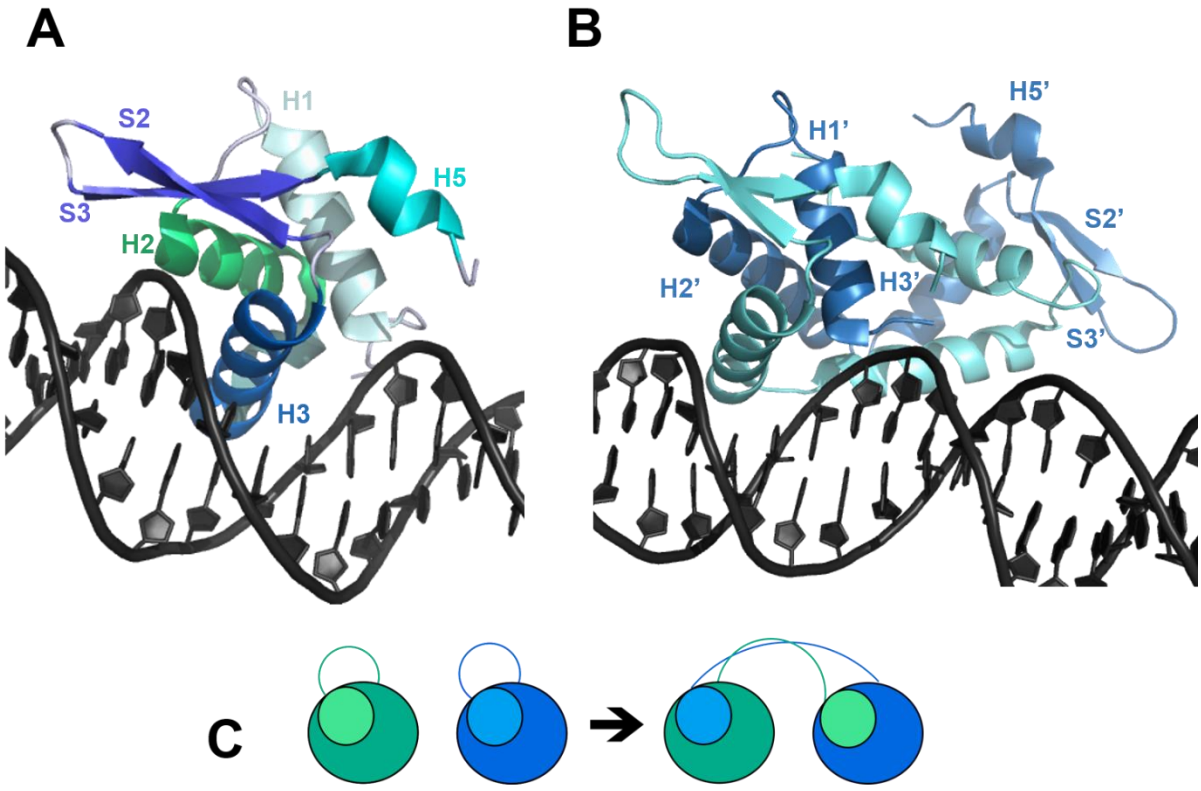


Figure 2: The structure of the FOXP2 FHD when bound to DNA.
A) the FOXP2 monomer bound to DNA, with helices labelled H and strands labelled S. **B)** the domain swapped dimer of FOXP2, where one monomer in cyan and the other in blue. The helices and strands of this second monomer are labelled H1', H2', H3', H5', S2' and S3' to distinguish them from the corresponding sections of the first monomer. The images were generated using structure 2a07 on PDB (Stroud *et al.*, 2006). The software used was the PyMOL molecular graphics system version 1.8, Schrödinger, LLC. **C)** Schematic of two monomers becoming a domain swapped dimer, where one domain is exchanged upon dimerisation. This image was created using Inkscape 1.0.

The relevance of the FHD in homodimerisation may still be relatively unknown, however, it has been shown to be involved in heterologous interactions. For example, SATB1 – another transcription factor involved in neural development - interacts with the FHD of FOXP2 (Estruch *et al.*, 2018).

1.3.6. DYNAMICS

A protein's dynamics are usually imperative for its ability to function. For example, an enzyme might need to change conformations after binding to its substrate in order to carry out its catalytic activity (Yang *et al.*, 2014). Many structural proteins, such as actin, function by moving across a surface, which is only possible if the domains of the protein itself are dynamic (Pollard *et al.*, 2000). Likewise, transcription factors require flexibility to function – this includes determining specificity to DNA binding (Chu *et al.*, 2011), alteration of DNA (Craggs *et al.*, 2014) and regulating repression and activation of gene expression (Morris *et al.*, 2018).

When in the absence of DNA, the isolated FOXP2 FHD is more flexible than when in the presence of DNA (Morris *et al.*, 2018). This difference is especially pronounced in helix 2, helix 3 (which is the main interface for DNA binding) and in the β-sheet wing region, all of which are more solvent exposed when in the absence of DNA (Morris *et al.*, 2018). This flexibility seems to be a trait of the FOXP subfamily, and it has been seen in the NMR structure for FOXP1 as well (Chu *et al.*, 2011). While the entire FOX family has been seen to be flexible in these regions of the FHD, they exhibit fluctuations in the microsecond timescale for FOXP1 as opposed to only the pico- to nanosecond timescale observed in other FOX proteins (Jeon and Lee, 2011). As these regions are involved in domain swapping, this extra flexibility in the FOXP proteins is likely to be responsible for enabling this function (Chu *et al.*, 2011). The apo form of isolated FOXP2 FHD occurs as a monomer, and not all studies of the DNA-bound form have observed a domain-swapped dimer (Stroud *et al.*, 2006; Morris and Fanucchi, 2016). Therefore, its ability to form a domain swapped dimer may be situationally specific, enabling it to perform specific functions under the correct conditions (Morris *et al.*, 2018).

When DNA binds FOXP2, there is a reduction in solvent exposure in the wing 1, H2 and H3 regions, which implies an increase in the structure of the backbone (Morris *et al.*, 2018). In part, the flexibility of the FHD is a result of the hinge loop region, where alanine 539 has replaced a proline across the FOXP2 subfamily, allowing for an increase in flexibility which is

also associated with its ability to form the domain swapped dimer (Stroud *et al.*, 2006). Replacing this alanine with the proline found in other FOX proteins was seen to decrease the flexibility of the FHD globally, with less difference between the apo and the DNA-bound forms (Morris *et al.*, 2018). This change had an effect on the DNA binding affinity of the FHD, as well as on its transcriptional activation activity. Given that the W1 region, the C-terminal region and the region between H2 and H3 have high levels of flexibility in many FOX proteins, and tend to have large amounts of variability within the family, it has been suggested that the dynamics in these regions are largely responsible for DNA binding specificity (Chu *et al.*, 2011; Morris *et al.*, 2018). Therefore, the dynamics of FOXP2 are likely essential for it to bind the correct DNA sequence for the appropriate time periods, and changes in its conformation to make a domain-swapped dimer under the right conditions may alter its function slightly, forming part of the regulatory network of gene expression.

1.3.7. DNA BINDING

FOXP2 is a versatile transcription factor in terms of its DNA binding properties. While it does bind DNA with specificity – where it recognises specific sequences of DNA and binds to these via specific interactions – FOXP2 has the ability to bind several unrelated sequences (Webb *et al.*, 2017). Its interaction with these sequences has varying affinities, suggesting that it may differentially regulate the different sequences it binds to *in vivo*. This could result in varying extents of transcriptional repression for different genes when FOXP2 is present at one concentration. As specificity and affinity of binding to DNA are determining factors for FOXP2's function, it is important that they are also understood (Webb *et al.*, 2017). Furthermore, the protein-protein interactions with which FOXP2 is involved are likely to alter its DNA binding behaviour.

The quaternary structure of FOXP2 does appear to have an influence on its ability to bind DNA. Compared to the wild type FHD, which binds to DNA with a K_D of 66.94 ± 4.34 nM, a monomeric mutant (A539P) has a decreased affinity for DNA, with a K_D of 102 ± 4.97 nM (Morris and Fanucchi, 2016). This suggests that the quaternary structure of FOXP2 could potentially regulate its repression or activation activity, as, depending on whether it is monomeric or dimeric, it may bind to DNA more tightly, which could affect its function. While it is possible that dimerisation influences DNA binding, the ability of the FOXP2 FHD to bind

DNA may not be influenced by its ability to dimerise, however. A FOXP1 mutant that does not bind DNA can still dimerise without major energy changes in the process involved (Medina *et al.*, 2016) and since the FOXP2 and FOXP1 FHDs are relatively conserved, it is likely that this is the same for FOXP2.

Besides a potential change in affinity, it is possible that homodimerisation of FOXP2 occurs at the FHD in order to bridge DNA, as the DNA binding residues in the domain swapped dimer are located on opposite ends of the protein (Stroud *et al.*, 2006). The protein dimers may allow simultaneous binding of DNA, potentially localising interchromosomal genes (Stroud *et al.*, 2006). Therefore, given the increased affinity for DNA of the FOXP2 FHD dimer compared to the monomer, as well as its potential in bridging DNA, its quaternary structure is potentially relevant to its transcriptional regulatory properties (Sin *et al.*, 2014).

1.3.8. PROTEIN-PROTEIN INTERACTIONS

The ability of FOXP2 to efficiently bind to and regulate the correct genes may be heavily influenced by its protein-protein interactions (Sin *et al.*, 2014). Indeed, protein-protein interactions between transcription factors have been found to assist with transcriptional regulation in a number of ways. These include localisation to a particular DNA sequence. One such example is how the transcription factor Pho is required to recruit the polycomb complex to the correct DNA sequence (Schuettengruber *et al.*, 2007). Another example is that many transcription factor interactions are involved in controlling the PTMs taking place in the nucleus. For example, for histone deacetylase (HDAC) (involved in chromatin remodelling and cell replication) to be activated, it needs to be acetylated, the process of which is associated with its interaction with YY1 (Gordon *et al.*, 2006). Specifically, YY1 has been found to associate with HDACs 1, 2 and 3 (Yang *et al.*, 1996; Yang *et al.*, 1997; Thomas and Seto, 1999; Wang *et al.*, 2008). Heterologous interactions may also result in a binding partner which is actively involved in gene regulation, such as recruitment of general transcription factors (Gordon *et al.*, 2006). The polycomb repressive complexes PRC1 and PRC2 are examples of this, where recruitment of both triggers transcriptional silencing through histone modification (Alfieri *et al.*, 2013).

FOXP2 is capable of both homo and hetero-oligomerisation, both of which are apparently able to influence its activity (Li *et al.*, 2004; Estruch *et al.*, 2016; Usui *et al.*, 2017). Binding to

either itself or other proteins may help FOXP2 bind DNA more tightly, or may alter its specificity for certain sequences of DNA, by altering the structure of either FOXP2 or DNA to increase the favourability of binding (Li *et al.*, 2004). Altered DNA binding could potentially result in a change of the targets themselves, or alternatively a change in the expression of a target. Furthermore, FOXP2 may alter other proteins, regulating transcription independently of its DNA binding ability.

1.3.9. HETEROLOGOUS INTERACTIONS

Besides having the ability to interact with itself, FOXP2 can interact with other proteins in complex regulatory networks. This includes interactions with two other FOXP proteins, FOXP1 and FOXP4, which share a high level of homology with FOXP2. FOXP1, FOXP2 and FOXP4 are all expressed in the brain, and expression frequently overlaps with each other both spatially and temporally (Li *et al.*, 2004). While their high level of homology does suggest some extent of functional redundancy, where they are capable of binding the same sequences, the exact pattern of combination with each other can lead to different results. For example, depending on the interactions between the different FOXP proteins, a gene may be activated rather than repressed, as occurs when only FOXP2 is present (Sin *et al.*, 2014). Alternatively, that gene may be repressed more strongly (Pignata and Romano, 2014). This enables detailed regulation of expression, which may be necessary for the development of the brain (Konopka *et al.*, 2009).

Outside of the FOX family, several binding partners of FOXP2 have been identified. These include a wide variety of proteins with various functions, reflecting the diverse involvement of FOXP2 in physiological pathways. For example, FOXP2 has been found to be a component of the NuRD complex- a complex of proteins involved in histone modification (Chokas *et al.*, 2010). The interaction of FOXP2 with other members of the NuRD complex, such as P66β, strongly suggests that FOXP2 is involved in altering chromatin accessibility (Chokas *et al.*, 2010). Other protein-protein interactions include those with proteins involved in PTMs: FOXP2 interacts both with PIAS proteins, which promote SUMOylation, and with SUMO proteins, which are directly responsible for SUMOylation of proteins (Estruch *et al.*, 2016). These interactions imply that FOXP2 is post-transcriptionally modified, and that its

modification may alter its binding to either DNA or to other proteins, changing its function to what is relevant at the time of SUMOylation (Estruch *et al.*, 2016).

Other partners of FOXP2 may act as corepressors, or may prevent it from binding to DNA – for most of these proteins, however, the actual mechanism of interaction is unknown (Estruch *et al.*, 2018). When the functions of FOXP2 and its binding partners are identified, however, the role of the interacting partner can provide information about the relevance of the interaction between FOXP2 and itself. Several of the proteins that bind to FOXP2, for example, TBR1, are involved in neural development, and, therefore, are likely to interact with FOXP2 to regulate neural development (Estruch *et al.*, 2018).

Yin Yang 1 (YY1) is one such binding partner. While the exact nature of the FOXP2-YY1 interaction is unknown, the two proteins have been shown to interact using a Bioluminescence Resonance Energy Transfer (BRET) assay. In this assay, both proteins were transfected into mammalian cells and one was fused with luciferase, while the other was fused to yellow fluorescent protein (Estruch *et al.*, 2018). Because of the fusion proteins, when FOXP2 and YY1 interacted with each other, their close proximity caused resonance energy transfer between the two fused proteins, resulting in a wavelength shift in emitted light (Estruch *et al.*, 2018). By transfecting mammalian cells with versions of FOXP2 with its C- or N-terminus truncated, the various regions of FOXP2 that interact with YY1 have been identified, because if a segment vital for the interaction was missing, the proteins would no longer interact (Estruch *et al.*, 2018). The YY1-FOXP2 interaction occurred across most areas of FOXP2, excluding the far N-terminal region (residues 1-58) (Estruch *et al.*, 2018). This suggests FOXP2 may possess multiple binding sites for YY1. While the BRET assay is a relatively reliable method to detect protein-protein interaction, it does not guarantee their ability to interact in isolation – other proteins might be required for the interaction to occur.

As both FOXP2 and YY1 are involved in neural development, it is possible that their interaction is necessary for the correct developmental processes to take place in the cell. In order to understand how the interaction of these two proteins may influence these processes, it is important to understand more about the manner of their interaction – for example, whether YY1 affects the DNA binding capabilities of FOXP2, and how strong the interaction between both proteins is. This would give insight into why the two proteins interact, as well as what pathways their binding might influence.

1.4. YY1

1.4.1. YY1 FAMILY

Yin Yang 1 (YY1), also known as NF-E1, UCRBP and CF1, belongs to the C2H2 family of zinc finger proteins (Gordon *et al.*, 2006). This protein is a transcription factor with the ability to both repress and activate transcription. This resulted in it being called Ying Yang 1, after its dual function. The C2H2 family consists of a wide variety of proteins containing zinc fingers whose zinc-binding residues consist of two cysteines followed by two histidines (Iuchi, 2001). Many of the members of this zinc finger family are transcription factors and are commonly found in the human brain (Ruppert *et al.*, 1988). Besides the number of zinc fingers distinguishing the members of the C2H2 family from each other, they can also be identified according to the exact spacing between the zinc binding residues (Iuchi, 2001). YY1 is a GLI-Krüppel related protein, meaning that it contains three zinc finger motifs with the GLI subgroup consensus sequence, and one zinc finger motif with the consensus sequence belonging to the Krüppel subgroup (Shi *et al.*, 1991).

YY1 has a few closely related transcription factors. YY2 and REX are homologues of YY1 which also occur in placental mammals and are the result of retrotransposition of YY1. Pho and Phol are orthologues of YY1 which are found in *Drosophila*, both of which are associated with the polycomb complex (Kim *et al.*, 2007). In addition to the zinc finger, YY1 does have some conserved regions within its N-terminal domain when compared to these family members. These N-terminal regions are conserved amongst YY1, YY2, REX1, Pho and Phol, which are relatively closely related with one another (Kim *et al.*, 2007).

1.4.2. YY1 FUNCTION

YY1 is a ubiquitously expressed transcription factor. It is heavily involved in many developmental processes, including X inactivation (Jeon and Lee, 2011), DNA repair (Gordon *et al.*, 2006), cell proliferation and differentiation (Gordon *et al.*, 2006; Deng *et al.*, 2010), smooth muscle development, neural development (Gabriele *et al.*, 2017), apoptosis (Deng *et al.*, 2010) and myoblast differentiation (Shi *et al.*, 1997). Its role in neural development in

particular overlaps with that of FOXP2, especially in its involvement in cell differentiation and proliferation. Like with FOXP2, mutations in YY1 are capable of leading to intellectual disorders (Gabriele *et al.*, 2017).

The lack of YY1 in both alleles results in lethality. Haploinsufficiency and pathogenic mutations can still result in viable embryos, but causes Gabriele-de Vries syndrome – a condition affecting cognitive abilities, delaying development, and resulting in multiple congenital abnormalities including in the brain, uterus, heart and skeleton (Nabais Sá *et al.*, 2019). Among the behavioural effects of this disorder are ASD and schizoaffective disorder. Less severe effects of altered YY1 include anxiety and attention-deficit/hyperactivity disorder (ADHD) (Nabais Sá *et al.*, 2019). This not only indicates a key role of YY1 in the brain, but also that it is highly important in other developmental pathways.

YY1 is important not only in development but also in adults. Further functions of YY1 in the brain later in life are associated with neuroprotective pathways. Downregulation of YY1 is one of the factors seen in Parkinson's disease (Pal *et al.*, 2016), while upregulation is a factor observed in Alzheimer's disease (Aubry *et al.*, 2015). This suggests a need for tight control of the protein, as its influence in expression pathways is clearly high.

At least one of YY1's roles is involved in the cell cycle. For example, YY1 is one of the transcription factors which regulate p53, an important tumour suppressor protein which is associated with apoptosis (Sui *et al.*, 2004). In glioma tumour cells, YY1 has been implicated as a suppressor of p53, resulting in a lack of induced cell death, and thereby enabling tumour cell proliferation (Gao *et al.*, 2018). The mutation T372R is located in the DNA binding region of YY1, and is associated with pancreatic neuroendocrine tumours. This mutation alters the binding specificity of YY1, resulting in an altered function influencing calcium channels (Cromer *et al.*, 2015).

However, YY1 does not always act as tumour stimulator. It can be a tumour suppressor depending on the context. For example, in pancreatic tumours, it inhibits SOX2, which is transcription factor involved in the proliferation of cancer stem cells (Zhang *et al.*, 2017). In breast cancer, YY1 represses FEN1, which sensitises the cancer cells to DNA damaging agents, therefore, supressing the tumour (Wang *et al.*, 2015). This contradictory nature of YY1 in cancer suggests a strong influence of interacting partners in its function. Changes in the interactome of YY1 might well be the cause of changes in its behaviour (Sarvagalla *et al.*, 2019).

Besides its role in cancer and neural diseases, YY1 is also implicated in the regulation of retroviral expression. It has been associated with repressing the expression of viral genes, preventing them from entering the lytic cycle. These include HIV-1, adeno-associated virus (AAV) and Epstein-Barr virus infections (Nieborak and Górecki, 2016).

In order to regulate the diverse number of processes in which it is involved, YY1 not only has multiple gene and protein targets, but also several ways in which it can control expression. When regulating transcription, it can act either as a repressor, an activator, or an initiator of transcription, depending on the conditions and on its interacting partners (Shi *et al.*, 1991; Sui *et al.*, 2004). The capacity of YY1 to both repress and to activate transcription can be achieved either when bound to DNA or when unbound to nucleic acid and instead bound to other proteins (Gordon *et al.*, 2006; Deng *et al.*, 2010). Depending on what it is interacting with, its regulatory role can change (Deng *et al.*, 2010).

1.4.3. YY1 REPRESSION

YY1 is responsible for the repression of several targets, the most well studied being viral promoters such as the AAV P5 promoter (Gordon *et al.*, 2006). It does, however, also repress many human genes, which include immunoglobulin (Shi *et al.*, 1997) and the chromatin remodeller HCFC1, which is also involved in neural disorders when altered (Gabriele *et al.*, 2017). There are currently several models of repression for YY1, which may all be valid (***Figure 3***) (Gordon *et al.*, 2006).

Firstly, the displacement model suggests that YY1 directly binds the same sites on promoters as other proteins, blocking activators from binding (Shi *et al.*, 1997; Gordon *et al.*). The second model suggests that YY1 interferes with the function of a transcriptional activator, either by binding the transcriptional activator, by masking general transcription factors from the activator, or potentially by bending DNA to prevent activator binding to the promoter (Shi *et al.*, 1997; Wilkinson *et al.*, 2010). Finally, YY1 may interact with a corepressor, which affects other factors, such as chromatin structure. This is supported by the finding that YY1 interacts with HDAC proteins, which deacetylase histones (Thomas and Seto, 1999; Wang *et al.*, 2008), altering the chromatin structure to make the DNA less accessible to other transcription factors.

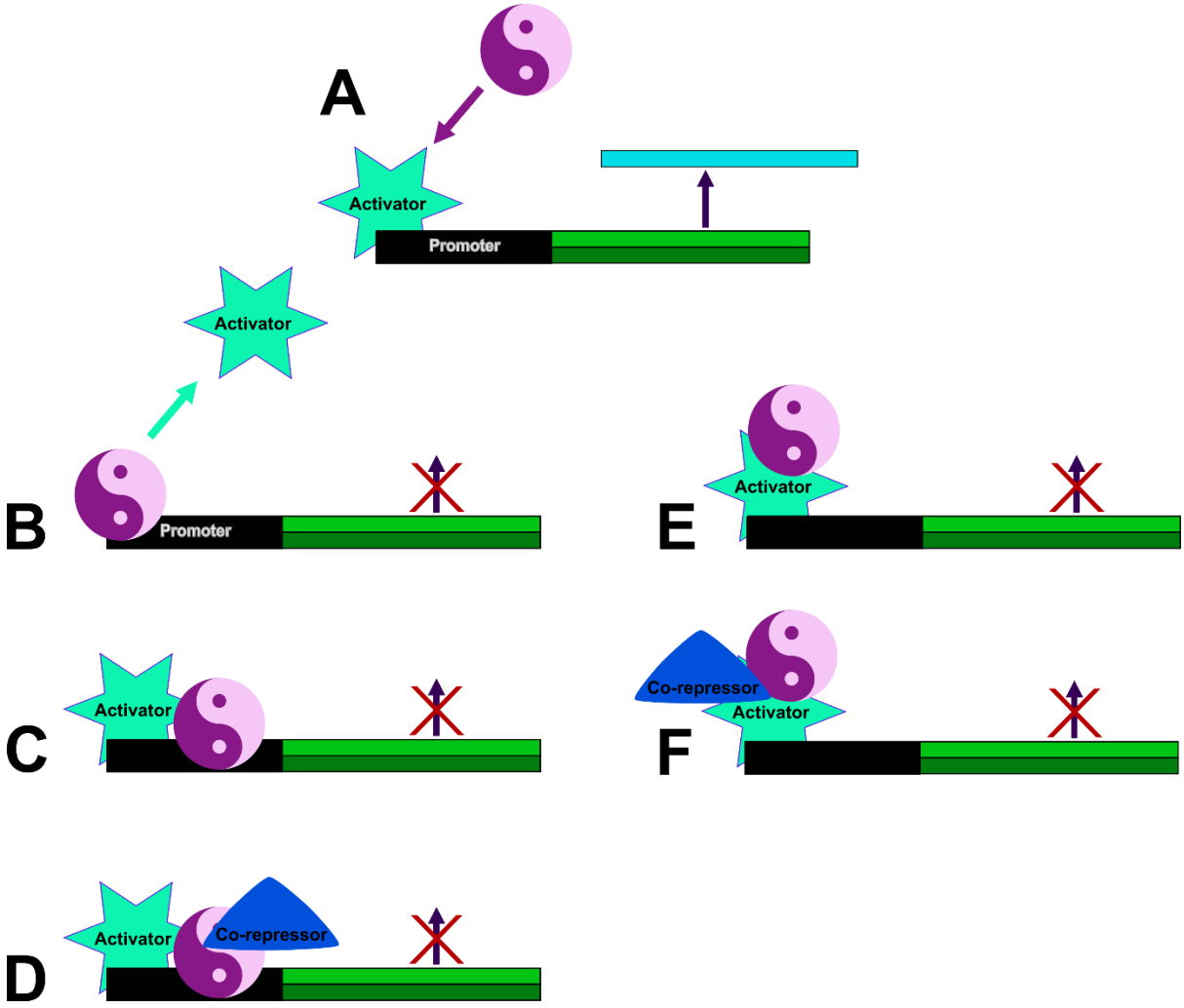


Figure 3: The models describing YY1 translational repression.
YY1 is represented by the purple yin yang sign. The green boxes are the DNA downstream of the promoter region and the blue is mRNA. **A)** Before YY1 interacts with the DNA or activator, gene expression occurs. **B)** The displacement model of YY1 repression, where YY1 binds the same promoter sequence. **B**, **C** and **E)** The model whereby YY1 interferes with the transcriptional activator, which has been proposed to occur in several different ways. YY1 may bind the activator and DNA together, just the activator, or may bind the activator as well as a co-repressor, interfering with the activator's function. **D** and **F)** The third model, where YY1 binds a co-repressor to mediate repression. These co-repressors may directly interfere with transcription or may mediate processes such as chromatin remodelling. This figure was created using Inkscape 1.0.

Two regions of YY1 are associated with repression (***Figure 4***). Most domain mapping studies have determined that one of the repression domains occurs in the zinc finger region (Thomas and Seto, 1999). However, these studies have given conflicting results concerning the exact residues involved within this region. The N-terminal repression region of YY1 lies between

amino acids 170 and 200 (Sui *et al.*, 2004). The two repression domains may work together, because deletion of either one produces the same effect (Thomas and Seto, 1999).

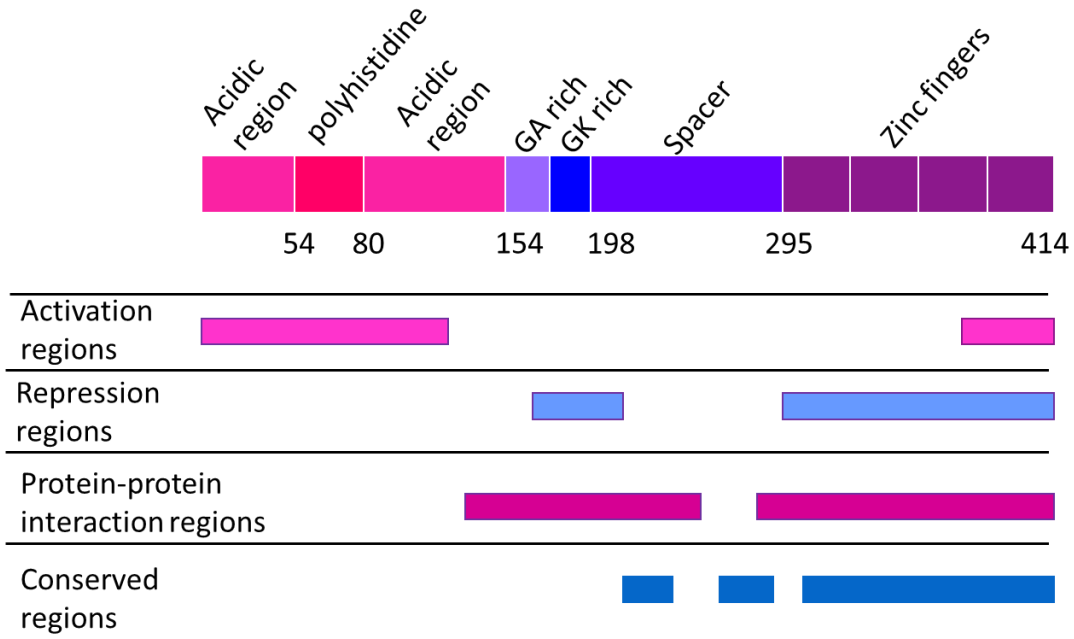


Figure 4: Sections of YY1's sequence and their associated functions.
The sections of YY1 are illustrated at the top. The bars below represent the functions (activation, repression and protein-protein interactions) associated with various segments of YY1. The protein-protein interaction domains indicate the most frequently found residues, and some interactions occur throughout the rest of the sequence. The conserved regions across YY1 proteins are also shown in order to compare function with conservation. This Image was made using PowerPoint (2013).

1.4.4. YY1 ACTIVATION

YY1 is also capable of activating genes depending on the relevant promoter and its interaction partners. While YY1 represses the AAV P5 promoter most of the time, in the presence of the adenovirus E1A oncoprotein it activates the viral promoter instead. YY1 is also involved in the activation of ribosomal proteins (Gabriele *et al.*, 2017). Therefore, it is capable of both activating and repressing the same gene, or only activating or repressing a specific gene. Like

with repression, the mechanism of inducing transcriptional activation has several models (***Figure 5***).

YY1 has been suggested to activate transcription directly by interacting with general transcription factors, stimulating transcription. Evidence for this model includes the presence of two acidic activation domains in YY1, as well as its ability to interact with general transcription factors such as TBP, which make up the transcriptional machinery necessary for transcription to occur (Houbaviy *et al.*, 1996).

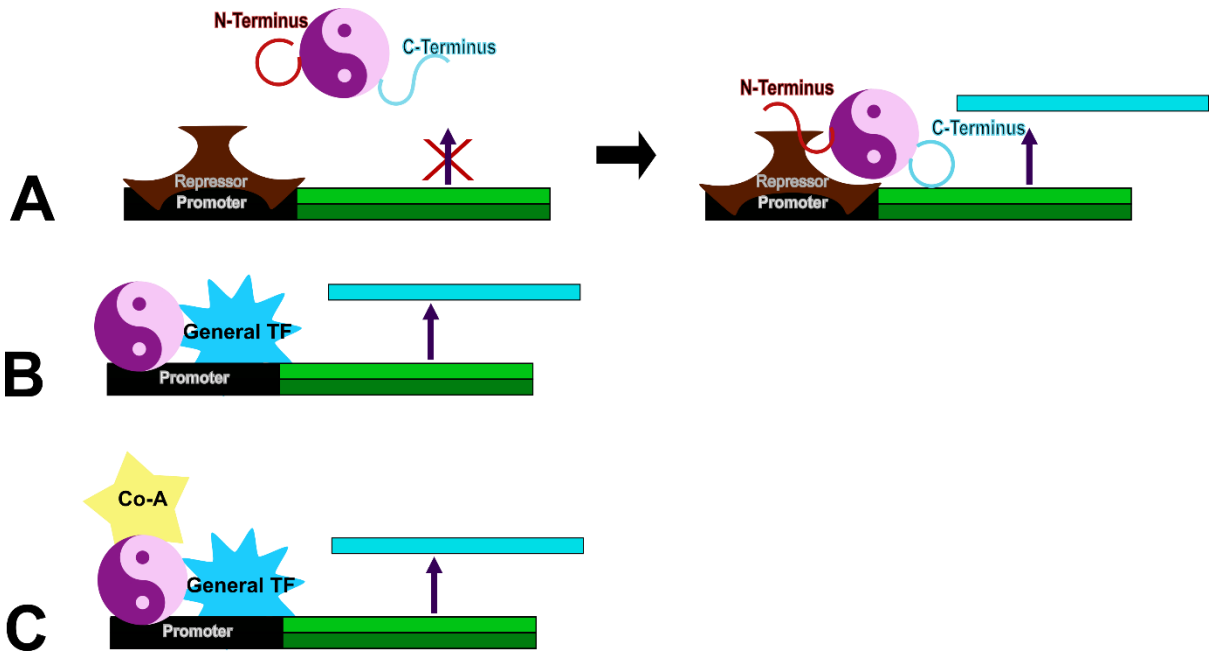


Figure 5: Transcriptional activation models for YY1.
A) The indirect activation model: YY1 undergoes a structural change which results in the exposure of an activation-relevant sequence. Activation occurs when the N-terminus (activation domain) becomes open in conformation and the C terminus (repression domain) becomes closed. **B)** The direct model of activation by YY1, where YY1 directly stimulates transcription. **C)** The model where YY1 interacts with coactivators after binding DNA, which function to activate transcription of the gene. Co-A stands for co-activator. This diagram was created using Inkscape 1.0.

The second model proposed for transcriptional activation suggests that YY1 interacts with cofactors, which may result in a structural alteration of YY1 (Gordon *et al.*, 2006). This could result in the unmasking of a section of the YY1 sequence which is responsible for activation via protein-protein interactions (Thomas and Seto, 1999). The third model also involves the

recruitment of cofactors, but in this case the interaction brings the cofactor to the target promoter to activate it, possibly by acetylating the DNA – this is implied by the interaction between YY1 and acetyltransferases such as CBP and p300 (Thomas and Seto, 1999; Gordon *et al.*, 2006).

Like with the repression domains of YY1, two activation domains have been identified – one in the N-terminal region, and the other in the C-terminal (Thomas and Seto, 1999). Again, the exact location of these is under dispute. The N-terminal activation domain has been concluded to be bipartite (Thomas and Seto, 1999), while most are in agreement that the C-terminal activation domain occurs at the extreme C-terminus of the protein (Gordon *et al.*, 2006).

1.4.5. YY1 INDUCTION

Besides its dual function to both activate and repress transcription, YY1 is also capable of acting as an initiator element. It has been found to initiate transcription when bound to TFIIB and the large subunit of RNA polymerase II *in vitro* (Shi *et al.*, 1997).

1.4.7. YY1 STRUCTURE

YY1 consists of 414 amino acids and has a predicted MW of 44 kDa. It contains an intrinsically disordered N-terminal region and a more structured zinc finger domain made up of four zinc fingers at the C-terminus (***Figure 4***). The structures for the C-terminus and a section near the middle of the protein have been determined using X-ray crystallography (amino acids 293-414 and 199-228) (***Figure 6***, ***Figure 7***) (Houbaviy *et al.*, 1996; Alfieri *et al.*, 2013), but the structure of the full length YY1 is relatively unknown, other than that it appears to display disordered tendencies, especially in the N-terminal region outside of the DNA binding domain (Górecki *et al.*, 2015). While it does not appear to have many structural motifs, there are distinct sections within YY1 that, when deleted, result in the inability of the protein to participate in intermolecular interactions, repression and activation activity (***Figure 4***) (Sui *et al.*, 2004). This is relatively typical of intrinsically disordered regions.

1.4.7.1. N-TERMINAL REGION

The N-terminal region of YY1 is intrinsically disordered and lacks homology with known protein domains (Kim *et al.*, 2007; Górecki *et al.*, 2015). It contains many sections of repeating residues and several protein binding sequences. This includes an acidic region between residues 43 and 53, which contains a stretch of glutamic and aspartic acid residues (***Figure 4***) (Deng *et al.*, 2010). These repeated residues may be important for interactions with positively charged proteins such as histones. Furthermore, YY1 contains a homopolymeric stretch of histidine residues at position 70-80, which is present in several other transcription factors (Shi *et al.*, 1997). While it has been suggested that this stretch may help to neutralise adjacent amino acids and play a role in assisting repression, this has not been proved (Shi *et al.*, 1997). It has, however, been shown that this stretch of histidines is important in localising YY1 to nuclear speckles (Salichs *et al.*, 2009), which are hubs at which regulatory processes for gene expression occur, such as preparation of mRNA (Galganski *et al.*, 2017). Further along the YY1 sequence, at amino acids 170-201, there is a high percentage of glycine and alanine residues (Shi *et al.*, 1997). This region has been shown to have transcriptional repression activity (Gordon *et al.*, 2006).

The conserved sections in the disordered N-terminal region of YY1 include the amino acids between 203 and 226, as well as those between 250 and 298 (***Figure 4***) (Kim *et al.*, 2007). Both conserved sections occur in what is generally referred to as the spacer region, which was named as such because it could not be identified by homologous structures (amino acids 198-295).

Little importance was initially associated with the spacer region because the sequence was predicted to have little structure and does not share homology with any known domains in proteins outside of YY-related proteins. Therefore, both of these conserved regions appear to be specific to YY-related proteins (Kim *et al.*, 2007). The spacer region has since been shown to be important for some protein-protein interactions in YY1, as its deletion can result in a lack of the ability of p53 to bind to it, while deletion of other YY1 regions does not have this effect (Sui *et al.*, 2004).

Within this broader spacer region, residues 199-228 have been identified as important for protein-protein interactions involving members of the polycomb group of proteins (Wilkinson *et al.*, 2006). These proteins form distinct complexes with each other, and are involved in

transcriptional silencing (Wilkinson *et al.*, 2010; Alfieri *et al.*, 2013). They accomplish this by modifying histones through members with either methyltransferase, E3 ligase, or deubiquitinase activity (Alfieri *et al.*, 2013). This stretch of YY1 within the spacer region has been called the recruitment of polycomb (REPO) domain (Wilkinson *et al.*, 2006).

The structure of the REPO domain of YY1 bound to the polycomb protein MBTD1 has been solved in a crystal structure, where the YY1 peptide of the REPO domain binds to its partner as two antiparallel β strands joined by a β hairpin loop, rather than the random coils that had been previously predicted (***Figure 6***) (Alfieri *et al.*, 2013). This is likely due to a shift from an unstructured conformation upon binding to a partner, as is commonly observed for intrinsically disordered proteins, which identifies the REPO domain as a likely molecular recognition factor (discussed in section 1.1.2.). Whether this particular motif has multiple binding partners, and whether it assumes the same structure for each, has yet to be explored.

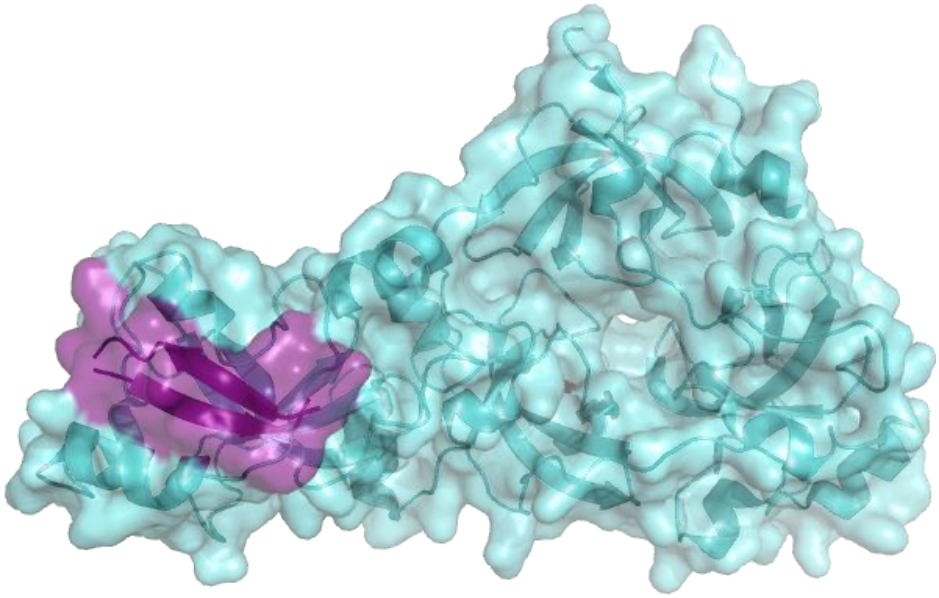


Figure 6: REPO domain of YY1 bound to the 4MBT domain of MBTD1.
The REPO domain of YY1 is shown in purple, forming two β strands which insert into a hydrophobic groove in the first MBT motif of the 4 MBT domain (blue). PBD accession code: 4c51 (Alfieri *et al.*, 2013). Image rendered using PyMOL molecular graphics system version 1.8, Schrödinger, LLC.

The second section of the spacer region is also conserved (Kim *et al.*, 2007). Given the ability of the REPO domain to form structure, it likely that YY1 also either contains some structure or acts as a molecular recognition factor. Overall, the intrinsic disorder of the N-terminal region implicates it strongly in protein-protein interactions, and may prove to act as a hub for these, recruiting several members of large complexes.

Besides interactions with other proteins, the disordered N-terminal regions of YY1 has also been associated with RNA binding (Sigova *et al.*, 2015). Further work needs to be done in order to confirm this, although intrinsically disordered proteins have previously been observed to bind RNA (Varadi *et al.*, 2015). This ability might be relevant to simultaneous binding of YY1 to *Xist*RNA and the X chromosome during X-inactivation (Wai *et al.*, 2016).

1.4.7.2. C-TERMINAL REGION

The GLI subgroup to which YY1 belongs refers to zinc fingers with the consensus sequence [Y/F]X**C**X_3G**C**X_3[F/Y]X_5LX_2**H**$X_{3\text{-}4}$**H**[T/S]GEKP, while the Krüppel subgroup members possess the consensus sequence [Y/F]X**C**X_2**C**X_3FX_5LX_2**H**XRX**H**TGEKP (Ruppert *et al.*, 1988), where the zinc binding residues are highlighted in bold. Besides indicating relatedness between proteins, specific consensus sequences for zinc fingers can be related to either what type of molecule the zinc finger binds, or what sequence of nucleic acid it recognises (Iuchi, 2001).

Proteins belonging to the C2H2 zinc finger protein family differ according to both the number of zinc fingers present, as well as the spacing between the zinc binding residues (Iuchi, 2001). The number of zinc fingers can influence how many types of molecules the protein can interact with simultaneously – this includes other proteins, DNA and RNA (Iuchi, 2001; Hashimoto *et al.*, 2017). The same zinc finger may be able to bind more than just one of these, or it may be more specific towards one (Wai *et al.*, 2016; Hashimoto *et al.*, 2017). The more zinc fingers in a protein, however, the more intermolecular interactions with different types of molecules will be possible (Iuchi, 2001). As YY1 has four zinc fingers (Houbaviy *et al.*, 1996), it is possible that its zinc finger domain is capable of interaction with other marcromolecules simultaneously with its DNA interaction, as proteins with more than three zinc fingers have been known to be able to do this (Iuchi, 2001). However, this has found not to be the case for DNA or RNA –

they compete with each other for binding to the zinc finger domain in YY1 (Wai *et al.*, 2016). There are a large number of different zinc fingers, all of which have different structures. Their ability to bind zinc in order to stabilise their structures suggests that binding other molecules, including proteins, is a result of shape and charge complementarity (Laity *et al.*, 2001). In the case of DNA binding, each zinc finger typically associates with three adjacent base pairs on the DNA when binding specifically (Choo and Klug, 1994; Hashimoto *et al.*, 2017).

1.4.6. DNA STRUCTURING AND NUCLEIC ACID BINDING FUNCTIONS

Although there is evidence that does indicate that YY1 is involved in the direct modes of transcriptional regulation described above, it has also been found to have several other functions related to DNA binding, which influence transcriptional activation and repression more indirectly. One of these functions is regulation of DNA structure itself, which is intrinsically linked to the ability of other transcription factors to bind to promoter sequences. One such structural alteration is chromatin accessibility, controlled by histone modifications such as acetylation and methylation (Li *et al.*, 2007), where YY1 interacts with several proteins responsible for such processes (Yang *et al.*, 1997; Rezai-Zadeh *et al.*, 2003; Li *et al.*, 2007). These interactions suggest a contribution to YY1's role in gene repression as well as activation by means of modulating the accessibility of chromatin to other transcription factors.

Besides its involvement in the formation of euchromatin and heterochromatin, YY1 has been observed to have involvement in the DNA looping found between distal enhancer regions and promoter regions (Weintraub *et al.*, 2017). YY1 oligomers are postulated to bind both regions, bringing them into proximity and thus forming enhancer-promoter loops, (Weintraub *et al.*, 2017). This phenomenon has been found to be important for transcriptional control.

To add to the complexity of YY1's functions, it has been observed to bind to several different forms of nucleic acid, each of which suggests a different function. Besides double stranded DNA binding (Houbaviy and Burley, 2001), YY1 has also been found to interact with Holliday junctions (Wu *et al.*, 2007). In conjunction to YY1's involvement with ATPases of the INO80 complex, this Holliday junction interaction implicates YY1 in DNA repair activities (Wu *et al.*, 2007; López-Perrote *et al.*, 2014). Other than Holliday junctions, some evidence has also suggested that YY1 may bind to single-stranded DNA (Wai *et al.*, 2016).

Evidence has been found which indicates that YY1 binds to RNA – this occurs both *in vitro* and *in vivo* (Sigova *et al.*, 2015; Wai *et al.*, 2016; Weintraub *et al.*, 2017). Part of this function in the cell is to bind non-specifically to *Xist* RNA, as part of YY1's X-inactivation function (Jeon and Lee, 2011). It has also been associated in the cell with pre-RNA at the sites of transcription (reviewed in Rambout *et al*., 2015). A stabilising role has also been suggested for RNA binding in terms of oligomer formation (Weintraub *et al.*, 2017), which may be especially significant to the role YY1 has in the formation of membrane-less organelles such as nuclear speckles (Galganski *et al.*, 2017).

Exactly how YY1 binds to RNA is still under a certain amount of debate. RNA has been observed to bind to the YY1 N-terminus in some cases (Sigova *et al.*, 2015), but not in all. In other cases, it binds to the zinc finger regions, but cannot bind simultaneously to DNA (Wai *et al.*, 2016). Given that YY1 seems to function to bind to *Xist* RNA and to DNA at the same time in order to perform X-inactivation (Jeon and Lee, 2011), the N-terminal region's capacity to bind to RNA sequences does seem plausible and warrants further investigation.

1.4.8. YY1 DNA BINDING

Amino acids 298-397 of YY1 constitute four C2H2 type zinc finger motifs, for which there is one crystal structure (***Figure 7***) (Gordon *et al.*, 2006). This section of the protein is more structured as well as more conserved compared to the N-terminal region. The zinc fingers are involved in sequence specific DNA binding and are important for transcriptional repression and possibly also activation (Shi *et al.*, 1997). Typical of zinc fingers of the C2H2 family, all four motifs conform to the typical secondary structure of two beta sheets and one alpha helix (Houbaviy *et al.*, 1996; Iuchi, 2001). This structure is stabilised by the zinc binding residues (Iuchi, 2001). The two histidine residues are located in the alpha helical region, which is responsible for most DNA contacts, as well as for the specificity with which each finger binds to DNA (***Figure 7***) (Houbaviy *et al.*, 1996).

When binding DNA, YY1 associates with the consensus sequence CG**CCAT**nTT, where n is any base (Kim *et al.*, 2007). The core sequence necessary for association is CCAT (Shi *et al.*, 1997). YY1 does show specificity towards this binding sequence (Kim *et al.*, 2007), but it is not the only one that it is capable of binding. A longer consensus sequence containing the same core GCCG**CCAT**TTTG has also been discovered for YY1 (Kim and Kim, 2009). This

sequence has been found in several, but not all, of YY1's physiological binding sites, including the *Peg3* and *Xist* genes (Kim and Kim, 2009). YY1 has higher affinity to this sequence compared to the shorter variant, which is likely due to an extended availability for zinc finger one's binding interactions (Kim and Kim, 2009). Another alternative sequence is the core binding sequence ACAT, which is less common (Shi *et al.*, 1997).

All four zinc fingers bind to the major groove of the DNA with different levels of contact (Houbaviy *et al.*, 1996). Zinc fingers two and three of YY1 both make extensive interactions with both the phosphate backbone and bases (***Figure 7***), while finger one is involved in mostly backbone contacts and finger four has only a few interactions with the nucleic bases (Houbaviy *et al.*, 1996).

The binding affinity of YY1 and its zinc fingers to DNA has been studied by a variety of groups with varying results. Many of these used electrophoretic mobility shift assays (EMSAs) to determine the K_D, which varied from the nanomolar to low mircromolar range (Usheva and Shenk, 1996; Lace *et al.*, 2009). More accurate techniques such as isothermal calorimetry (Allan *et al.*) and fluorescence anisotropy were in better agreement, and suggested a relatively weak K_D (~0.5-1 μM) when compared to many other transcription factors with K_Ds in the nanomolar range (Houbaviy and Burley, 2001; Golebiowski *et al.*, 2012). This affinity was observed to vary with changing salt concentrations, which falls within expected behaviour for DNA-protein interactions, given the number of polar contacts involved (Golebiowski *et al.*, 2012). It is possible that cofactors or other components such as PTMs are required for higher affinity binding of YY1 to DNA, or that this relatively low affinity is required for its function.

The thermodynamics and kinetics of DNA binding have been studied for the zinc finger region of YY1. The binding reaction is enthalpically driven, confirming that many polar interactions are involved in the binding process (Houbaviy and Burley, 2001). The binding reaction results in a large heat capacity change, which was speculated to result from large conformal changes in either the DNA binding domain DBD or in the DNA (Houbaviy and Burley, 2001). The stoichiometry for the DBD was calculated to be about 1, while the entropy did increase at 20 °C (0.98 kcal/mol^{-1}) (Houbaviy and Burley, 2001). Surface plasmon resonance (Milo *et al.*) has been used to determine binding kinetics of YY1 to three DNA sequences. The on rates of these ranged from about 1.03 - 3.41 x 10^6 $M^{-1}s^{-1}$ while the off rates ranged from 0.152 – 0.319 s^{-1} (Golebiowski *et al.*, 2012). These are relatively rapid (Golebiowski *et al.*, 2012). Interestingly, this low specificity and rapid on and off rate trend is associated with intrinsically disordered

protein interactions (Uversky and Dunker, 2010). Zinc finger motifs do contain structure, but the only crystal structure of this DBD is when it is bound to DNA. It is possible that it is more unstructured (likely to a much lesser degree than the N-terminal is) when unbound, which does explain the large change in heat capacity.

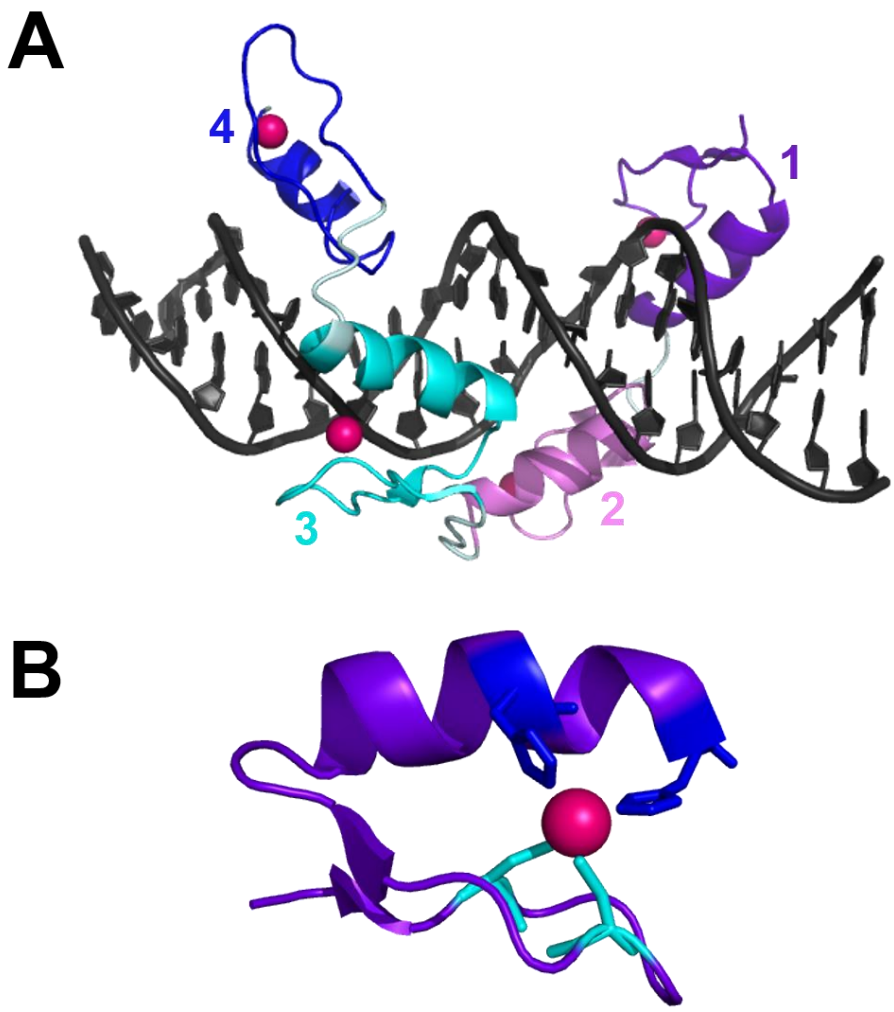


Figure 7: The crystal structure of the zinc finger domain of YY1 bound to DNA.
A) the full zinc finger domain of YY1 bound to DNA, as solved by Houbavity *et al*. (1996). The DNA is represented in dark grey, while the zinc fingers of YY1 (1-4) are distinguished by their different colours and the numbers next to them. The sections between the motifs are represented in light blue. The locations of the zinc ions bound to each structure are represented in hot pink. **B),** a close up of zinc finger 1 is shown with the DNA omitted. Zinc binding cysteine residues are in cyan and histidine residues are in blue. The zinc is represented by the pink sphere. The PDB code for this structure is 1UBD (Houbaviy *et al.*, 1996); it has a resolution of 2.5 Å and the images were generated using PyMOL molecular graphics system version 1.8, Schrödinger, LLC.

Unfortunately, reliable kinetic and thermodynamic information on DNA binding is not available for the full length YY1, due to either stability constraints, or due to a lack of an appropriate model to fit data (Houbaviy *et al.*, 1996; Houbaviy and Burley, 2001). The latter reason implies that the N-terminal of YY1 might add further complications to the DNA binding mechanism of YY1.

1.4.9. YY1 PROTEIN-PROTEIN INTERACTIONS

1.4.9.1. HOMOLOGOUS INTERACTIONS

Many transcription factors have been found to have the capacity to form homo-oligomers, which are frequently necessary for their function (Li *et al.*, 2004; Weintraub *et al.*, 2017; Thulo *et al.*, 2021). The oligomeric state of a transcription factor can influence its propensity to bind certain partners, and, therefore, affect its transcriptional regulation activity (Li *et al.*, 2004). Furthermore, an oligomer can exhibit altered behaviour and even altered structure compared to the monomer, which can diversify the protein's function (Weintraub *et al.*, 2017), especially when oligomerisation only occurs in certain conditions (for example, changes in concentration).

There have been some mixed reports about the oligomeric state of YY1. Relatively early studies concerning YY1 structure and DNA binding activity did not include much in the way of structural characterisation. Later in 2011, Golebiowsky *et al.* published a size exclusion chromatography (Cromer *et al.*) elution profile for both full length YY1 and its isolated DNA binding domain. This profile, along with a few others since published, included one monomeric peak for both full length YY1 and the DBD (Górecki *et al.*, 2015). This conflicted with several other findings published elsewhere (Wu *et al.*, 2007; López-Perrote *et al.*, 2014; Weintraub *et al.*, 2017), which indicated that YY1 forms oligomers, both *in vitro* and *in vivo*. These conflicting findings were also based on SEC experiments, both preparative and analytical, in which YY1 was eluted in several peaks. This suggested oligomeric heterogeneity, where the main peaks observed were calculated as 158 kDa, 440 kDa and 670 kDa. These sizes are significantly larger than the expected YY1 molecular weight, which is 45 kDa – although, YY1 is not a globular protein, (Górecki *et al.*, 2015), which makes it difficult to compare YY1 elution times with those of the globular standards used. Protein-protein interactions of

intrinsically disordered proteins are typically tightly controlled by PTMs. Therefore, it is possible that the conflict in observations regarding the oligomeric state of YY1 result from differences in the PTMs present in YY1 expressed in different hosts and conditions.

The YY1 oligomeric mixture was observed using cryo-electron microscopy (López-Perrote *et al.*, 2014), which suggested that the two smaller peaks found in SEC correspond with concentration dependent mixtures of dimers, which further associate into tetramers. The larger YY1 complexes were reported to have aggregate-like conformations, and were, therefore, not further investigated. At a resolution of 20 Å, relatively little detail could be garnered from the images of YY1 produced. However, the symmetry of the smaller complexes suggested a dimeric conformation, which was bell shaped. For the larger species, the overall length seemed to increase two-fold, which suggests a tetramer. Details on how the dimeric subunits were associated with each other could not be sufficiently observed. Therefore, López-Perrote *et al*. (2014) suggested two potential models on how each unit might be aligned. One has them aligned in a row laterally, while the other proposes the two dimeric pairs interacting with each other on a back to back basis. It is possible that the YY1 tetramers are heterogeneous, and could associate in more than one manner.

The ability of YY1 to self-associate has also been observed *in vivo*, suggesting that the oligomers observed *in vitro* are significant. Furthermore, RNA seems to stabilise the formation of YY1 oligomers *in vivo* (Weintraub *et al.*, 2017), and has also been found to encourage the binding of YY1 at active regulatory elements (Sigova *et al.*, 2015). This suggests that homo-oligomerisation may be an important factor influencing YY1 DNA binding, which may be regulated by the RNA expressed by the genes to which YY1 binds (Weintraub *et al.*, 2017).

Some studies have been conducted on whether YY1 forms oligomers when bound to DNA, although they have not been extensive. Wu *et al*. (2007) observed that YY1 binds Holliday junctions in multiple sized bands, which suggests that it may bind to this DNA conformation in several oligomeric states. The same study did not observe multiple species of YY1-bound DNA when double stranded oligonucleotides were bound to YY1. When using isolated oligomeric YY1 samples, however, another study demonstrated that these did bind to DNA (also as one species), and also that these oligomers seemed to have an altered specificity for DNA compared to the monomeric forms previously studied (López-Perrote *et al.*, 2014). The larger YY1 complexes were reported to have bound to DNA not containing the YY1 consensus

sequence (López-Perrote *et al.*, 2014). Whether there is a shift in oligomeric state upon DNA binding, however, has not been observed.

There are many possibilities for how self-association might alter or regulate a protein's ability to regulate gene expression. These oligomers could have altered binding capabilities to other proteins, to RNA or to DNA. One effect which has been observed is an apparent alteration in DNA specificity when oligomerisation occurs (López-Perrote *et al.*, 2014). Besides this, the exact role of YY1 oligomerisation in DNA binding has not been fully explored. However, it has been proposed that YY1 dimers may function by regulating DNA structure – specifically, by binding to distal enhancer and promoter regions, and then to each other, thus looping the DNA and activating or repressing transcription (Weintraub *et al.*, 2017). YY1 has been found to self-associate within these DNA regions, and its removal has been found to severely affect transcription of many genes, suggesting that this may be an important function of homo-oligomerisation (Weintraub *et al.*, 2017). Whether it is the only one, however, has not been fully explored.

1.4.9.2. YY1 HETEROLOGOUS INTERACTIONS

YY1 has many interacting partners which are involved in diverse processes. These include RNA polymerase (Shi *et al.*, 1997), histone modifying proteins such as HDAC2(Thomas and Seto, 1999), DNA repair proteins of the INO80 complex (Wu *et al.*, 2007) and proteins involved in the cell cycle such as p53 and Hdm2 (Sui *et al.*, 2004). When binding to YY1, a large proportion of its partners appear to interact roughly within two main regions, suggesting that YY1 contains two protein binding hotspots (Thomas and Seto, 1999; He *et al.*, 2007). The first occurs in the central region of the protein, at about residues 150-200, which contains both GA and GK rich regions, along with a portion of the spacer region (Thomas and Seto, 1999). The second occurs within the zinc finger region and some of its flanking spacer region (Thomas and Seto, 1999). For some proteins, such as E1A and TBP, both domains appear to be important for interactions to occur (Shi *et al.*, 1997). Other proteins, however, interact with only one of the protein binding domains, such as in the case of HDAC2, which interacts with the central hotspot (Yang *et al.*, 1996), and CREB, which interacts with the C-terminal hotspot (Galvin and Shi, 1997).

While several protein recognition sequences specific to its binding partners have been found to be present in the sequence of YY1, relatively little structural information has been found concerning how YY1 interacts with its binding partners. The exception to this is the crystal structure of the YY1 REPO domain bound to the 4MBT domain of MBTD1, a polycomb protein (Alfieri *et al.*, 2013). This interaction is heavily dependent on hydrophobic interactions, where the two stranded beta sheet of the YY1 REPO domain binds to a hydrophobic pocket in in the 4MBT domain of MBTD1 (Alfieri *et al.*, 2013). This interaction is furthered by polar interactions (Alfieri *et al.*, 2013).

Furthermore, it is possible that the C-terminal region of YY1 associated with protein-protein interactions involves different zinc fingers depending on the protein (Iuchi, 2001). As a zinc finger can be specific for DNA, RNA and proteins, and YY1 has four zinc fingers, it is possible that this domain can simultaneously bind DNA and proteins (Houbaviy *et al.*, 1996; Iuchi, 2001). Alternatively, if a protein bound to zinc fingers 2 or 3, which are most involved in interacting with DNA, it is possible that YY1 becomes inhibited from binding DNA.

The ability of cofactors to bind to YY1 appears to be a vital aspect of the regulations of YY1's dual function (Deng *et al.*, 2010; Wilkinson *et al.*, 2010). Understanding more about how YY1 interacts with a variety of its partners would give a significant amount more insight into how it represses and activates transcription. Given its many functions and potential binding sites, it is highly likely that YY1 acts as a hub for multiple interactions and complex formations.

Despite the identification of a number of binding partners, relatively little is known about the exact mechanism by which YY1 binds to most of its partners. While the location of interaction on YY1 has been mapped for several of its interacting partners, frequently only association and the effect on transcription have been investigated. FOXP2 is one such binding partner (Estruch *et al.*, 2018). While the domains of FOXP2 that interact with YY1 have been investigated, nothing more is known about the interaction between these two proteins (Estruch *et al.*, 2018). For a better understanding of how their interaction might influence neural development, the association between the two transcription factors needs to be investigated.

RATIONALE

Transcriptional regulation is a universally important process which is still not fully understood. A single transcription factor can have be involved in such diverse processes that the complexities of the mechanisms behind these processes become difficult to unwind. This is largely due to the formation of a diverse interaction network consisting of proteins, RNA and DNA (Uversky, 2019). An understanding of this transcriptional interactome is necessary in order to find treatments for disorders and diseases caused by the mis-regulation of gene expression. Mutation or mis-expression of a transcription factor can disrupt its interactions with other proteins and nucleic acids, resulting in a change in gene expression in important pathways (Nabais Sá *et al.*, 2019). The resulting disorders can severely affect the lives of those who have them, as well as their families.

Two of the transcription factors implicated in disorders arising from altered transcriptional control are YY1 and FOXP2. Both of these transcription factors are essential in the development of many tissues and they are also both heavily involved in neural development (Bowers and Konopka, 2012; Gabriele *et al.*, 2017). Both YY1 and FOXP2 are reliant on their macromolecular interactions to control gene expression (Weintraub *et al.*, 2017). Therefore, to gain an understanding of how YY1 and FOXP2 are implicated in ASD and speech development, among other disorders, a full understanding of their macromolecular interactions is necessary. As such, the macromolecular interactions between YY1, DNA and FOXP2 were explored in this study, to better understand their roles in transcriptional regulation and thus to identify how diseases and disorders associated with each protein are caused and might be treated.

YY1 is a ubiquitously expressed transcription factor associated with multiple cellular pathways via several mechanisms (Gordon *et al.*, 2006). DNA repair (Wu *et al.*, 2007), transcriptional initiation (Shi *et al.*, 1997), liquid-liquid phase transitions (Salichs *et al.*, 2009), chromatin modification (Thomas and Seto, 1999) and DNA structuring (Weintraub *et al.*, 2017) are all examples of such processes. To achieve such a wide range of functions, YY1 relies on its nucleic acid binding activity together with its homo- and heterologous protein-protein interactions (Wu *et al.*, 2007; Jeon and Lee, 2011; Weintraub *et al.*, 2017). While these processes have each been identified and explored to various extents, the associations between

YY1 and other proteins including its associations with itself have not been well studied. Since these associations are a critical step in granting this protein its regulatory control, it is important to be able to describe these associations in detail.

Before one can fully address these YY1 interactions, it is necessary to understand the structure and folding of YY1 alone and how this structure changes upon interaction with nucleic acid or other proteins. There is some evidence to suggest that YY1 has some degree of intrinsic disorder (Górecki *et al.*, 2015), however, there is not sufficient detail to draw any real conclusions regarding this. In this study, the focus is on obtaining clarity on the regions of intrinsic disorder, their dynamics, their structure and how they form associations upon protein-protein or protein-DNA interaction. YY1 oligomerisation has been proposed to function in DNA looping (Weintraub *et al.*, 2017) and if this can be proved, it will pave the way for more detailed interpretations about this protein's function. It is clear from the literature that YY1 oligomeric status is an important contributing factor to its function. YY1 oligomers have been observed to exhibit lower specificity DNA binding compared to monomeric species (López-Perrote *et al.*, 2014) and YY1 oligomers have also been shown to be stabilised by RNA binding (Weintraub *et al.*, 2017). Despite these interesting discoveries, the mechanism of YY1 oligomerisation remains unexplored.

An understanding of how the structure, oligomeric and DNA-binding behaviour of YY1 inter-relate will facilitate a deeper understanding of the mechanism by which it accomplishes its functions within the transcriptional interactome. This will lead to a better understanding of how YY1 is regulated and how it leads to the many disorders with which it has been associated.

In addition to YY1's homologous protein-protein interactions, its heterologous protein-protein contacts are vital for its functions. This is in part due to YY1's role in recruitment of other transcription factors. YY1 has been suggested to interact with FOXP2 *in vivo* (Estruch *et al.*, 2018). However, whether the two proteins interact in isolation or as part of a large complex is unknown. Additionally, the relevance of this interaction has not previously been explored. By observing their interaction and how it affects the DNA biding behaviour of each protein, a better understanding of how the FOXP2-YY1 interaction may play a role in control of gene expression could be achieved. The exploration of this interaction could help better determine how both proteins regulate processes with which they are mutually involved. This includes neural developmental processes related to ASD and speech disorders.

Therefore, to understand the manner in which YY1 and FOXP2 influence transcription, it is necessary to both determine if and how they bind to their partners, and how this interaction influences their other binding capacities. In this way, we can determine how each transcription factor's interactions alter gene expression, and therefore possibly affect phenotype.

AIM AND OBJECTIVES

3.1. AIM

The aim of this project is to investigate the binding mechanisms of YY1 with regards to its homologous and heterologous interactions by determining the YY1-YY1, YY1-DNA and YY1-FHD binding affinities, dynamics, and the subsequent effect of the protein-protein interactions on DNA binding and hence transcriptional regulation.

3.2. OBJECTIVES

- To transform T7 cells with FOXP2 FHD and YY1 or TCP-YY1.
- To express and purify FOXP2 FHD and YY1 or TCP-YY1 using liquid chromatography.
- To investigate and compare the secondary, tertiary and quaternary structures of YY1 and FOXP2 FHD in the presence and absence of DNA using circular dichroism (CD), intrinsic tryptophan fluorescence, and size exclusion chromatography respectively and to use this to outline how YY1 uses oligomerisation to facilitate its interactions.
- To use HDXMS to characterise the dynamics of YY1 when in isolation and also when bound to DNA, so as to determine changes undergone by the protein upon DNA binding.
- To characterise the interactions between YY1 and FOXP2 FHD using fluorescence anisotropy under different conditions through determining the K_D.
- To use EMSAs and fluorescence anisotropy to characterise the interaction between each protein and their respective DNA binding sequences. EMSAs will also be used to confirm that each protein is specific to its own DNA sequence. By using these techniques, relative binding affinities and specificities can be observed.
- To determine whether the FHD-YY1 interaction affects the K_D of the YY1-DNA interaction by using fluorescence anisotropy and supershift assays.

METHODS AND MATERIALS

4.1. PROTEIN CONSTRUCTS

In order to study the interaction between YY1 and the FOXP2 FHD in isolation, it was necessary to make different constructs of each protein. For all proteins, the sequences were codon optimised for *E. coli* before being added to a vector (Genscript, USA).

YY1 was expressed with an N-terminal hexahistidine tag (His tag) to enable purification using immobilised metal affinity chromatography (IMAC) (**Table 1**). While YY1 contains an intrinsic histidine rich region, the lack of previous structural studies made it unclear as to whether these residues were exposed enough to the solution to bind a column. Therefore, the construct used was kept consistent with most of those in the literature – an N-terminal non-cleavable His tag was added (Golebiowski *et al.*, 2011). This construct was referred to as WT-YY1.

An additional construct of YY1 (TCP YY1) was also created for fluorescent labelling purposes, as the wild type could not be covalently labelled via a group such a maleimide, which binds to cysteine residues. Although YY1 does contain cysteine, all eight of these make up zinc finger motifs, and are therefore vital for the correct folding and DNA binding activity of YY1. Given the functional importance of these cysteine residues, as well as the relatively large number of them (which would make stoichiometry of binding unsuitable), labelling WT cysteine residues was not an option. Instead, a new tag had to be added to the protein, which introduced a site specifically bound by certain labels. A tetracysteine peptide (TCP) tag (CCKAEAACC) was introduced after the N-terminal His tag, with a stretch of linker amino acids in between (**Table 1**).

The WT FHD construct also contained an N-terminal His tag, with a thrombin cleavage sequence, so the tag could later be removed from the protein (**Table 1**).

Table 1: The YY1 and FOXP2 constructs used in this study

Protein	Construct	Residues	Construct
FHD	FOXP2 FHD	502-585	MASMTGGQQMGMHHHHHHSSGLVPRGSFHD
YY1	WT YY1	1-414	HMGSSHHHHHHYY1
	TCP-YY1	1-414	HMGSSHHHHHHASMTGGQQMCCKAEAACCGGYY1

Sequence colour codes: Grey indicates linker regions; purple indicates the His tags, blue is for thrombin cleavage sites, green for the TCP tag and pink for the actual protein (not sequence).

4.2. TRANSFORMATION

All constructs were cloned into pET-11a vectors in between the BamHI and NdeI sites (***Figure 8***) by Genscript (USA). The plasmids were then used to transform T7 *Escherishia coli* cells (New England Biolab, USA). The competent cells stored at -80 °C were defrosted on ice before 6 µl of 100 ng/µl of pET-11a vector was added to them. After flicking this mixture lightly five times, the cells were incubated on ice for thirty minutes before undergoing heat shock for 35 seconds at 42 °C. The cells were then immediately incubated on ice for 5 minutes, before 950 µl sterile Super Optimal broth with Catabolite repression medium (SOC, containing 20 g/l tryptone, 5 g/l yeast extract, 4.8 g/l $MgSO_4$, 3.603 g/l dextrose, 0.5 g/l NaCl and 0.186 g/l KCl) was added to support cell replication. This solution was incubated at 37 °C while being shaken at 230 rpm for 1.5 hours.

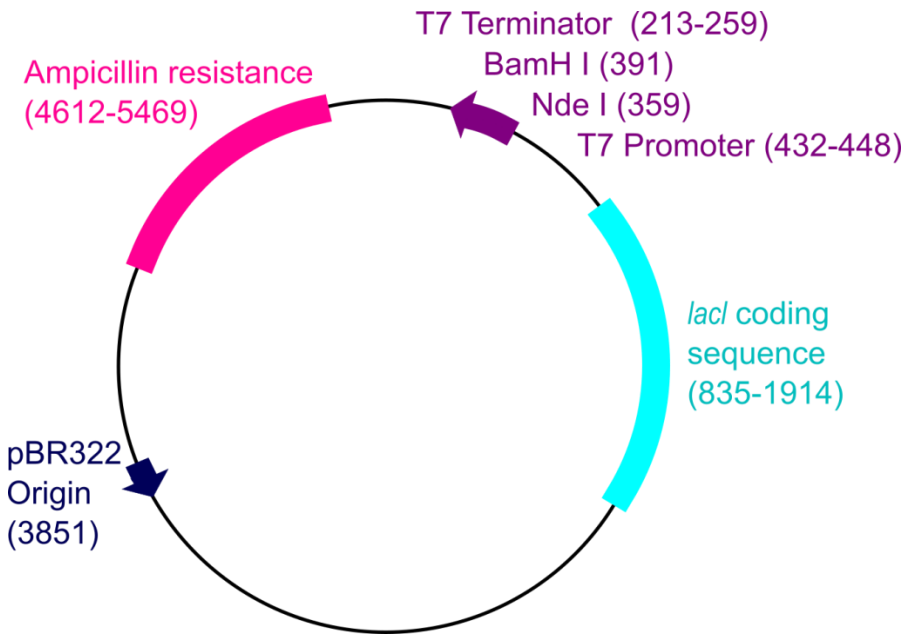


Figure 8: Diagram of a pET11-a plasmid with sites of interest.
The codon optimised sequence of YY1 was inserted in between the BamHI and NdeI sites in the purple region, between the T7 promoter and terminators. The *lacI* coding sequence allowing for controlled overexpression is shown in blue, the ampicillin resistance sequence in pink and the pBR322 Origin in dark blue. The numbers indicate the base pair number in the plasmid sequence. This image was made using Inskape 1.0.

From the transformed cell solution, 100 µl was spread onto sterile agar plates containing lysogeny broth (LB) with 0.1 mg/ml ampicillin to select for successfully transformed cells as the pET-11a vector confers antibiotic resistance. Furthermore, negative control plates with no cells were prepared so as to confirm growth had not come from contamination. The plates were incubated at 37 °C overnight, after which they were removed and appropriate colonies were picked – where possible, at least one from each plate was used.

Single, isolated colonies were transferred into 50 ml of sterile LB containing 0.1 mg/ml ampicillin, and grown at 37 °C while shaking at 230 rpm until they reached an OD_{600} of 0.6, where they were either induced for expression or made into 1:1 glycerol stocks. A portion of this solution was also used in a miniprep to extract the plasmid for sequencing by Inqaba Biotech (RSA) to ensure that transformation was successful and contained the correct sequence.

4.3. YY1 INDUCTION TRIALS

YY1 was induced under differing conditions in order to optimise the conditions for its soluble expression. Two temperatures were tested (37 and 20 °C), along with four different isopropyl β–D–1–thiogalactopyranoside (IPTG) concentrations at each temperature: 0, 0.2, 0.5 and 1 mM. Furthermore, the difference in expression when using LB or 2x Yeast Tryptone broth (YT) (1.6% (w/v) tryptone; 1% (w/v) yeast extract; 0.5% (w/v) NaCl) was tested at 20 °C. For all other trials, 2x YT was used. Samples from each condition were taken at the time points of 0, 1.5, 3, 4, 5, 6, 11 and 24 hours for the trials at 37 °C. Because a lower temperature would result in slower expression, fewer time points were required for the 20 °C expression trial: 4, 6 and 22 hour samples were taken for this trial. When inducing with IPTG, $ZnCl_2$ was also added to the flasks at a working concentration of 100 μM, as recommended by Golebiowski *et al*. (2011). This was to help protein folding, as YY1 is a zinc finger protein and therefore requires zinc ions in order to fold.

In order to lyse the cells, samples were sonicated at 45% 3 times for 3 seconds and rested on ice in between. The samples were subsequently separated into insoluble and soluble fractions by centrifuging each at 10 000 x g for two minutes. They were they resolved using sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS PAGE), and the most favourable condition at which soluble YY1 was expressed was selected.

4.4. EXPRESSION

For large scale expression, FOXP2 FHD was expressed as previously described by Blane and Fanucchi (2015), as follows. A 500 ml flask containing 100 ml of YT media was autoclaved and inoculated with a FOXP2 FHD glycerol stock and 0.1 mg/ml ampicillin. The culture was grown at 37 °C in a shaker incubator set to 250 rpm overnight. At least 1.5 L of YT media was prepared and dispensed into expression flasks. The volume of media in each flask did not exceed more than 1/5 of the flask's capacity. The media was autoclaved and allowed to cool after which each flask was inoculated with a 1 in 50 dilution of overnight culture and 0.1 mg/ml ampicillin. The cultures were grown at 37 °C in a shaker incubator set to 250 rpm until an OD_{600} of 0.6. was reached. The cultures were then cooled to 20 °C before being induction with 0.5 mM IPTG. The cultures were grown at 20 °C in a shaker incubator set to 250 rpm for 22 to

24 hours after which the cells were harvested by centrifugation at 5000 xg at 4 °C for 15 minutes. The resulting pellets containing the FHD producing cells were resuspended into 50 ml of equilibration buffer (**Table 4**) per 1 L of culture and stored at -20 °C.

YY1 was expressed under these same conditions, except expression was induced with 0.2 mM IPTG in addition to 100 μM $ZnCl_2$, which was required for the zinc fingers.

4.5. SDS PAGE

SDS PAGE is a non-native electrophoretic technique used to separate proteins by size. It is a commonly used technique when determining whether a protein has been overexpressed, or when observing how pure it is. Samples are prepared with sodium dodecyl sulfate (SDS), which is an ionic surfactant. It binds to the hydrophobic groups of proteins, causing them to linearise, while coating them with an overall negative charge with a consistent mass to charge ratio. This results in all proteins in the mixture migrating towards the anode, their separation affected by size rather than charge. Additionally, the samples contain the reducing agent β-mercaptoethanol, which removes any disulphide bridges which may alter the apparent size of a protein under non-reducing conditions. The samples are heated to ensure that all proteins in the sample are fully denatured.

Depending on the size of the target proteins, certain variations of this technique are more appropriate than others. The glycine method, as described by Laemmli (1970), makes use of glycine as trailing ions, and does not require separate cathode and anode buffers. This technique is appropriate for larger proteins, such as YY1. For smaller proteins such as the FHD, however, a stronger resolving power is required, making the tricine method (Schägger and von Jagow, 1987) more appropriate. As the name suggests, this method makes use of tricine as a trailing ion, which has a slightly lower pk_a, causing it to stack the bands on the separating gel more tightly.

Samples for glycine and tricine gels were made up in 1:1 and 1:3 ratios of protein:reducing sample buffer, respectively. They were thoroughly mixed, sonicated for three one second bursts, heated at 94 °C for five minutes, vortexed for ten seconds and spun down briefly before being loaded onto the gel, to ensure thorough denaturation and homogeneity.

YY1 was resolved via the glycine method, while the FHD was resolved using the tricine method. The gel and buffer compositions for both glycine and tricine are shown in **Table 2**. For both methods, a Bio-Rad MiniPROTEAN^{TM} Tetra Cell electrophoresis set (USA) was used to polymerise the gels and to resolve the samples. The gels were polymerised in a clean assembled casting stand, where the separating gel mixture was allowed to polymerise while covered by a layer of isopropanol to exclude air. The stacking gel was then added and allowed to polymerise with the combs in place. YY1 separating gels contained 12.5% acrylamide, while FHD separating gels contained 16% acrylamide. The composition of the gels for each method is described in **Table 2**.

After polymerisation, the gel tank and cassettes were assembled, and the relevant buffers were added (**Table 3**). Ten microliters of each sample was added to the wells after they had been rinsed with the relevant buffer. The gels were resolved at 20 °C at 60 V while passing through the stacking gel, and the voltage was increased to 160 V once the samples reached the stacking gel until the dye front had reached the bottom of the gels.

Table 2: SDS PAGE composition for glycine gels and tricine gels.

Buffer	Composition	Glycine/ Tricine	Glycine 4% Stacking (Per 7.11 ml)	Glycine 12.5% Separating (per 15 ml)	Tricine 4% Stacking (Per 12 ml)	Tricine 16% Separating (per 15 ml)
Monomer solution	30% (w/v) acrylamide 2.7% (w/v) bis-acrylamide	Glycine	940 µl	6.25 ml	-	-
Separating gel buffer	1.5 M tris-HCl pH 8.8	Glycine	0	3.75	-	-
Stacking gel buffer	500 mM tris-HCl pH 6.8	Glycine	1.75 ml	-	-	-
10% SDS	10% (w/v) SDS	Glycine	70 µl	150 µl	-	-
AB-3	49.5% (w/v) Acrylamide 3% (w/v) bis-acrylamide	Tricine	-	-	1 ml	5 ml
3x Gel buffer	3 M Tris 1 M HCl pH 8.45 0.3% (w/v) SDS	Tricine	-	-	3 ml	5 ml
dH_2O		Both	4.3 ml	4.75 ml	8 ml	5 ml
To initiate gel polymerisation:						
APS	10% (w/v) ammonium persulfate	Both	35 µl	75 µl	90 µl	50 µl
TEMED	N,N,N′,N′-Tetramethyl Ethylenediamine (TEMED)	Both	15 µl	7.5 µl	20 µl	12.5 µl

Table 3: Buffers used during glycine and tricine methods of SDS PAGE.

Buffer	Composition	Tricine/ Glycine	Purpose
Tank Buffer	25 mM Tris-HCl pH 8.3 192 mM glycine 0.1% (w/v) SDS	Glycine	Running buffer
Anode buffer	100 mM Tris 22.5 mM HCl pH 8.9	Tricine	
Cathode buffer	100 mM Tris 100 mM Tricine 0.1% (w/v) SDS pH ~8.25	Tricine	
Reducing sample buffer (Glycine)	125 mM Tris-HCl pH 6.8 4% (w/v) SDS 20% (v/v) glycerol 10% (v/v) β-mercaptoethanol	Glycine	Sample preparation
Reducing sample buffer (Tricine)	150 mM Tris-HCl pH 7 12% (w/v) SDS 30% (v/v) Glycerol 6% (v/v) β-mercaptoethanol 0.05% Coomassie blue	Tricine	
Stain Solution	0.1% (w/v) Coomassie blue R 10% (v/v) acetic acid 50% (v/v) methanol 40% (v/v) dH_2O	Both	Stains and fixes bands of protein
Destain solution	0% (v/v) acetic acid 50% (v/v) methanol 40% (v/v) dH_2O	Both	Removes stain in gel that is not bound to

	protein, making protein bands visible

4.6. PROTEIN PURIFICATION

4.6.1 PREPARATION OF CELL LYSATE

Both the YY1 and the FHD cell solutions stored at -20 °C after expression were defrosted and lysed in 0.1 mg/ml lysozyme, while 1 mM phenylmethylsulfonyl fluoride was added to inhibit degradation of the proteins. To allow time for the lysozyme to take effect, this mixture was incubated at 20 °C while slowly being inverted for 45 minutes. It was then allowed to cool on ice before being sonicated in seven twenty five second bursts (or until the solution was sufficiently less viscous), with at least forty seconds of sitting on ice in between each sonication. The cells were then incubated for thirty minutes at 20 °C with deoxyribonuclease I before they were centrifuged for 20 minutes at 20 000 x g. The supernatant was further filtered using a 0.45 μM steri-filter before purification via chromatography.

4.6.2. IMAC

IMAC is a technique used to isolate proteins with histidine rich regions. As these are relatively rare in the proteome, a protein with a terminal polyhistidine tag can be isolated from other proteins in cell lysate relatively easily (Petty, 2001). The stationary phase consists of the column resin, which is charged with a cation such as Ni^{2+} or Co^{2+}, both of which bind to the imidazole group of the histidines with different degrees of specificity (***Figure 9***). The cell lysate is passed through the column, and proteins without an attached His tag elute in the mobile phase (Petty, 2001). Non-specific binding to the column can be further removed by using a wash step with higher amounts of components such as salt or imidazole (Arnau *et al.*). The mobile phase is then changed to a buffer with a high concentration of imidazole, which competitively binds the stationary phase, causing the isolated protein to elute.

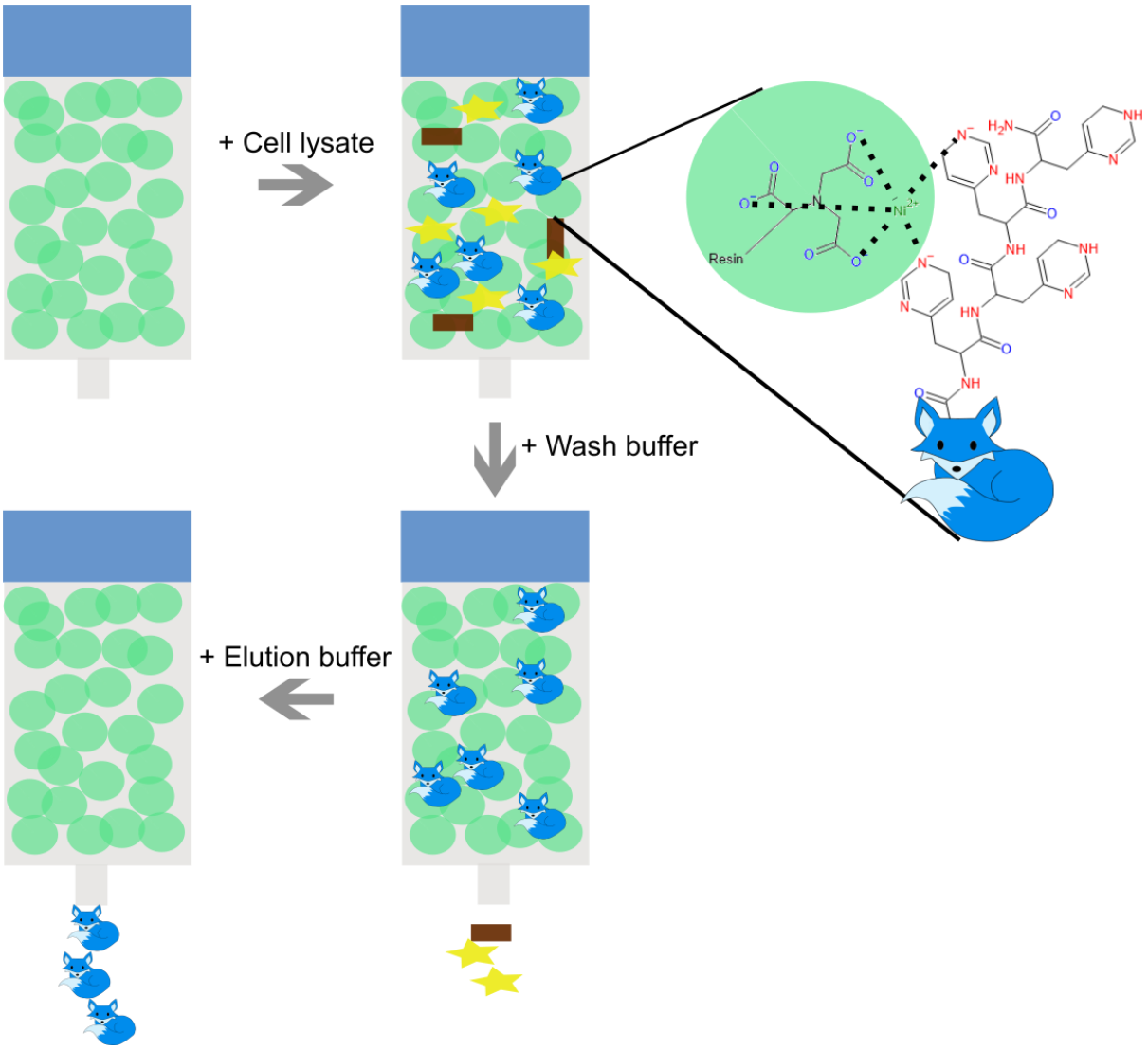


Figure 9: Diagram of the processes that occur during IMAC.
Resin beads are shown as green circles, while the column they are contained in is grey and blue. FHD is represented by a blue fox, while other contamination is shown as brown blocks and yellow stars. On the right, a zoomed in version of how the imidazole of the histidine groups attached to a protein bind to the Nickel, which is bound by the resin. Image made with InkScape 1.0 and chemical structure with ChemSketch (Freeware) 2021.1.1.

All constructs were purified using IMAC. The filtered supernatant was loaded onto a 5 ml HiTRAP column (GE Healthcare, USA) charged with either Ni^{2+} (FHD) or Co^{2+} (YY1 constructs). The same protocol as previously described by Blane and Fanucchi (2015) was used for the forkhead domain, while purification trials were required for YY1. The composition of all chromatography buffers are detailed in **Table 4**.

Table 4: Buffers used during preparative chromatography

Buffer	FHD	YY1 constructs	Technique
Equilibration buffer	20 mM Tris-HCl, pH 7.5 500 mM NaCl 30 mM imidazole	20 mM Tris-HCl, pH 7.5 500 mM NaCl 50 mM imidazole	IMAC SEC
Detergent-salt wash	20 mM Tris-HCl, pH 7.5 1.5 M NaCl 30 mM imidazole	20 mM Tris-HCl, pH 7.5 1.5 M NaCl 50 mM imidazole 0.5% TWEEN 20 1% Triton-X	IMAC
Imidazole	N/A	20 mM Tris-HCl, pH 7.5 500 mM NaCl 100 mM imidazole	IMAC
Elution buffer	20 mM Tris-HCl, pH 7.5 500 mM NaCl 300 mM imidazole	20 mM Tris-HCl, pH 7.5 500 mM NaCl 800 mM imidazole	IMAC
Cleavage buffer	20 mM Tris-HCl, pH 8 2 mM KCl 100 mM NaCl	N/A	Thrombin Cleavage

An ÄKTA prime pump system (GE Healthcare, USA) was used for IMAC, and the absorbance of the solution leaving the column at A_{280} was monitored using the AKTA's internal detector. The column was equilibrated with ten buffer volumes of equilibration buffer before the supernatant of the cell lysate containing FOXP2 was injected. After the absorbance at A_{280} returned to baseline, ten column volumes of detergent-salt wash buffer used to remove any DNA and protein contamination. Pure FOXP2 was eluted using elution buffer and collected in fractions.

YY1 was purified in a similar manner as described above, except it had two wash steps – one containing salt and detergent, to try and disrupt interactions between YY1 and other proteins or DNA, while the other wash contained an increase in imidazole to help remove other proteins bound non-specifically to the column (**Table 4**). Additionally, the equilibration buffer used had a higher concentration of imidazole compared to that for the FHD to prevent contaminants binding non-specifically to the column (**Table 4**). This was possible as YY1 contains histidine rich regions and thus bound the column very tightly. YY1 was eluted in 800 mM imidazole by allowing the column to incubate in the elution buffer for 45 minutes to an hour before eluting, and then repeating this process once the A_{280} absorbance peak begun to tail again. This was to help YY1 detach from the column, as it was bound very tightly. Equilibration buffer was used in between each buffer change of IMAC to return conductance and absorbance readings to the baseline, remove any traces of detergents and to ensure the correct salt concentration before elution.

4.8.2. EMSA

An electrophoretic mobility shift assay (EMSA) is a technique used to observe binding between proteins and DNA (Garner and Revzin, 1986). It is a native gel technique in which DNA in the presence of protein is resolved. Free DNA will travel the furthest distance on the gel, while if it becomes bound to protein the overall size will increase, making migration through the gel pores slower (Fried, 1989). This results in bands becoming resolved higher up on the gel in comparison to the free DNA (***Figure 10***). Therefore, when viewing the position of the DNA bands, one can observe whether the protein has bound the DNA. Furthermore, one can observe whether multiple higher order structures are being formed because larger DNA-protein complexes will create bands that shift higher up on the gel compared to smaller binding complexes.

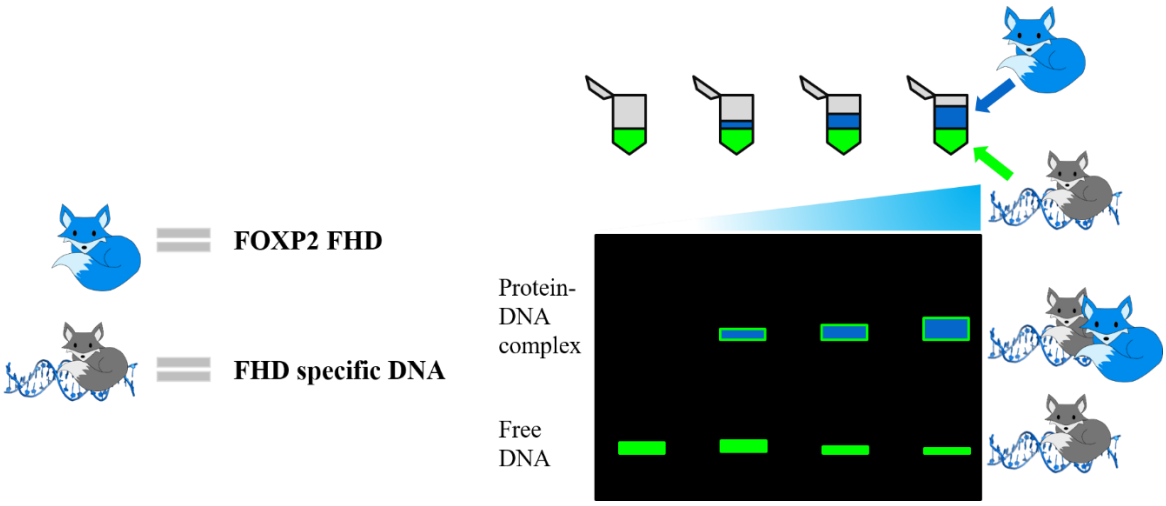


Figure 10: Diagram of an EMSA
Samples of the same concentration of DNA and increasing protein concentrations are each resolved in a separate lane. If there is a DNA-protein interaction, the protein-DNA complex separates from the free DNA complex, moving less distance from the wells due to size. DNA staining is used to visualise bands. Image created using InkScape 1.0 and Powerpoint (2013).

EMSAs were used not only to test the activity of each protein with its relevant DNA sequence, but also to test the specificity of binding each protein had for the other DNA sequences. An 8% native acrylamide gel was used for each EMSA. Each gel was made according to recipe in **Table 5**. The gels were set up using the same procedure as in SDS PAGE, but the buffer used in the tank was 1X Tris-borate-EDTA buffer (TBE) buffer. Samples were prepared with a constant amount of DNA (0.5 μM), while each lane had increasing protein concentrations. To each sample, 2 mM dithiothreitol (DTT) was added, and one fifth of the volume for each consisted of 5 x EMSA binding buffer (10 mM HEPES, 1 mM $MgCl_2$, 100 mM KCl, 10% (w/v) glycerol, 0.1 mg/ml bovine serum albumin, pH 7.5). Samples were allowed to incubate at 4 °C for at least 30 minutes before use to allow for binding. The gels were resolved using a voltage of 160 V for one and a half hours at 4 °C.

Table 5: EMSA gel composition table.

Component	Composition	Per 1 Gel
Triethylene glycol (TEG)	100% TEG	1.5 ml
Monomer solution	30% (w/v) acrylamide 2.7% (w/v) bis-acrylamide	2 ml
5 x TBE	445 mM tris-HCL 445 mM boric acid 10 mM EDTA pH 8.3	1.5 ml
dH_2O	-	2.5 ml
To initiate polymerisation reaction:		
10% APS	10% APS	50 µl
TEMED	-	10 µl

4.6.3. THROMBIN CLEAVAGE

When possible, it is preferable to cleave the His tag off a construct after purification. This is especially important for small proteins, as the addition of the tag can result in altered function or behaviour of the protein (Booth *et al.*, 2018). Therefore, a cleavage site specific to an appropriate protease, such as thrombin, can be introduced in between the tag and the actual protein sequence if no such cleavage sites exist in the actual protein sequence. The protein is incubated in thrombin until it has been cleaved, after which the thrombin, cleaved protein, uncleaved protein and the His tag need to be separated. This is easily done using a benzamidine column and IMAC column in tandem (***Figure 11***) (Blane *et al.*, 2018). The benzamidine column specifically binds thrombin, while the IMAC column specifically binds the His tag and the uncleaved His tagged protein, while the cleaved protein passes through (***Figure 11***).

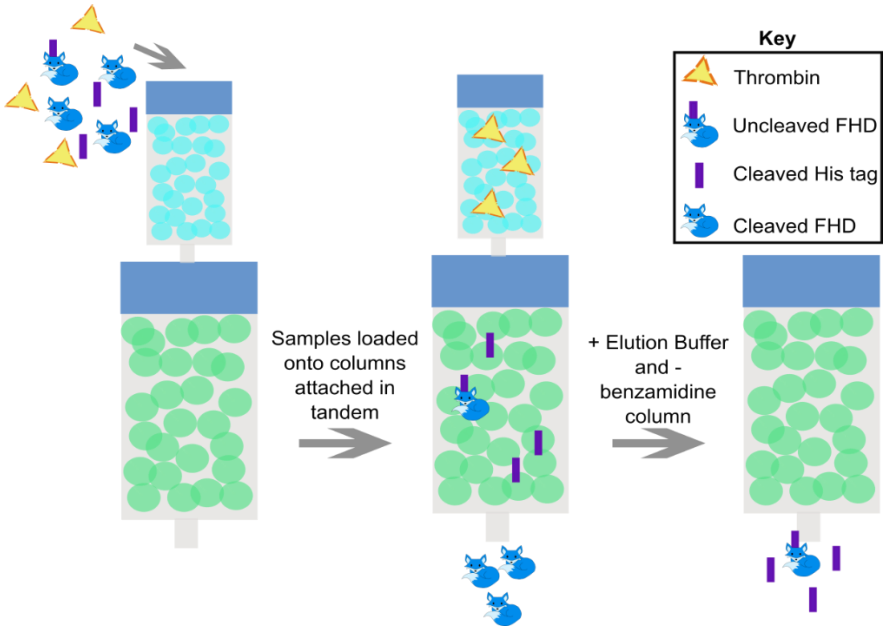


Figure 11: Separation of cleaved FOXP2 FHD from uncleaved FHD, His tags and thrombin
The blue resin indicates a benzamidine column, which binds to the thrombin in the solution. Green resin indicates the IMAC column with nickel attached. After cleavage, the FOXP2 solution is loaded onto the benzamidine column attached to the IMAC column (left), which are in equilibration buffer. The cleaved FHD passes through both columns and elutes, while thrombin and uncleaved FHD or His tags attach to their respective columns (middle). To remove uncleaved FHD, elution buffer is added after the benzamidine column has been removed.

Cleavage of the FHD was done as described in Blane and Fanucchi (2018). After IMAC purification, the fractions collected during IMAC which contained the FHD were pooled together and dialysed against cleavage buffer overnight. In order to remove any aggregates, the dialysed protein was centrifuged at 4 °C at 12000 x g for ten minutes. The supernatant containing soluble protein was then incubated in 1 unit of thrombin per mg of FOXP2 FHD for four hours at 20 °C. Every hour, a sample for 16% tricine SDS PAGE was taken in order to confirm cleavage had occurred.

A Ni^{2+} charged IMAC column was connected in tandem to a 1 ml HiTrap benzamidine (GE Healthcare, USA) fast flow column (GE Healthcare, USA) and equilibrated with equilibration buffer. Cleaved protein passed straight through the columns, but the His tag and uncleaved FHD only dissociated from the IMAC resin after elution buffer was added. Samples taken from the fractions collected of the cleaved and uncleaved peaks were resolved on a 16% tricine SDS PAGE to check for purity and cleavage.

4.6.4. SIZE EXCLUSION

Size exclusion chromatography is a technique which separates proteins by size. It makes use of a resin which consists of beads with small irregular spaces in them (Gámez-Valero *et al.*, 2016). Smaller proteins fit into these pores, while larger proteins do not, or fit into fewer of them. As a result, proteins with a higher hydrodynamic volume move faster through the column, as they essentially have less volume to cover compared to the smaller proteins (Gámez-Valero *et al.*, 2016). This is a useful technique to further purify a protein solution if it is contaminated by differently sized components.

As YY1 was not always sufficiently pure after IMAC, a second purification step was done using size exclusion chromatography to obtain suitably pure protein. A HiLoad Sephadex 16/600 200pg column (GE Heathlcare, USA) was used to purify YY1. The column was equilibrated with FHD equilibration buffer after which 2 ml protein was injected and any resulting peaks collected until one column volume (120 ml) of buffer had passed through the column.

Following each purification step, the fractions collected were resolved on a 12.5% glycine SDS PAGE gel to confirm their purity.

4.7. PROTEIN AND DNA CONCENTRATION DETERMINATION

Both YY1 and the FOXP2 FHD contain tyrosine and tryptophan residues, which are capable of absorbing light within the UV-Vis range (Lakowicz, 2006). Generally, pure proteins have a maximum absorbance at 280 nm, while aggregates scatter light at 340 nm and DNA absorbs at 260 nm (Lakowicz, 2006). The amount of light a protein absorbs is directly related to its concentration, as described by the Beer-Lambert law:

Equation 1: $A = \varepsilon c l$

Where A is the absorbance in absorbance units, ε is the extinction coefficient of the protein in $M^{-1}cm^{-1}$, c is the concentration of the protein in molar and l is the path length of the sample through which the light travels.

The extinction coefficient is calculated based on a protein's sequence, where the number of amino acids which absorbed light are used to determine how much the protein will absorb light at 280 nm per unit concentration per path length (Gasteiger *et al.*, 2005). The following equation is used:

Equation 2: $\varepsilon_{280} = (N_{Tyr} * 1490) + (N_{Trp} * 5500)$

Where N_{Try} and N_{trp} refer to the numbers of tyrosine and tryptophan residues in the protein, respectively; and 1490 and 5500 $M^{-1}cm^{-1}$are the extinction coefficients of each amino acid (Gasteiger *et al.*, 2005). Cystine residues involved in disulphide bonds also absorb some light, and can be added to the equation when not all cysteines are reduced (Gasteiger *et al.*, 2005). When working with YY1 and the FHD, however, the cysteines have generally been found to be reduced and therefore these residues were not included.

The absorbance spectra of each protein was measured between the wavelengths 240 and 350 nm using either a Thermo Scientific™ Nanodrop™ or a JASCO V-630 spectrophotometer. Proteins samples were measured in triplicate and subtracted from blank spectra containing only buffer. The 340 nm absorbance was subtracted from the 280 nm absorbance in order to discount any contributions to the further calculations from aggregates. After the triplicate readings were averaged, the Beer-lambert law was used to calculate the protein concentration, where the extinction coefficient for YY1 was 19940 while for the FOXP2 FHD it was 22460 $M^{-1}cm^{-1}$, as calculated by ExPASy ProtParam (Gasteiger *et al.*, 2005).

Lastly, the ratio between the 280 and 260 absorbance values was calculated to ensure that the contribution to the absorbance peak was from protein, and not partly because of DNA contamination. No buffers used in this process contained high concentrations of components such as imidazole, which can absorb light at 280 nm and thus interfere with accurately determining the protein concentration.

The concentration of DNA oligonucleotides was also determined using the same spectroscopic methods, with the appropriate extinction coefficients (**Table 6**)

As the YY1 absorbance spectrum had a low A_{280}: A_{260} ratio, to check for DNA contamination in the purified YY1 solution, a 1% agarose gel was used, as this would detect any large pieces of DNA in the sample. The gel was stained with a 1: 10 000 dilution of the DNA stain SybrGold DNA stain (Invitrogen, Life technologies).

Table 6: DNA oligonucleotide sequences used and the proteins to which they are specific.

DNA Name	Sequence	Specific protein	Base pairs	Extinction coefficient (260 nm) ($M^{-1}.cm^{-1}$)	Techniques
Nelson	5'-TTAGG**T**G**TTTAC**TTTCATAG-3' 3'-AATCC**ACAAATG**AAAGTATC-5	FOXP2	20	310960	EMSA
CNTNAP	5'-TTTGAT**TGTTTAC**TTTGTTC-3' 3'-AAACTA**ACAAATG**AAACAAG-5	FOXP2	20	310960	EMSA, Supershift EMSA, Intrinsic Fluorescence
AAV P5	5'-AGGGT**CTCCATTTT**GAAGCG-3' 3'-TCCCA**GAGGTAAAA**CTTCGC-5'	YY1	20	320048	EMSA, Supershift EMSA, Intrinsic fluorescence, fluorescence anisotropy, HDXMS
T-BOX Pal	5'- ATT**TCACACCT*AGGTGTGAA***AT-3' 3'- TA***AGTGTGGA***T**CCACACTT**TA-5'	TBR1	22	383828	EMSA

* Bold and underlined sites indicate the actual binding site or consensus sequence in the DNA.

4.8. DNA AND PROTEIN BINDING STUDIES OF YY1 AND FOXP2

4.8.1. DNA OLIGONUCLEOTIDES

Several DNA oligonucleotide sequences were used for this study, as it was necessary to use the appropriate cognate binding sequence for each protein in addition to a control for non-specific binding. For DNA specific to YY1, the AAV P5 promoter sequence that had previously been used in the crystal structure of YY1 DBD (Houbaviy *et al.*, 1996) was used (**Table 6**). This sequence belongs to the promoter site for transcriptional initiation of AAV proteins when entering the lytic cycle (Houbaviy *et al.*, 1996). YY1 binds to this as part of its role in repressing expression of this gene. Both single stranded (5' to 3') and double stranded DNA of this sequence was used.

The cognate sequence for FOXP2 FHD was the Nelson sequence (**Table 6**) (Nelson *et al.*, 2013), which has been used extensively in DNA binding studies with the protein (Blane and Fanucchi, 2018; Thulo *et al.*, 20210; Morris and Fanucchi, 2016), and has a high affinity to it. The Nelson sequence, however, contains random flanking regions to the sides of the core binding site which are not necessarily physiologically relevant.

Therefore, the promoter regions and upstream regions of several genes that FOXP2 is known to bind were scanned for any base pairs that matched with the core binding sequence of the Nelson sequence. From these, the CNTNAP2 promoter region contained several sections with the Nelson core binding site and some of its flanking region. Therefore, one sequence similar to the Nelson DNA was selected from this gene and used to determine FOXP2 binding on a more physiologically relevant sequence. This DNA was referred to as CNTNAP2 DNA. The CNTNAP2 promoter has been found to be directly regulated by FOXP2 in neural developmental pathways (Adam *et al.*, 2017).

Finally, a control DNA oligonucleotide was used that neither FOXP2 nor YY1 should be specific for: the T-box Palindrome (Pal) DNA sequence is specific for T-box domain containing transcription factors (Müller and Herrmann, 1997). TBR1 is one of the transcription factors which is a known binding partner of FOXP2 (Deriziotis *et al.*, 2014). As the T-box Pal sequence was very different from the FOXP2 or YY1 specific DNA sequences, it was used to confirm that each protein did have some level of specificity of DNA binding.

4.8.2. SUPERSHIFT ASSAY

A supershift assay was also done, where both YY1 and FOXP2 were added to the samples containing DNA specific for one of the proteins. This was to see the effect of the protein-protein interaction on the protein-DNA complexes. If the protein-DNA band shifted up or an extra, larger band was seen where the second protein was added, it could be surmised that a protein-protein-DNA complex was being formed. Alternatively, if no shift was seen, this would suggest that the protein-protein complex does not occur in the presence of DNA.

To observe the effect of the FHD on the YY1-DNA interaction, a constant concentration of 0.5 µM AAV P5 promoter DNA was used. Every four lanes contained increasing concentrations of YY1: 0, 1, 2.5 and 5 µM (**Table 7**). These were repeated four times, except the final repeat skipped the 5 µM concentration, due to limitation of lanes. Each set of YY1 repeats further had a fixed concentration of FHD – for the first, no FHD was added, then in the next set 2.5 µM was added to each lane, then 5 µM, then 10 µM.

Table 7: Concentrations of proteins and DNA used in a supershift gel to observe how FHD affects the specific YY1-DNA interaction.

Lane	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
[AAV P5 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
[YY1] (µM)	0	1	2.5	5	0	1	2.5	5	0	1	2.5	5	0	1	2.5
[FHD] (µM)	0	0	0	0	2.5	2.5	2.5	2.5	5	5	5	5	10	10	10

In order to determine the effect of YY1 on the FHD-DNA interaction, CNTNAP2 DNA was used, and FHD concentrations increased repeatedly every four lanes, at the same concentrations

as YY1 in the above setup. YY1 was instead added in increasing concentrations per set of increasing FHD – starting at 0 µM, then increasing to 2.5, 5 and 7.5 µM (**Table 8**).

A Bio-Rad (USA) PROTEAN^{TM} II xi Cell gel cassette was used to allow the gels more space and time to separate compared to normal EMSAs, as the gel was 16 x 16 cm. Twenty five microliters of sample was added to each lane, and the gel was resolved for either seven or eight hours at 160 V and four degrees Celsius.

Table 8: Concentrations of proteins and DNA used in a supershift gel to observe how YY1 affects the specific FHD-DNA interaction.

Lane	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
[CNTNAP2 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	1	2.5	5	0	1	2.5	5	0	1	2.5	5	0	1	2.5
[YY1] (µM)	0	0	0	0	2.5	2.5	2.5	2.5	5	5	5	5	7.5	7.5	7.5

4.8.3. PROTEIN LABELLING

4.8.3.1. NTA ATTO 550

For interaction studies using fluorescence anisotropy, YY1 needed to be labelled with a fluorophore. Because it contains both a His tag as well as an intrinsic histidine region, YY1 was labelled with NTA ATTO 550 (Sigma-Aldrich, RSA) (***Figure 12***), which is a fluorophore associated with Nitrilotriacetic acid (Shav-Tal *et al.*). Like with IMAC resin, the NTA interacts with histidine tags (Zhao *et al.*, 2010). Besides the complementarity of this label to the chemical groups of YY1 which made binding possible, ATTO 550 was selected for its properties as a fluorophore. With good photostability, a quantum yield of 0.8 and a fluorescent lifetime of 3.6

ns, it is able to provide a good anisotropy signal with precision (ATTO-Tech, ND). The drawback of this label was that it does not bind covalently to the protein, and the labelling reaction is, therefore, reversible. This was addressed by using freshly labelled YY1 for each binding assay, and also by using an alternative label to ensure consistent results.

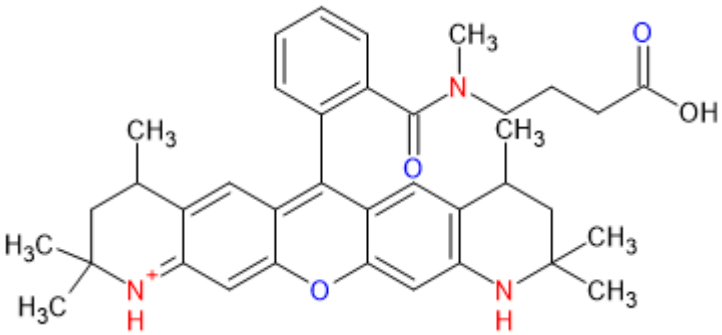


Figure 12: ATTO 550 structure.
The fluorescent label ATTO 550 has ideal properties for anisotropy studies, with a high quantum yield and good photostability. Figure generated in ChemSketch (Freeware) 2021.1.1.

For the labelling reaction, NTA ATTO 550 was added to a 10-15 µM solution in a molar ratio of 3 moles NTA ATTO 550 per 1 mole YY1. As NTA ATTO 550 was solubilised in dimethyl sulfoxide (DMSO), care was taken not to exceed 1% of the volume of protein when added, as higher percentages of DMSO could compromise the stability of YY1 (Chan *et al.*, 2017). The dye was left to bind YY1 overnight at 4 °C, after which it was centrifuged at 13 000 x g for thirteen minutes to remove any precipitant. Excess label was then removed using a Pierce™ Dye removal column (ThermoFisher Scientific, USA).

To determine the protein and the dye concentrations of the remaining NTA ATTO 550-bound YY1, an absorbance spectrum was taken for each. NTA ATTO 550 has an absorbance maximum at 554 nm, with an extinction coefficient of 120 000 ($M^{-1}.cm^{-1}$) (ATTO Tech, ND). To correct for the signal at 280 nm contributed by the label rather than the protein in the solution, the following equation was used:

Equation 3: $$A_{protein} = A_{280} - (A_{554}/10)$$

Where $A_{protein}$ is the actual absorbance at 280 nm contributed by the protein only, A_{280} is the total absorbance value at 280 nm as measured by the spectrophotometer and A_{554} is the maximum absorbance of NTA ATTO 550. The 554 nm absorbance is divided by a factor of

ten because this is the correction factor of ATTO 550 for wavelength of 280 nm – the dye is known to contribute to the 280 nm absorbance at an intensity that is ten percent of the dye's maximum absorbance. Therefore, that ten percent is subtracted to obtain the protein concentration. The labelling efficiency was then calculated by dividing A_{554} by $A_{protein}$, where zero would indicate no labelling occurred while one would indicate fully labelled protein.

4.8.3.2. TCP-YY1 ASCY3 LABELLING

In order to ensure that YY1 showed the same behaviour regardless of what label was used on it, a short tetracysteine peptide (TCP) tag (CCKAEAACC) was added to YY1 to create a TCP-YY1 construct. The TCP tag is a construct that specifically binds to biarsenical multiuse affinity probes such as AsCy3 (***Figure 13***), allowing specific and efficient labelling of proteins (Alexander and Schepartz, 2014; Hecht *et al.*, 2018). The TCP tag on YY1 was added near the N-terminus of YY1, immediately downstream of the N-terminal His tag followed by a GS linker, as adding it to the C terminal end may interfere with DNA binding.

AsCy3 is a biarsenical dye which forms disulphide bonds with the cysteines in the TCP tag (Alexander and Schepartz, 2014). These covalent bonds, once formed, are much less likely to dissociate again from the protein once bound. Compared to ATTO 550, AsCy3 has a lower quantum yield at 0.28 when bound to protein – when unbound, this value is lower by a factor of about 6 (Cao *et al.*, 2007). Furthermore, AsCy3 also has a lower fluorescence lifetime compared to ATTO 550, at about 1 ns when bound to protein and 0.3 ns when unbound (Hecht *et al.*, 2018). Therefore, it is a less suitable label for anisotropy, which requires long enough fluorescent lifetimes for an effective change in complex size to be apparent (Lakowicz, 2006).

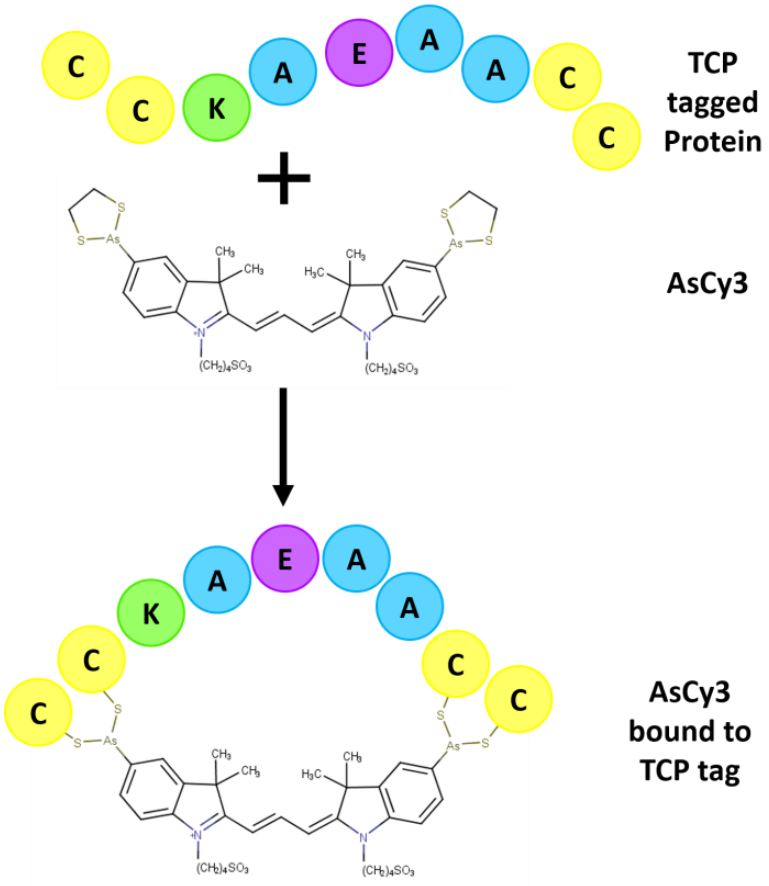


Figure 13: A schematic of AsCy3 labelling the TCP tag to label TCP-YY1.
AsCy3 is used to label a protein via the TCP tag. The AsCy3 and protein are reduced using β-mercaptoethanol and Tris (2–carboxyethyl) phosphine (TCEP). This allows for the AsCy3 to form disulphide bonds with the four cysteines in the TCP tag. Figure created using Chemsketch (Freeware) 2021.1.1. and PowerPoint (2013).

TCP-YY1 was labelled with ASCy3 by adding 10 μM AsCy3 to a mixture of 2 μM YY1, 1 mM TCEP, 1 mM β-mercaptoethanol, 20 mM HEPES and 100 mM NaCl, as described by Hecht *et al*. (2018). The labelling process was monitored by observing the fluorescence spectrum of AsCy3 when excited at 560 nm. When not bound to protein, it has a maximum emission wavelength of 568 nm, whereas this shits to 576 nm when bound to protein (Alexander and Schepartz, 2014). Therefore, the labelling reaction could be observed at one hour intervals to determine whether it occurred and if it was complete. As it was found to take 7-9 hours, labelling was done overnight at 4 °C. Because of the red shift in emission of AsCy3, the unbound dye was not removed from the solution, as it would not contribute to the anisotropy readings at 576 nm.

4.8.4. FLUORESCENCE ANISOTROPY

Fluorescence anisotropy is a useful technique used to determine the binding affinity of two molecules. It is a spectrophotometric method that relies on the concept of Brownian motion (Otto *et al.*, 1994). Molecules are not static in solution – they are constantly tumbling. Small molecules tend to be able to tumble faster than larger molecules. When polarised light of the relevant wavelength is passed through a solution of molecules labelled with a fluorophore, the excitation of those fluorophores will depend on whether each of them are in the correct orientation to absorb that photon (Otto *et al.*, 1994). The exited state of most fluorophores then lasts for about 1 to 10 ns, which is a much longer time period than small molecules, which rotate in the picosecond range (Lakowicz, 2006). Therefore, the fluorophore has usually changed orientations by the time it emits, resulting in the polarisation of the emission being different from what it would have been if the molecule were in its original exited position (Lakowicz, 2006). This results in low detection of emission of light at the corresponding parallel and perpendicular angles to the exited light. As such, the anisotropy of small molecules is low (Lakowicz, 2006).

The anisotropy value of a fluorophore is calculated in the following manner:

Equation 4: $$r = \frac{I\parallel - GI\perp}{I\parallel + 2GI\perp}$$

Where r is the anisotropy value, $I\parallel$ is the intensity of vertically polarised emission, $I\perp$ is horizontally polarised emission intensity and G is the correction factor of the instrument, which accounts for instrumental bias towards its own preferential detection of horizontally or vertically polarised light (Lakowicz, 2006). The equation assumes that the sample is excited using vertically polarised light (Lakowicz, 2006).

Macromolecules such as proteins have lower rotational diffusion, also in the nanosecond range, which results in an increase in anisotropy as it is more likely for emission to occur when the fluorophore is still in an appropriate orientation (Lakowicz, 2006). This technique can therefore be used in binding studies, where a small molecule such as an oligonucleotide can be labelled with a fluorophore and a DNA binding protein can be added to the solution in increasing

increments (Perez-Howard *et al.*, 1995). If the protein binds to the DNA on its own, the overall size of the complex will be much larger than the DNA, thereby increasing the anisotropy value because of the decrease in depolarisation (Perez-Howard *et al.*, 1995). By using increasing concentrations of the protein, the binding affinity can be determined by constructing a binding curve.

While EMSAs were used to determine if YY1 binds to its DNA, and could show the differently sized complexes, an estimation of the affinity with which YY1 binds DNA is not expected to be accurate from this technique. Therefore, fluorescence anisotropy was used to get a better estimation of the affinity between YY1 and its DNA, in order to better understand how YY1 regulates transcription via its DNA binding capacity. This was done in two ways: firstly, by titrating YY1 into labelled DNA, and secondly, by titrating DNA into labelled YY1.

Further fluorescence anisotropy was done to investigate both the homo- and hetero-association of YY1. For homo-oligomerisation studies, unlabelled YY1 was titrated into NTA ATTO 550 labelled YY1 at three different salt concentrations to observe the effects of ionic strength on its oligomerisation. Likewise, for hetero-association studies, NTA ATTO 550 labelled YY1 was used, while FOXP2 was titrated into the sample. This was also done at increasing ionic strengths, as well as in the presence of saturating amounts of AAV P5 DNA in order to see if the presence of DNA affected the protein-protein interaction. The low salt YY1-FHD curve was repeated with TCP-YY1 labelled with AsCy3 instead of NTA ATTO 550 in order to ensure reproducibility across labels. Additionally, during anisotropy experiments, fluorescence spectra were taken intermittently so as to ensure quenching was not causing too low a signal to reflect an accurate anisotropy reading.

All anisotropy studies were conducted on a Perkin Elmer LS-50B Luminescence spectrometer connected to a LKB Bromma (2219 Mulitemp II) water bath. The excitation and emission band widths were set at 5 nm, and an integration time of 3 seconds was used for all curves. The G factor was calculated for each sample before adding the titrant, the final volume added of which never exceeded 10% of the original sample volume, so as to prevent significant changes in concentration of the labelled molecule.

Rhodamine X (ROX) labelled AAV P5 DNA was used at a concentration of 1 μM, in order to obtain an appropriate signal intensity. Samples using this DNA were excited at 580 nm, and the emission wavelength measured was 507 nm, where the emission intensity was at its maximum. NTA ATTO 550 labelled YY1 samples were prepared at a concentration of 0.2 μM

labelled YY1, excited at 554 nm and the emission wavelength used was 576 nm. AsCy3 labelled samples were used at a concentration of 0.2 μM TCP-YY1, exited at 560 nm and their emission measured at 576 nm.

All binding curves were done in buffers consisting of 100 mM NaCl and 10 mM HEPES at pH 7.5. The YY1 into ATTO 550-YY1 and the FHD into ATTO 550-YY1 curves were repeated in additional concentrations of NaCl: 300 and 500 mM. Lastly, the FHD into DNA-ATTO 550-YY1 curve was done at a salt concentration of 100 mM and a DNA concentration of 10 μM.

All binding curves were analysed using the software SigmaPlot (version 12.0), where they were fit to the simplest ligand binding model available, for single site binding. This was because the actual stoichiometry of the binding events were unknown. In these cases, the simplest model with an appropriate fit is usually the best choice.

4.9. STRUCTURAL ANALYSIS

4.9.1. CIRCULAR DICHROISM

Circular dichroism makes use of circularly polarised light, which interacts with chiral molecules. As proteins are made up of chiral amino acids, the peptide bonds linking them together are able to absorb this light (Kelly *et al.*, 2005). Depending on the secondary structure, where specific backbone patterns such as helices are present, the protein will interact in a distinctive manner with the circularly polarised light (Kelly and Price, 2000). This leads to a distinct spectrum in the far UV range (at and below 240 nm) being produced for each secondary structural element – where, depending on the structural composition, each protein will have a characteristic spectrum (Berova *et al*., 2000). This can be used to calculate or to estimate the percentage of each secondary structure component present in the protein (Whitmore and Wallace, 2004).

The secondary structure of YY1 was observed using far UV CD in order to observe its secondary structural composition. The CD spectrum of TCP-YY1 was compared to YY1 in order to ensure the added tag did not change the secondary structural composition of the protein. All CD spectra were taken with 5 μM YY1 in 50 mM $NaSO_4$ and 10 mM HEPES, pH 7.5, except where otherwise stated.

Besides determining folding, CD was also used to compare structural changes of YY1 when bound to AAV P5 DNA. These spectra were taken using 10 μM YY1 and 10 μM DNA. To minimise noise, the samples were diluted in MilliQ immediately before reading were taken, resulting in a buffer concentration of 25 mM $NaSO_4$ and 5 mM HEPES, pH 7.5.

The CD spectrum of 10 μM FOXP2 FHD in 50 mM $NaSO_4$ and 10 mM HEPES, pH 7.5 was also observed to make sure that it remained the same as has previously been found (Morris and Fanucchi, 2016).

All far UV CD was performed on a Jasco 1500 circular dichroism spectrophotometer, at a temperature of 20 °C unless otherwise stated. Excitation and emission bandwidths were set to 5 nm. Both the ellipticity in mdeg and the HT voltage – which is used to monitor the levels of noise and quality of the data – were monitored for each sample.

When plotting and analysing the CD spectra, the millideg unit was converted into mean residue ellipticity [θ] (MRE) in order to facilitate comparisons between sets of data, as this is essentially a standardised measure of each residue in a protein. To do this, the following equation was used:

Equation 5: $MRE\ [\theta] = \theta/10Cnl$

Where θ is the measured ellipticity in millidegrees, *C* is the protein concentration in molarity, *n* is the number of residues and *l* is the path length.

4.9.2. INTRINSIC TRYPTOPHAN FLUORESCENCE

Protein fluorescence occurs when the electrons of the aromatic amino acids become exited by a particular wavelength of light. As the wavelength possesses the correct amount of energy to be absorbed by an electron, the electron is able to become exited and move up an energy level, having the energy to move further from the nucleus (Lakowicz, 2006). The exited electron then releases the energy again at a longer wavelength, moving back to the ground state. The energy emitted is referred to as fluorescence (Lakowicz *et al.*, 1983). While phenylalanine, tyrosine and tryptophan are all capable of fluorescence because of their aromatic rings, the intensity

emitted by tryptophan far exceeds that of the other two (Ghisaidoobe and Chung, 2014). Tryptophan is excited when exposed to light at 295 nm, and emits the light at wavelengths between 300 and 350 nm (Ghisaidoobe and Chung, 2014). The exact wavelength and intensity of emission depends on the polarity of the local environment of the tryptophan residue (Sultan *et al.*, 2006).

YY1 contains two tryptophan residues, one at position 205 and the other at position 224, both of which occur in the spacer region, which is expected to be unstructured and relatively solvent exposed. By investigating the wavelength at which YY1 emits light after excitation at 295 nm, it was possible to determine the environment of the residues (Sultan *et al.*, 2006). Both tyrosine and tryptophan residues can be excited at 280 nm, which further allowed YY1's tertiary structure to be probed on a more global level, as it contains six tyrosine residues located throughout the protein. If a residue occurs in a very hydrophobic environment, the wavelengths emitted will be shorter than if they are in a polar environment (Lakowicz *et al.*, 1983).

This technique was used to give some insight into the local tertiary structure in areas where the YY1 tryptophan and tyrosine residues are present. Any changes in the tertiary structure upon DNA binding was observed by comparing the spectra for YY1 in the presence and absence of 10 μM DNA, which was a saturating concentration. Two DNA sequences were used: P5 AAV DNA, which was specific to YY1, and CNTNAP2 DNA, which it did not bind to specifically. This was done in order to observe whether there was a difference in the fluorescence emission between specific and non-specific DNA interactions in YY1. Furthermore, YY1 fluorescence spectra were also taken in the presence of 6 M Guanidinium chloride, in order to compare its native emission to its denatured emission spectrum – unfolding caused by denaturation should cause a shift in the wavelength of emission, as well as the intensity.

The emission spectra of TCP YY1 and the wild type were also compared, to ensure both were similarly folded, and that the tag did not interfere with the folding of the protein. Fluorescence spectra were also taken for FHD, in order to compare with previous results to confirm that it was likely folded the same and therefore native. All fluorescence spectra were done in a buffer containing 100 mM NaCl and 10 mM HEPES, as well as the additives mentioned above when necessary. YY1 samples contained 5 or 10 μM YY1, while FHD was measured at a concentration of 2 μM.

A Jasco FP-6300 fluorimeter was used for all intrinsic fluorescence studies. All the samples were excited at 280 nm. Additionally, the DNA bound and APO samples were also excited at 295 nm. The excitation and emission bandwidths were set to 5 nm.

4.9.3. SIZE EXCLUSION CHROMATOGRAPHY

Size exclusion chromatography (SEC) is a technique used to investigate the quaternary structure of a protein and its hydrodynamic radius. It uses a beaded matrix to separate proteins according to their size and shape (Gámez-Valero *et al.*, 2016). The beads contain small pores inside them, into which smaller proteins can fit. As the larger proteins move through the larger pores between the beads, they elute from the column faster than the smaller proteins, which are retarded in the column due to becoming delayed in the small pores (Gámez-Valero *et al.*, 2016). By passing standards of known size through the column, a standard curve can be used to accurately determine the radius of the protein of unknown size, assuming that it is globular. Additionally, if a protein has more than one quaternary structure, it can be observed from more than one peak eluting.

Knowing the quaternary structure of a protein in isolation is important because it can give insight into how the protein functions. For example, a protein may form concentration dependant oligomers, or may form disulphide bonds between subunits either natively or non-natively (Merten *et al.*, 2012). In addition, binding to an interacting partner such as another protein or DNA could cause a change in the quaternary structure, or that interaction might only occur with one of the oligomeric forms (López-Perrote *et al.*, 2014). Therefore, quaternary structure is capable of heavily influencing a protein's ability to function.

Two chromatographic techniques are available for size exclusion – Fast Performance Liquid chromatography (FPLC), and High Performance Liquid Chromatography (HPLC). These two approaches differ mostly in terms of the pressure applied – FPLC has lower pressure, typically uses larger columns and quantities of proteins, and usually has lower resolution. HPLC, in comparison, operates at much higher pressures in a much smaller column, and is sensitive enough to use much lower concentrations. It does, however, have certain limitations, such as requiring very pure protein and, more importantly, usually needs to be done at a lower pH than 7.5. Both techniques were utilised in this study.

4.9.4.1. SE-FPLC

To determine the quaternary structure of YY1, a HiLoad® 16/600 Superdex® 200 pg size exclusion column was used, which can separate globular proteins in a size range between 10 and 600 kDa. The column was attached to an ÄKTA Prime, the pump which was kept at a constant flow rate of 1 ml/min. YY1 was centrifuged at 10 000 x g for twelve minutes to remove any aggregation before 1 - 2 ml were loaded onto the column, which was equilibrated in FHD equilibration buffer (**Table 4**). Eluted protein was collected in 2 ml fractions before samples of each were taken for analysis by SDS PAGE. The samples used included two different concentrations of YY1 (50 mM and 25 mM) in the presence and absence of 5 mM DTT to monitor any concentration-dependant or disulphide bonded oligomers. To compare these peaks with a standard size, a gel Filtration Cal Kit High Molecular Weight (Cytiva) was used. These were plotted into a standard curve from which the hydrodynamic volume of YY1 could be extrapolated, assuming it was globular in nature.

4.9.4.2. SE-HPLC

Size exclusion high performance liquid chromatography (SE-HPLC) was used to further observe any oligomeric changes of YY1 occurring at much lower concentrations than could be observed using SE-FPLC, as well as at differing concentrations of salt. A 4.6 mm x 300 mmYarra 3µm SEC-3000 SE-HPLC (Phenomenex, USA) column with a separation range of 6 to 700 kDa was used, connected to a SHIMADZU SP20A HPLC machine (Japan). The column was calibrated using a gel Filtration Cal Kit High Molecular Weight (Cytiva, USA). YY1 was resolved on the column at a range of concentrations (**Table 9**) in two buffers with different salt concentrations: 20 mM HEPES, pH 6.8 and either 300 or 500 mM NaCl. This was to observe the effect of ionic strength on oligomer formation. Absorbance values of the eluent were detected simultaneously at 254 and 280 nm and a flow rate of 0.35 ml/min was used.

Table 9: Samples of YY1 used in SE-HPLC for elution profiles at 300 and 500 mM NaCl

[NaCl] (mM)	Concentration YY1 (μM)						
300	2.25	4.5	10	15	20	24	29
500	2.25	4.5	10	15	20	-	-

4.10. YY1 DYNAMICS: HDXMS

Hydrogen deuterium exchange mass spectrometry (HDXMS) can be used to study the global dynamics of a protein, as well as the intactness of its secondary structure via its hydrogen bonding network. It can also be used to observe protein-ligand interactions where binding will cause a shift in the structural dynamics (Konermann *et al.*, 2011). While proteins are dynamic, some domains or stretches of residues tend to be more structured than others, leaving them less exposed to the solvent (Weis, 2016). Hydrogen atoms bound to electronegative atoms such as nitrogen, oxygen and sulphur are labile, allowing them to be exchanged with other hydrogens in the solvent (Konermann *et al.*, 2011). As a heavy isomer of hydrogen, deuterium can also be exchanged with those hydrogens when it is present in solution (Konermann *et al.*, 2011).

The rate of exchange (k_{ch}) between a hydrogen atom and a deuterium atom is dependent on several factors. Low temperatures decrease the exchange rate, and a pH of 2.5 minimises k_{ch} (Weis, 2016). As such, once the exchange has occurred, these conditions are employed to minimise back-exchange of deuterium to hydrogen during downstream processes (***Figure 14***). Additionally, the side chains flanking the relevant bond can influence the k_{ch}, and increases in pD can heavily increase it (Englander and Kallenbach, 1983). Across the entire protein, the actual rate of exchange is also influenced by hydrogen bonds and solvent exposure – where not only must a hydrogen be exposed to the solvent in order to be exchanged, but it must also be uninvolved in any hydrogen bonds, which have the same effect as burying the atom in an hydrophobic interior (Englander and Kallenbach, 1983). Because hydrogen bonds in the backbone are most involved in highly structured motifs, it is possible to indirectly probe the structure of a protein using HDXMS (Masson *et al.*, 2019).

The more exposed to the solvent a residue is, the more likely hydrogen is to be replaced with deuterium (Konermann *et al.*, 2011). Not all amino acid side chains contain an H-N, H-O or H-S bond; however, there is an N-H bond between every residue besides proline, making up part of the peptide backbone. Therefore, by monitoring the exchange of deuterium atoms in the backbone, the dynamics and structure of the entire protein can be probed by first dialysing the protein into deuterium, quenching the reaction, digesting the protein into peptides and then using mass spectrometry to detect which residues contain deuterium (***Figure 14***) (Masson *et al.*, 2019). Those that do would been more solvent exposed, while those that do not would be either more buried or involved in hydrogen bonds (Konermann *et al.*, 2011).

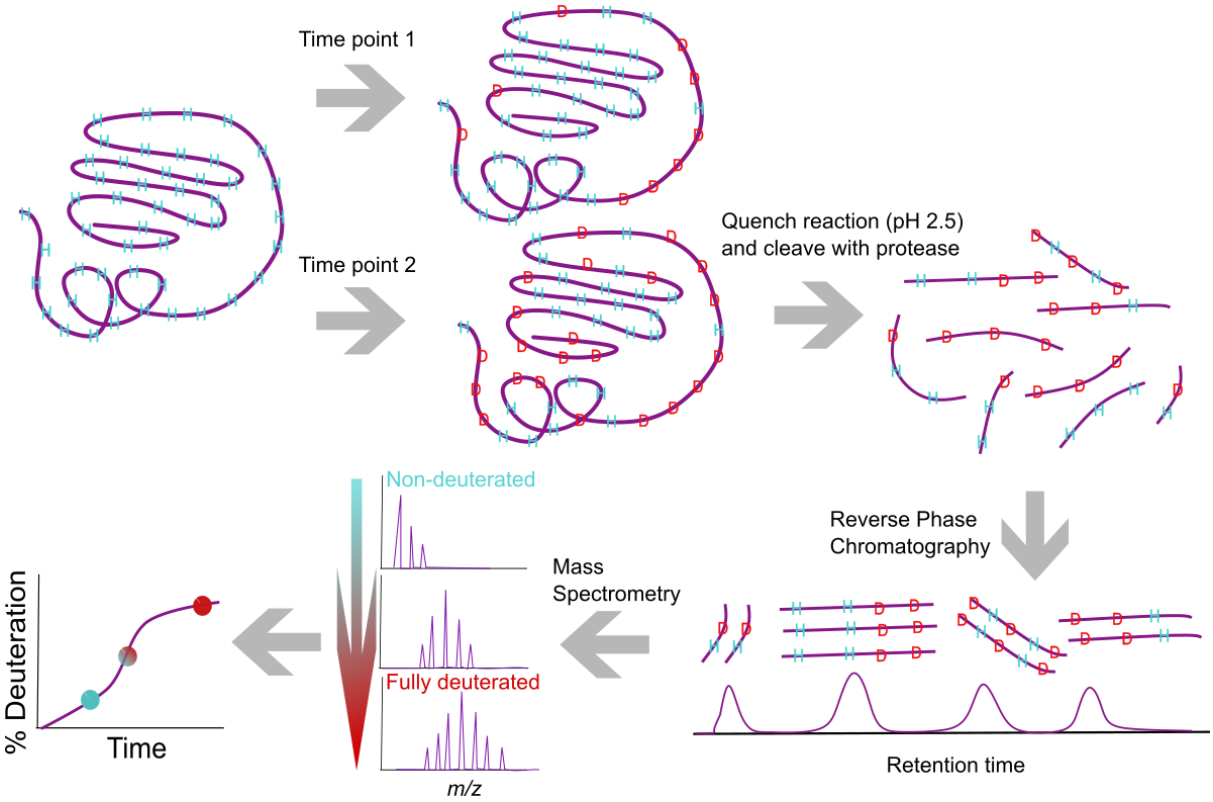


Figure 14: HDXMS workflow summary.
A protein sequence (purple) and its backbone hydrogens (blue) are exposed to deuterium for differing time points. The reaction is quenched at low pH 2.5 at 0 °C, then cleaved on a column with immobilised protease. LC-MS is done to determine the difference in deuterium uptake for each peptide at different time point, from which a deuterium uptake plot can be plotted. Image generated using InkScape 1.0.

While longer exposure to the solvent does allow for more hydrogen exchange, different regions of a protein will be affected less than others. Less accessible protein regions which are associated with high levels of structure do not maintain the same folded state all the time – they change conformations for periods of time, which can differ in length (Weis, 2016). This phenomenon can occur at a global or a local scale within the protein, where either small, select regions become open in conformation at a time, or very large portions (or the whole protein) unfold and become solvent accessible (***Figure 15***) (Weis, 2016). This results in the rate of overall deuterium exchange being dependent on the rates at which each particular conformation forms. The protein opens at a particular rate (k_{op}) to become solvent exposed, and closes at a particular rate (k_{cl}), while the exchange of hydrogen during the open, accessible state to deuterium also has a specific rate (k_{ch}) (***Figure 15***). Most proteins in their native state will tend to have a higher rate constant for forming a closed conformation compared to forming an open conformation (Weis, 2016). Therefore, the rate limiting step can depend on whether the conformational closing rate is greater or less than the rate of the deuterium exchange reaction.

If the protein is slower to enter its closed conformation compared to the time it takes for the exchange reaction to take place, then the hydrogens in each residue in that section of protein should be exchanged when it has entered the open conformation (Weis, 2016). Therefore, the overall rate of actual deuterium exchange k_{HX} can be determined by the equation below. This is referred to as EX1 kinetics (Weis, 2016).

Equation 6: $k_{HX} = k_{op}$

Conversely, if, as is more common, the closing rate exceeds that of the deuterium exchange rate, each opening event will result in only a portion of residues which have exchanged their hydrogens, as usually they will become inaccessible before having the chance (Weis, 2016). Therefore, the level of exchange will depend more heavily on how frequently the open conformation occurs, as seen in the below equation (Jensen and Rand, 2016). This is referred to as EX2 kinetics (Weis, 2016).

Equation 7: $k_{HX} = \frac{k_{op}}{k_{cl}} \times k_{ch} = K_{op} \times k_{ch}$

Therefore, adding separate time points for a protein's exposure to deuterium can demonstrate the dynamics which occur throughout the protein, as a change in conformation will usually result in a different level of exposure to deuterium. In this way, a region of the protein which is nearly almost always involved in hydrogen bonding of the backbone will show far less change over time compared to regions in the protein which are highly structure much of the time but have many short intervals where they lose that structure.

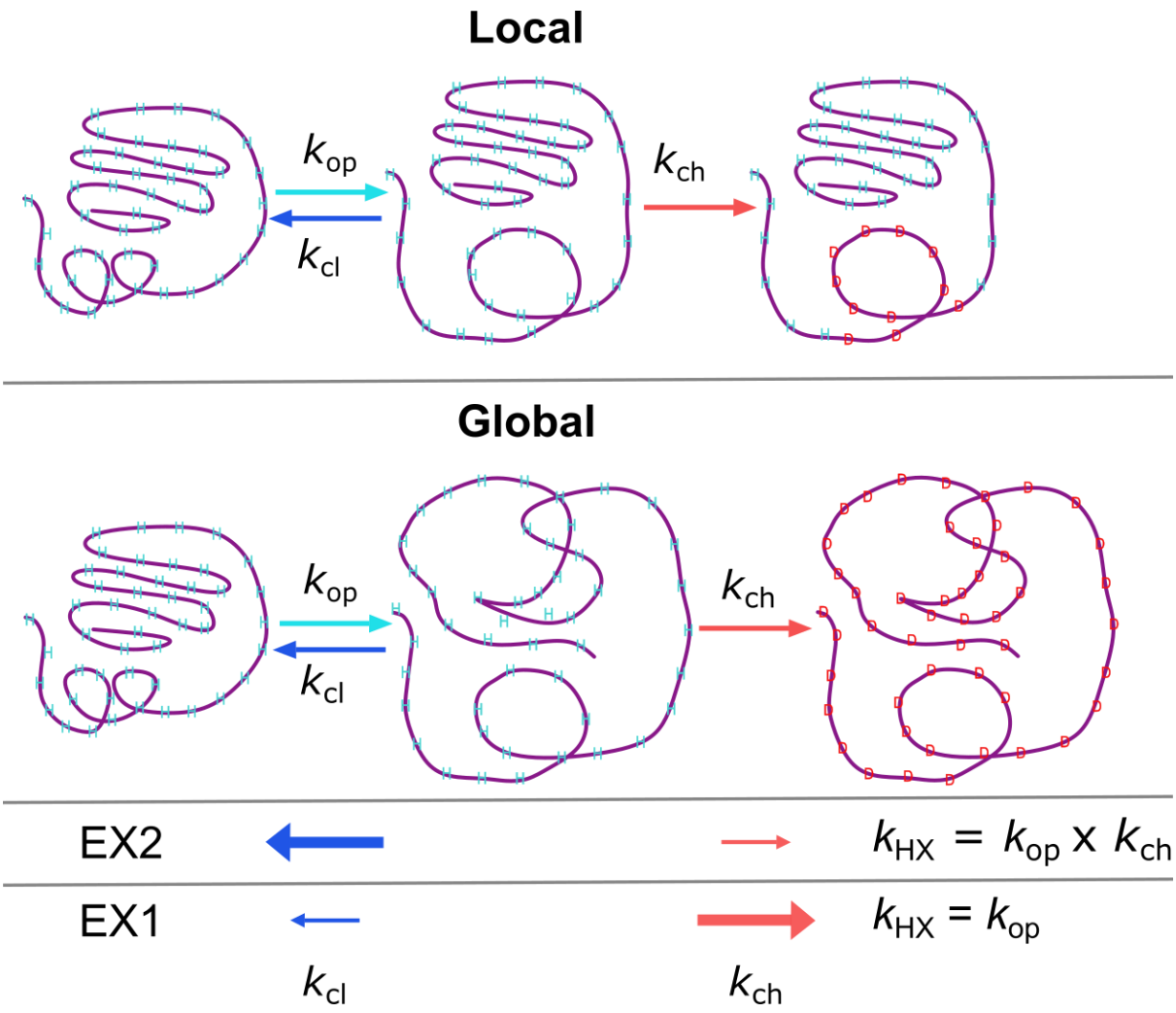


Figure 15: Kinetics of hydrogen-deuterium Exchange.
Local (top) and global (middle) unfolding mechanisms are shown with their associated steps of deuterium exchange. The protein conformation opens up and closes at rates k_{op} and k_{cl} respectively. Only the open conformation is available for the deuterium reaction, k_{ch}, to take place. The difference in the rates which cause EX1 and EX2 kinetics is shown at the bottom. This figure was created in InkScape 1.0.

Observing the dynamics of YY1 was done to provide an understanding of how structured the different domains are relative to each other. Knowing which regions are more flexible and which are less so will help to understand how each area of the protein may function; and their behaviour in solution can be compared to previous structural data. Furthermore, HDXMS was used on YY1 in the presence and absence of DNA, so the residues of YY1 that are involved in or affected by the interaction could potentially be identified.

4.10.1. SAMPLE ANALYSIS

HDXMS was done with the kind assistance of Drs Stoychev, Naicker and Buthelezi at the Council for Scientific and Industrial Research, Pretoria South Africa.

To allow for effective DNA binding, a low salt concentration buffer was used (10 mM HEPES, pH 7.5, 100 mM NaCl) to prepare samples. For DNA-bound YY1, samples were prepared at a 1:1 YY1:DNA ratio. Samples were diluted 8-fold in the same buffer which contained deuterium (or no deuterium but hydrogen instead, for the non-deuterated samples) to allow for deuterium exchange, which was then quenched by the addition of 3 M guanidinium chloride and 20 mM TCEP, pH 2.5 at 0 °C. Three time points were used to quench the samples: 15 s, 600 s and 3600 s. Additionally, fully deuterated samples were incubated overnight. After quenching the exchange reaction, YY1 was digested into peptides by passing it through a pepsin column (Applied Biosystems, USA) containing immobilised pepsin, which was stable in acidic conditions. These peptides were passed onto a Dionex (Thermo Scientific, USA) peptide trap (LC packing, ID:1.0 mm, phase C18PM) for desalting and then immediately separated using reverse phase liquid chromatography on a 50 x 2.1 mm C18 (Aeris Peptide 3.6 μm particle size) reverse phase column (Phenomenex, California, USA). Peptides were eluted using an acetonitrile gradient of 10-25% for 10 minutes at a flow rate of 200 μl/min before being analysed with mass spectrometry. The time of the processes after quenching were minimised, as back exchange does still occur, if at a much slower rate, at low pH and temperatures. The LC-MS process was done using an Agilent HPLC (California, USA) in tandem with a PALL Leap HDX robot (Leap Technologies, Florida, USA), with the above-mentioned chromatography columns contained in the 4 °C temperature controlled column compartment.

For mass spectrometry analysis, peptides separated with liquid chromatography were ionised in the gas phase and separated according to their mass in positive mode (**Table 10**). Non-deuterated samples were passed through MS1 and MS2 cycles, while deuterated samples were only subjected to MS1. All samples were analysed in triplicate, except for fully deuterated samples which were analysed in duplicate.

Table 10: Parameters used during Mass spectroscopic analysis of apo and DNA-bound YY1

Parameter	MS1	MS2
Cycle time (s)	3.5	3.5
No. samples	423	423
Ion source gas 1 (GS1)	30	30
Ion source gas 2 (GS2)	30	30
Curtain gas	30	30
Temperature (°C)	150	150
Ion spray voltage (V)	5500	5500
Charge state envelop range (Da)	300-1100 Da	100-1800
Accumulation time (s)	0.499	0.1
Time bins sum	4	8
Quadrupole 1 and 2 delustering potential (DP)	60	
Q1 Collision energy (CE)	10	

4.10.2. DATA ANALYSIS

Non-deuterated samples were used to identify peptides and create a peptide pool for further analysis of deuterated samples. PEAKS Studio version 6 was used to identify peptide peaks and their associated properties, including their sequences, retention times, charge information and peptide quality scores. This generated a peptide source which was imported into the analysis program HD Examiner (Sierra Analytics, USA) to create the peptide pool from which deuteration uptake was calculated after importing the MS1 data generated by the deuterated time points, as well as the relevant labelling times.

From the peptide pool and the MS1 data, the HDX-aminer software package (Sierra Analytics, USA) identified the relevant peaks for each peptide in each sample, and calculated the deuterium uptake of each time point based on the centroids of each peptide. As each centroid value is an indicator of the mass of a peptide based on its collective of isotopes, the non-deuterated centroids could be subtracted from each of the deuterated centroids to determine the average number of deuterium atoms per sample per time point. This was converted to percentage deuterium uptake, where the fully deuterated sample was 100% deuterated and was further used to account for deuterium back-exchange in each of the time points. To account for the back exchange, HD Examiner uses the following equation as described by Zhang and Smith (1993):

$$\textbf{Equation 8: } D = \frac{M_t - M_{0\%D}}{M_{100\%D} - M_{0\%D}} \times N$$

D indicates the actual number of deuterium atoms. M_t denotes the average mass of the peptide at the time point currently being calculated for. $M_{0\%D}$ and $M_{100\%D}$ are the non-deuterated peptide and the fully deuterated peptide, respectively. N is the number of exchangeable backbone hydrogens present in the peptide – this excludes proline residues, which don not have exchangeable hydrogens, and also the N-terminal of the peptide, which is much more prone to back-exchange compared with other residues.

Each peptide was manually validated by adjusting retention times when needed, adjusting the scope of m/z plots was taken into account to minimise noise, and setting outliers or poor quality time points with very low confidence scores to low confidence, which excluded them from all calculations.

From the percentage deuteration of each peptide, a deuterium uptake plot was generated, and from these plots, a heat map was generated. The heat map was calculated by subtracting overlapping peptide's deuterium uptake from each other, resulting in a higher resolution for deuterium uptake compared to each whole peptide. Depending on the level of overlap of the peptides, some single residue's deuterium uptake could be determined.

RESULTS

To understand how transcription factors such as YY1 and FOXP2 influence gene expression, it is necessary to understand how they work in isolation and in the presence of their interacting partners. This includes obtaining detailed knowledge regarding structure, which influences the resulting functions. It also includes characterising interactions between molecules: protein-protein, protein-DNA and protein-protein-DNA (If the latter exists). Therefore, both YY1 and FOXP2 were expressed, purified and characterised. The influence of DNA binding on the whole structure of YY1 as well as on the dynamics of the protein was investigated. Following this, the *in vitro* interaction between the FOXP2 FHD and YY1 was investigated to determine what role this may play in gene expression.

5.1. PROTEIN EXPRESSION

5.1.1. YY1

To the author's knowledge, YY1 had not previously been expressed natively in the pET-11a vector. It was, therefore, necessary to determine the optimal conditions for its expression. Several concentrations of IPTG were tested at 37 °C over 24 hours (***Figure 16***). No apparent YY1 overexpression could be seen when no IPTG was added, indicating that the induction system was tightly controlled. Overexpression was induced well at the lowest concentration of IPTG (0.2 mM), and did not show any apparent increase in efficiency in the samples induced at higher concentrations. Therefore, further expression trials were induced at 0.2 mM IPTG. At 37 °C, YY1 was overexpressed in large amounts as soon as three hours after induction. However, almost all of the protein was found to be insoluble. Therefore, the temperature was lowered to 20 °C before induction, in order to slow down expression and allow more time for the protein to fold. The lowered temperature was found to greatly increase the amount of soluble overexpressed YY1, with an optimal time of 20 hours (***Figure 16***). As less nutrients in the media can also sometimes be used to lower expression rate, LB media was tested as well as 2x YT media. Expression was higher in the more nutritious 2x YT media compared with the

LB media. Lastly, the amount of ampicillin used was decreased to test whether it was possible to use lower than typical amounts for previous expressions such as the FHD. It did not seem to make a significant difference in expression. Therefore, YY1 was expressed at large scale conditions at 20 °C for 20 hours after induction via the addition of 0.2 mM IPTG and 0.1 mM $ZnCl_2$ in 2x YT media containing 0.05 mg/ml ampicillin.

After overexpression, YY1 was purified using IMAC (***Figure 16***). Two challenges were encountered during this process: firstly YY1 attached to the column so strongly it was difficult to elute, and secondly it was frequently contaminated with other proteins. The former problem was addressed by incubating the bound protein with 800 mM imidazole for long periods of time (45 minutes-1 hour), allowing more YY1 to be eluted at once rather than in one long tail. This batch elution method allows more time for the imidazole to make contacts with the cobalt, giving it time to form an equilibrium and compete with the bound protein. To an extent, this solution to problem 1 also addressed problem 2 in that later peaks were usually cleaner than earlier elution peaks because less specifically bound contaminants eluted more easily. To further reduce contamination, additional steps were taken using washes containing detergent, high salt and imidazole, to disrupt nonspecific interaction between contaminants bound to either the column or YY1. While this did help to a large extent, a second purification step was still required for several fractions of each IMAC yield.

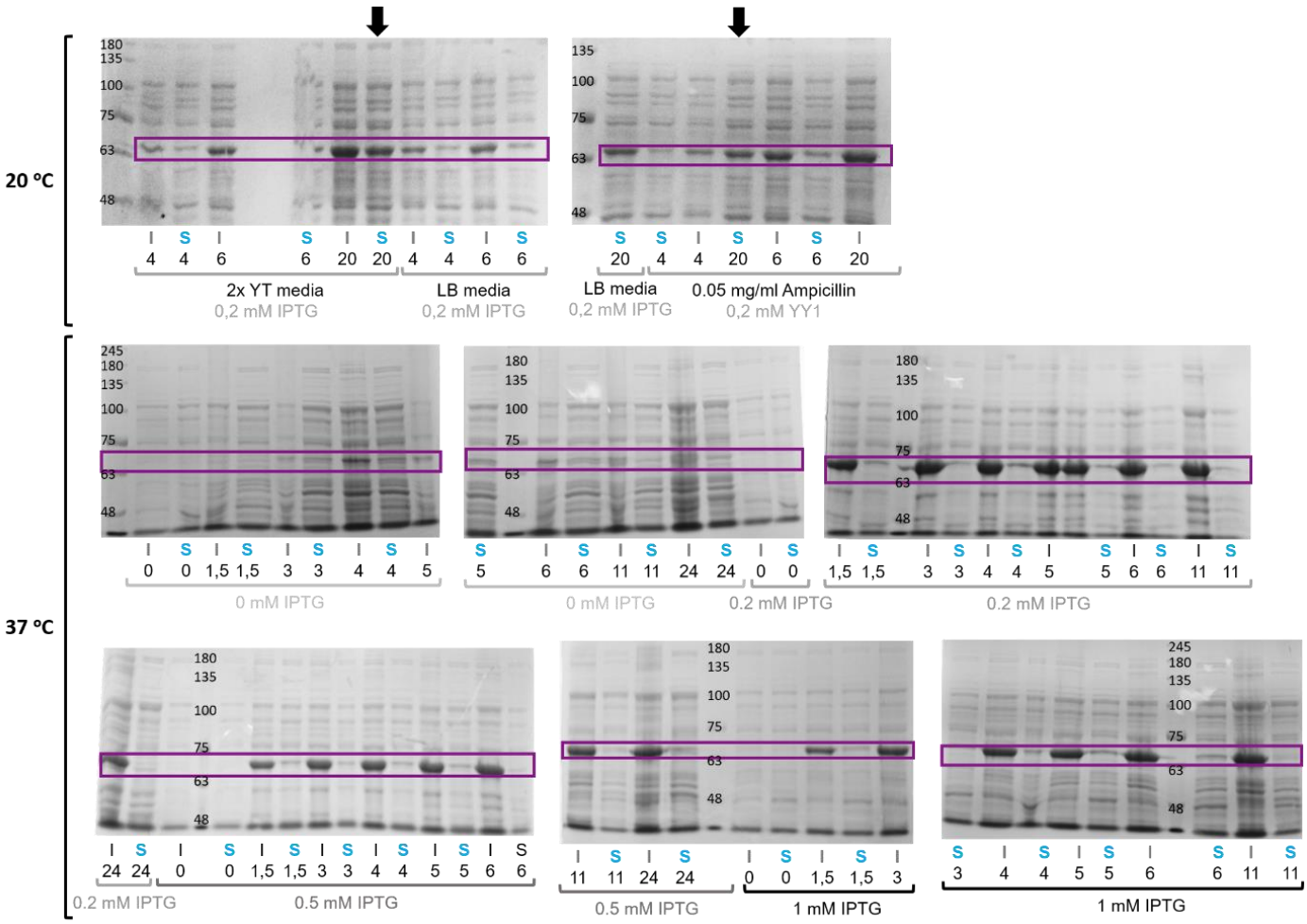


Figure 16: Expression Trials of YY1 at 20 and 30 °C.
The purple box shows the location of YY1 bands. I indicates the insoluble fraction of expressed protein while S indicates the soluble fraction. Numbers below **I** and **S** are the number of hours the cells were allowed to grow after induction. Black arrows show the preferred conditions (20 °C, 0.2 mM IPTG, 2x YT media) which express an acceptable level of soluble YY1. Image created using PowerPoint (2013).

Size exclusion chromatography was, therefore, used preparatively to further purify YY1 fractions which contained contaminants. Most fractions eluted from SEC did have improved purity, with some more so than others (***Figure 17***). YY1 eluted in two peaks: one large one within the void volume, and one much smaller one after this. When resolved on a gel, fractions collected from both peaks contained YY1, usually with similar levels of purity. To ensure that the much larger sized peak was not an amalgamation of aggregates, fractions from each one were kept separate and checked for activity using an EMSA (***Figure 21***). Both were found to be active, and in cases when the protein from the separate peaks was passed through the size

exclusion column again, the same phenomenon was observed, suggesting that YY1 exists in an equilibrium of oligomeric forms at concentrations ranging from about 15 to 60 µM.

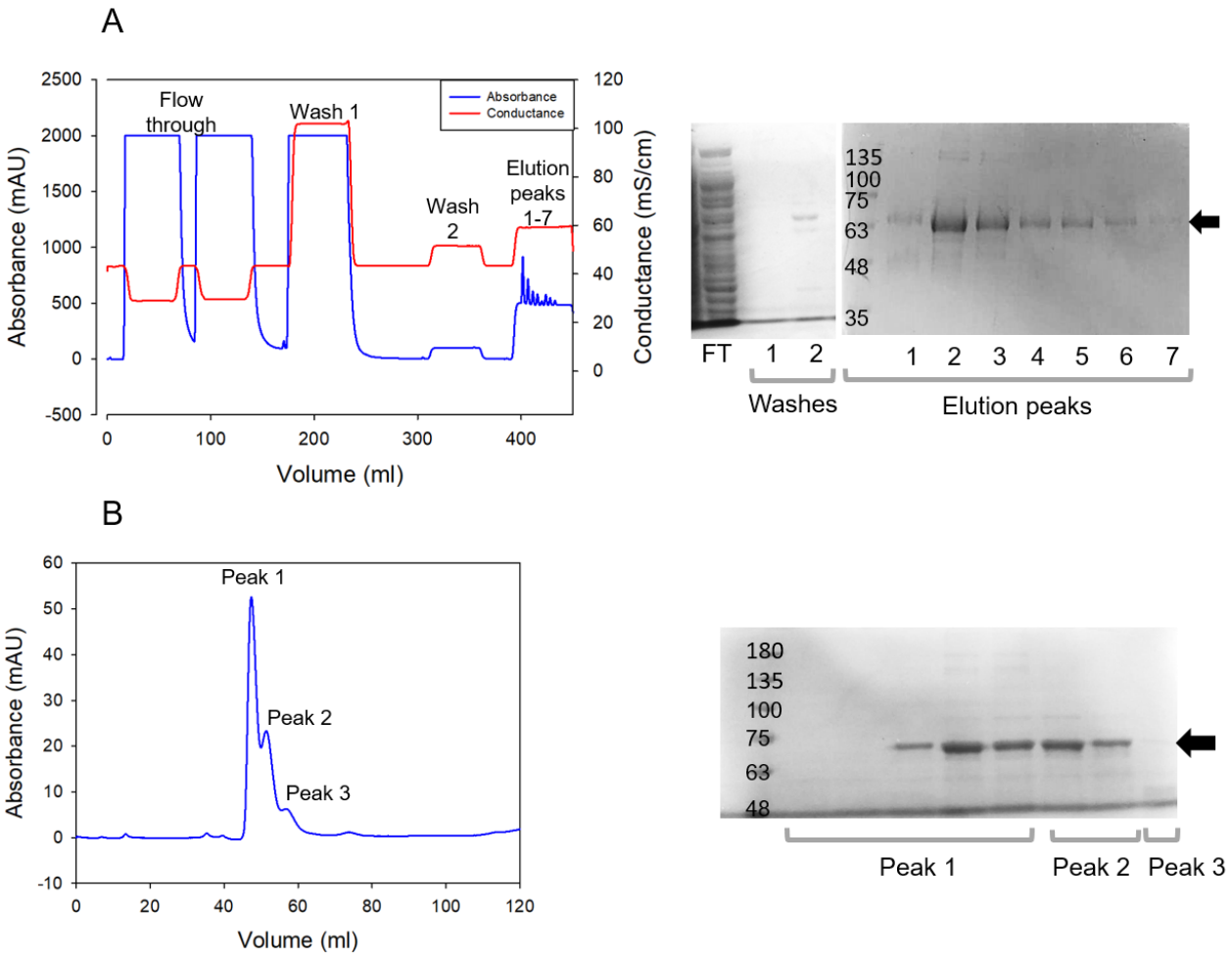


Figure 17: Purification profiles and corresponding SDS PAGE gels for YY1.
A) The YY1 IMAC profile on the left, with the UV absorbance at 280 nm in blue and conductance in mS/cm in red. The flow through is split into two peaks due to a pause in loading the cell lysate supernatant onto the IMAC column. The Detergent-salt wash follows, and is wash 1 on the gel to the right. Each elution peak in the elution step was collected and corresponds to the elution peaks on the gel. **B)** The size exclusion profile of contaminated YY1 IMAC fractions (for example peaks 2 and 3 in gel A) and the corresponding fractions on the gel on the right. The molecular weight marker marker used was BLUeye Prestained Ladder (Sigma-Aldrich, RSA). Graphs were plotted using SigmaPlot 12.0.

Fractions of YY1 resolved on SDS PAGE were calculated to represent a size of about 66 kDa, although YY1's predicted molecular weight is about 45 kDa. This is in line with previous findings in the literature, where it has generally been resolved between 65 and 68 kDa on SDS-PAGE and this is due to the sequence of YY1 (Shi *et al.*, 1997; Thomas and Seto, 1999). Much

of the YY1 sequence is typical of an intrinsically disordered protein, where it tends to contain a higher amount of charged and polar amino acids and fewer hydrophobic residues compared to globular proteins (Górecki *et al.*, 2015). SDS interacts with proteins via hydrophobic interactions, so in the case of many intrinsically disordered proteins, it binds at a lower ratio than one would find in a globular protein (Uversky and Dunker, 2010). This lower amount of SDS binding results in a lower mass-charge ratio than one would observe in the molecular weight marker standards, which results in slower migration through the gel and a larger apparent molecular weight (Uversky and Dunker, 2010).

5.1.2. FHD

Unlike YY1, FHD expression and purification is already an established protocol in the laboratory where this research was done (Blane and Fanucchi, 2015; Morris and Fanucchi, 2016). Therefore, no optimisation was required for expression or purification of this protein. The FOXP2 FHD was overexpressed when induced at 20 °C with 0.5 mM IPTG, and was also purified using IMAC. This resulted in a good yield of FHD albeit with some contamination (***Figure 18***). Because the FHD His tag was cleaved after this, a second round of purification was performed using a benzamidine column and an IMAC column in tandem. This not only separated the cleaved protein from the uncleaved portion, but also removed the contamination after the initial IMAC (***Figure 18***). This process resulted in pure FHD which could be used in downstream applications.

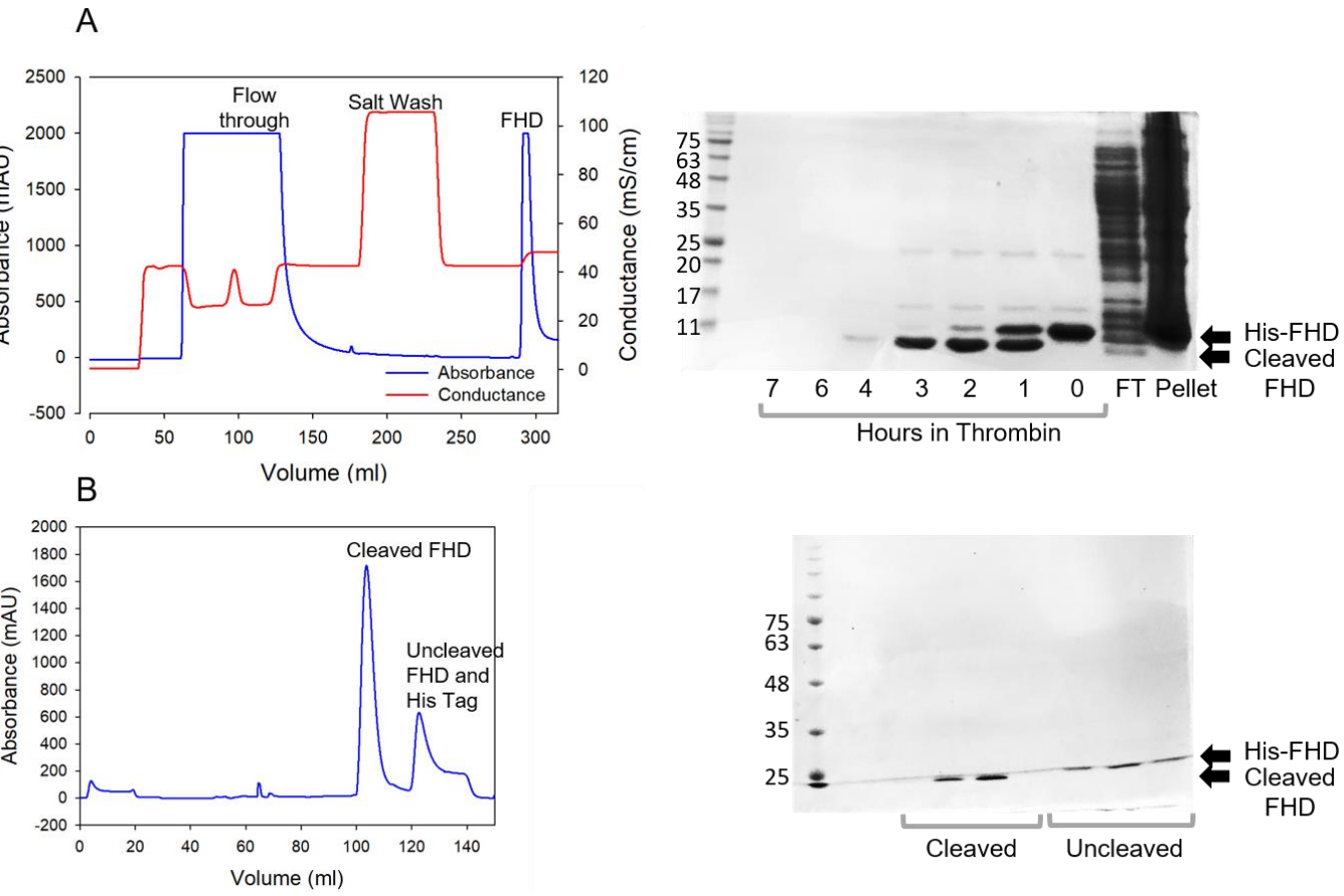


Figure 18: FHD purification elution profiles and corresponding gels.
A) The IMAC elution profile on the left, where the blue line indicates the absorbance while the red line indicates conductance. The flow through was collected before a high salt wash was used to remove lingering DNA contamination. The FHD was then eluted in 300 mM imidazole, and the resulting fraction of the protein can be seen on the gel on the right, labelled zero hours. After IMAC, the fractions were incubated with thrombin after dialysis against the correct buffer. **B)** The elution profile of the separation of cleaved and His-tagged FHD using a benzamidine column attached to an IMAC column. To the right is the gel showing the samples collected from the two peaks. The molecular weight marker marker used was BLUeye Prestained Ladder (Sigma-Aldrich, RSA). Graphs plotted with SigmaPlot 12.0.

5.2. PROTEIN CONCENTRATION AND DNA CONTAMINATION

After purification, each protein's absorbance spectrum was measured. Firstly, this was to determine the concentration. Secondly, it was to ensure that the protein had a typical absorbance spectrum of a pure protein. If it showed a peak at 260 nm or increased absorbance at 340 nm, it could indicate DNA contamination or aggregation, respectively. This was especially important to ascertain because both YY1 and the FOXP2 FHD are transcription factors which bind to DNA (Houbaviy *et al.*, 1996; Stroud *et al.*). Therefore, DNA contamination is likely and would interfere with any DNA binding experiments.

The absorbance spectrum of YY1 does not have a very clear peak at 280 nm, compared to the absorbance at 260 nm (***Figure 19***). This usually indicates the presence of DNA contamination. However, Purified YY1 was resolved on an agarose gel at 100 µM, and stained for DNA. No bands in the YY1 samples were seen, eliminating the possibility of large DNA molecules contaminating the YY1 (***Figure 20***). This was done again on an acrylamide gel, Once again, there was no trace of DNA contamination on this sample (***Figure 20***). Furthermore, unfolding and refolding YY1, size excluding it at 2 M NaCl and purifying it on a heparin column had no effect on the YY1 spectrum (data not shown), which decreases the likelihood of DNA contamination as a cause.

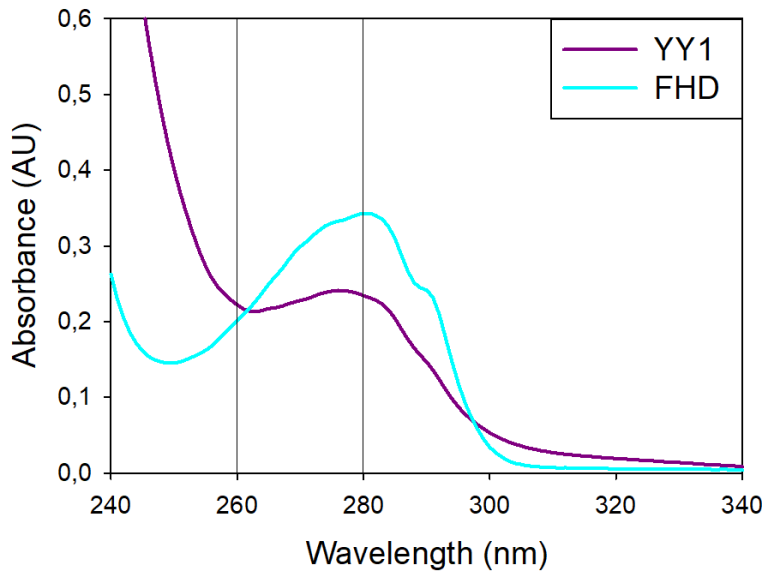


Figure 19: UV absorbance spectra of YY1 (purple) and the FOXP2 FHD (blue).
The vertical lines mark the positions of 260 and 280 nm. The A280:A260 ratios were 1.1 and 1.7 for YY1 and the FOXP2 FHD, respectively. Image was created using SigmaPlot 12.0.

Given the lack of DNA in the samples resolved on the gels, it is much more likely that the YY1 absorbance spectrum's low 280/260 ratio results at least in part from its low extinction coefficient, which is 19940 $M^{-1}cm^{-1}$ under reducing conditions. Additionally, YY1 contains eleven phenylalanine residues, which is many more than the number of tryptophan (2) and tyrosine residues (6) put together. While phenylalanine does not have a strong absorbance intensity, it does still contribute to the 260 nm absorbance value, and could, therefore, be one of the contributing factors (Lakowicz, 2006).

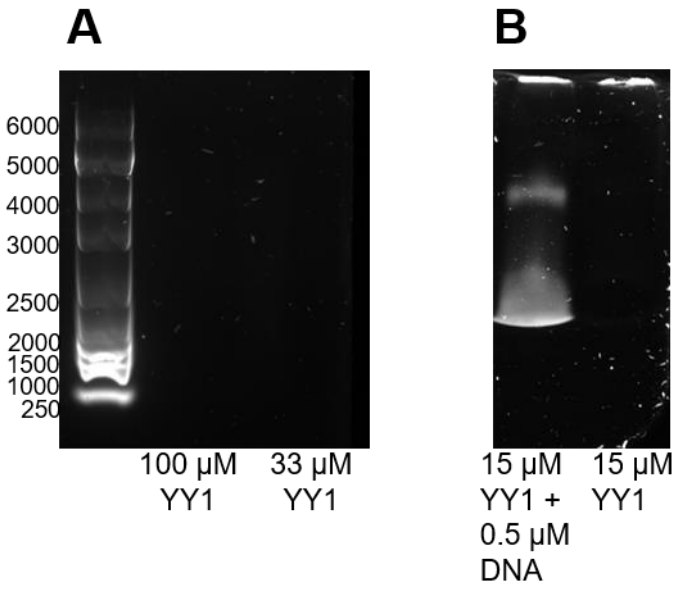


Figure 20: DNA gels contain YY1 to check for DNA contamination.
A) Agarose gel with DNA stained to check YY1 for DNA contamination. The O'GeneRuler 1 kb DNA Ladder (Thermo fisher scientific, EU) molecular weight marker is on the left-most lane, while YY1 samples are on the right. **B**) Left lane shows YY1 binding to AAV P5 promoter DNA, while right lane shows YY1 in the absence of added DNA, where no DNA is detected. Image made using Powerpoint (2013).

The FOXP2 FHD, on the other hand, had a far more conventional absorbance spectrum (***Figure 19***). Its peak absorbance was at 280 nm, and had an appropriate 260/280 value of 0.6, indicating that it likely did not contain DNA contamination and that it also behaved in the same manner as has previously been found (Blane and Fanucchi, 2015).

5.3. DNA BINDING STUDIES

The activity and specificity of both YY1 and FOXP2 FHD were tested using EMSAs. This was important to ascertain correct folding and functionality of each protein, and also to gain a better understanding of how they bind to DNA. The ability to bind a sequence associated with another transcription factor can be indicative of how these transcription factors might be related in their functions.

5.3.1. YY1 DNA BINDING

5.3.1.1. YY1 ACTIVITY AND DNA SPECIFICITY

YY1 was found to be active as it bound to its specific DNA sequence, the AAV P5 promoter DNA (***Figure 21 A***). A shifted band belonging to the YY1-DNA complex could be clearly seen on the EMSA gel. This shifted band increased in intensity as the concentration of YY1 increased, until a DNA to YY1 ratio of about 1:10, at which point YY1 was at a concentration of 5 μM and it appeared to have reached equilibrium with the DNA. At this and higher concentrations, there was very little unbound DNA, indicated by the faintness of the free DNA band. At lower concentrations of up to 2.5 μM of YY1, both a smaller and a larger YY1-DNA complex could be observed, although the former was much fainter. This suggests that YY1 may be able to bind to its DNA in several different oligomeric forms. As the protein to DNA ratio was increased further, this lower band was no longer apparent, indicating that the larger complex was more favoured at high YY1 concentrations. Even larger complexes which had not yet become visible on the gel in ***Figure 21*** were seen to form in other YY1- DNA EMSAs which had been run for longer, as well as the supershift assay, which is shown in ***Figure 39***.

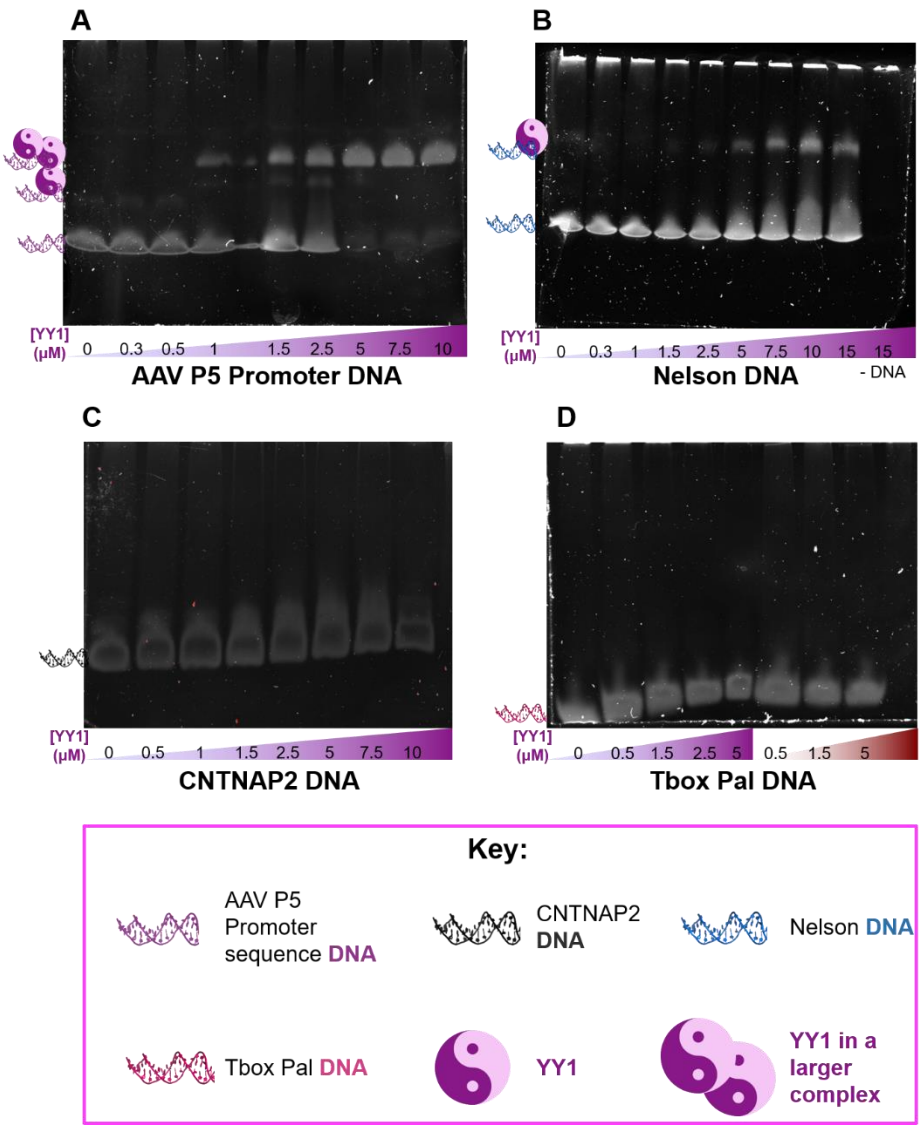


Figure 21: EMSAs of YY1 with several different DNA sequences.
A) Shows increasing YY1 concentrations per lane with a constant amount of AAV P5 Promoter DNA. **B**, **C** and **D)** show the same but with Nelson, CNTNAP2 and T-box Pal DNA sequences, respectively. Purple gradient triangles represent increasing YY1 concentrations, while red symbolises TCP-YY1. Symbols for the DNA and protein are shown at the bottom. Images made in InkScape 12.0 and Powerpoint (2013).

EMSAs were also performed to test the specificity with which YY1 binds to DNA. This was especially important for the FOXP2 FHD specific DNA sequences, because if both proteins bound these sequences, it would be more difficult to do studies involving YY1, FOXP2 FHD and DNA simultaneously (***Figure 22***). YY1 was found to bind the Nelson DNA sequence (specific to FOXP2), albeit at a lower affinity compared to its own AAV P5 sequence (***Figure 22***). As the Nelson sequence does not contain the consensus binding sequence belonging to

YY1, this was very likely a form of non-specific binding. In contrast, YY1 did not bind to the CNTNAP2 DNA sequence, which contains the same core binding sequence as the Nelson sequence, but different flanking regions. Therefore, it is very likely that YY1's interaction with the Nelson sequence results from the DNA flanking regions. The last sequence tested was T-box Pal DNA, which was used as a negative control to ensure YY1 specificity. Like the CNTNAP2 sequence, this DNA did not bind to YY1 (***Figure 22***). This confirmed that YY1 does bind to DNA specifically, although it is capable of non-specific interactions.

Aside from comparing YY1 to other DNA sequences, it was important to ensure that the protein was capable of activity after each process it underwent. Firstly, it was necessary to ensure that the different oligomeric complexes of YY1 purified with SEC were all active, rather than incorrectly folded. Therefore, peaks one and two which were confirmed to contain YY1 with SDS PAGE were compared at the same concentrations for their capacity to bind DNA (***Figure 22*** A). Both the larger and the smaller oligomeric forms of YY1 displayed DNA binding, indicating that both forms of the protein were active. The second YY1 peak eluted from SEC did have a higher intensity of bands, which may suggest a better level of activity. It is also possible, however, that the much larger oligomer had more DNA binding to it and could not enter the gel due to the size of the complex, as seen in section 5.8. Further studies of the isolated YY1 SEC peaks and their ability to re-equilibrate were done and are addressed in section 5.5.

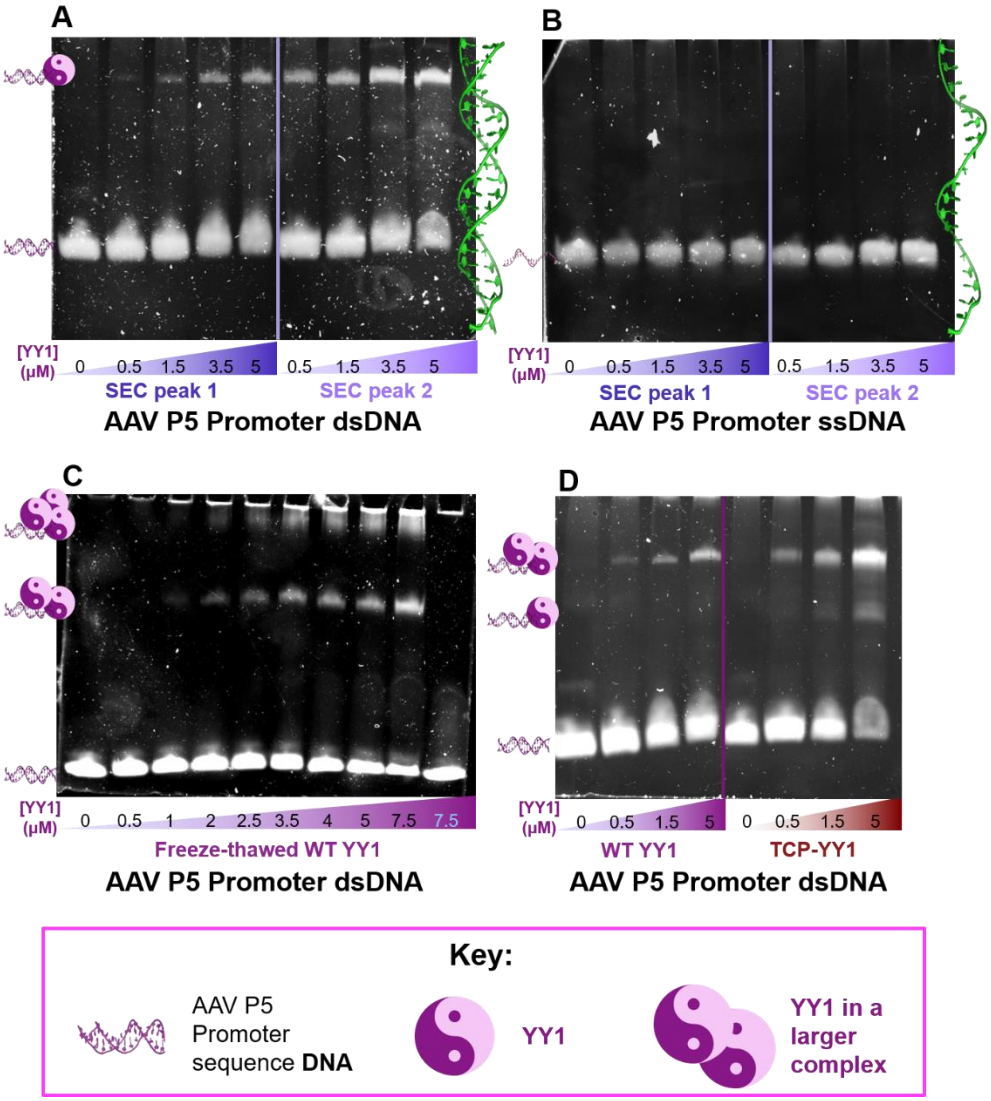


Figure 22: A comparison of YY1 binding for oligomers, single stranded DNA, Freeze thawed and TCP-tagged YY1.
A) Shows an EMSA where YY1 that eluted in the different SEC peaks that was collected during purification was incubated with double stranded AAV P5 DNA. **B)** Shows the same as **A**, except with single stranded DNA as opposed to double stranded. **C)** Shows freeze-thawed YY1 and AAV P5 DNA, except for the last lane, which is thermally melted protein which has been brought back to room temperature. **D)** Is an EMSA comparing the activity of WT YY1 and TCP tagged YY1. Images created using Pymol (pdb 1UBD), Inkscape 12.0 and Powerpoint (2013).

EMSAs of the separate size excluded elution peaks of YY1 were performed using both double and single stranded YY1 DNA. Although YY1 has previously been found to bind ssDNA, it did not bind to the single stranded 5'-3' AAV P5 promoter sequence here (***Figure 22***). This

suggests that its single stranded sequence binding requirements may differ from those of dsDNA.

Over time, YY1 was found to lose its activity when stored at 4 °C. Golebiowski *et al*. (2011) encountered the same problem at 4 and -20 °C, but found that YY1 was stable when stored at -80 °C. Therefore, YY1 was stored at -80 °C when it would not be used for several weeks. To make sure that it was in fact stable after the freeze-thaw process, an EMSA was done to ensure it retained its activity (***Figure 22***) which it did, as its activity remained at similar levels to the protein which was not frozen. This EMSA's free DNA migrated to the very bottom of the gel, allowing what may be a much larger DNA-protein complex to be visible at the top of the gel. Other YY1 EMSA gels did contain some suggestion of this higher band, but were not run far enough along the gel for it to become properly visible. Therefore, this behaviour was not taken as unique to freeze thawed YY1.

Like the freeze-thawed YY1, it was also necessary to ensure that TCP-YY1 was active and did not significantly change the DNA binding behaviour of the protein compared to the WT. Both protein constructs bound DNA at a concentration of 0.5 μM YY1, which was at a 1:1 YY1:DNA ratio.

5.3.1.2. YY1 DNA BINDING AFFINITY

While an EMSA can be used to observe different complexes and to confirm activity, it is a very low resolution technique. In order to determine a dissociation constant for the interaction, a more sensitive technique, where possible, is preferable. DNA binding assays with YY1 and the AAV P5 DNA sequence were done using fluorescence anisotropy to determine the binding affinity.

YY1 was first titrated into ROX-labelled DNA. As the YY1 concentration increased, so did the anisotropy value (***Figure 23***). Because the anisotropy value is dependent on the molecule's – or complex's – size, this increase showed that the overall size of the DNA increased, which indicates that YY1 was binding to it. However, this increased consistently until YY1 reached 20 μM. Unfortunately, it was not possible to increase this concentration further because YY1 was not stable at the high stock concentrations that would have been required to do so. In most cases, the fact that the binding curve does not yet reach saturation at the concentrations used,

would indicate a low DNA binding affinity. However, this contradicts the apparent binding affinity seen in the YY1-DNA EMSAs, which usually exhibit a much higher K_D than seen in anisotropy. Given that oligomeric complexes are observed in the EMSAs, which do seem to increase with increasing concentrations of YY1, it is possible that the increase in anisotropy is due larger YY1-DNA complexes forming, as opposed to a low DNA binding affinity.

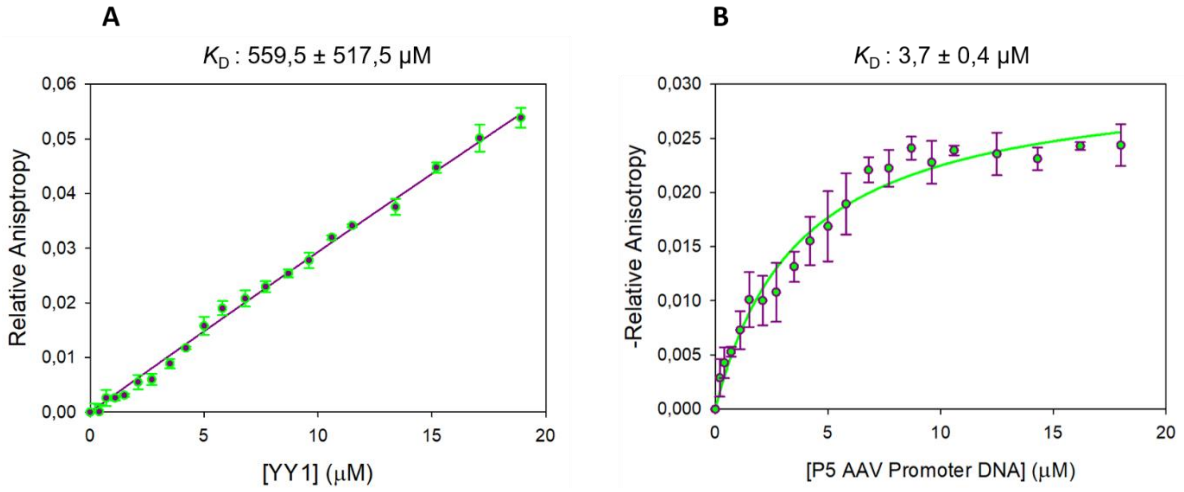


Figure 23: YY1 binding to DNA as seen with fluorescence anisotropy.
A) Shows YY1 titrated into ROX-labelled AAV P5 DNA. **B)** Shows DNA titrated into ATTO 550 labelled YY1. Negative relative anisotropy is shown on the Y-axis, meaning that the YY1 complex size decreases with the addition of DNA. Graphs were made using SigmaPlot 12.0.

To test this, AAV P5 promoter DNA was titrated into ATTO 550-labelled YY1, to prevent concentration-dependant oligomers from forming. This resulted in a decrease in anisotropy, which nevertheless formed an isotherm (***Figure 23***). When fitted to a ligand binding curve, the K_D was calculated to be 3.7 ± 0.4 µM. Although a decrease in anisotropy indicates a decrease in complex size, which is counter-intuitive to binding, the shape of the curve does suggest a binding event. The decreasing complex size observed as DNA binds could be a result of DNA disrupting the large oligomeric complexes YY1 forms, resulting in a smaller overall complex size.

5.3.2. FHD-DNA BINDING

Like with YY1, the activity and the specificity of the FOXP2 FHD was investigated using EMSAs. FOXP2 FHD bound to the Nelson sequence as expected (***Figure 24***), indicating that it was likely folded correctly and, therefore, active. Likewise, the FOXP2 FHD also bound to the CNTNAP2 sequence of DNA, to which it should also be specific. This indicates that this may be a site in the CNTNAP2 promoter to which the FOXP2 FHD binds. When together with the AAV P5 promoter sequence (specific to YY1), however, the FHD did not display any indication of interaction (***Figure 24***). Because of this specificity, it was possible to do further experiments to observe the interaction between YY1 and FOXP2 FHD in the presence of AAV P5 DNA (see sections 5.7 and 5.8).

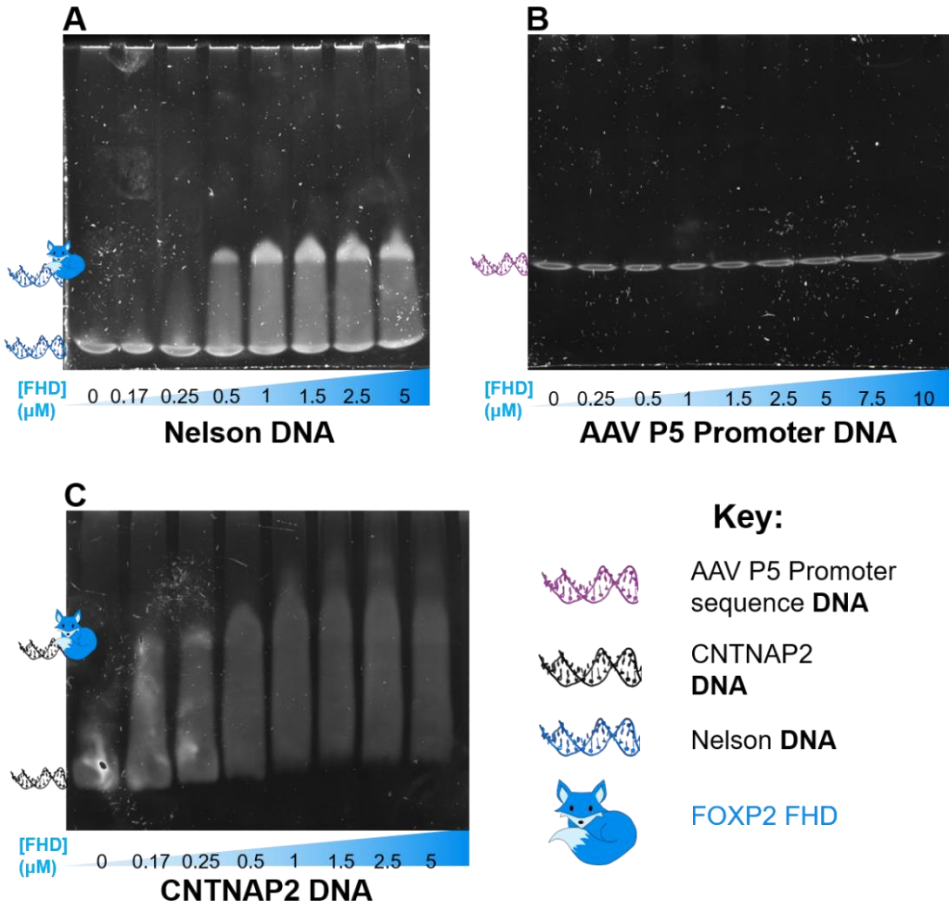


Figure 24: DNA binding assays of FHD to various DNA sequences.
A) Shows Nelson DNA being bound to the FHD. **B)** Shows AAV P5 promoter DNA in the presence of increasing FHD, where a complex is not formed. **C)** Is CNTNAP 2 DNA with FOXP2 FHD, where increasing amounts of FHD do result in DNA binding. Images created using PowerPoint (2013) and InkScape 12.0.

5.4. STRUCTURAL CHARACTERISATION

5.4.1. SECONDARY STRUCTURE

5.4.1.1. YY1 SECONDARY STRUCTURE

The secondary structure plays a large role in protein flexibility and overall protein folding, which in turn can help to elucidate how a protein functions. Some work has been done on the secondary structure of YY1, including the full length protein, the N-terminal in isolation, and the C-terminal region (Górecki *et al.*, 2015; Nieborak and Górecki, 2016; Chen *et al.*, 2019). The secondary structure of YY1 was investigated using far UV CD, in part to compare with the present literature. Additionally, it was done under different conditions which had not previously been investigated, so as to observe if these conditions induced a significant change in secondary structure.

The CD spectrum of YY1 shows a very slight trough at about 225 nm, but has its strongest signal at about 205 nm (***Figure 25***). Below this, the noise was generally increased to a point at which the data could not be confidently used. In part, this noise may have been due to some buffer interference. However, MilliQ only had slightly less noise, indicating this was a limitation of the available cuvettes and instrumentation. The spectra seem to suggest a strong presence of random coil (Kelly *et al.*, 2005). Unfortunately, random coils exhibit the lowest point of their trough below 200 nm, at which point in these spectra, the amount of noise is too high for use. However, other YY1 far UV CD have produced a similar shape with the same troughs, and calculated large amounts of random coil, followed by beta sheet, turn and then helical content (Górecki *et al.*, 2015; Nieborak and Górecki, 2016). No difference was seen between WT and TCP-tagged YY1 (***Figure 25***).

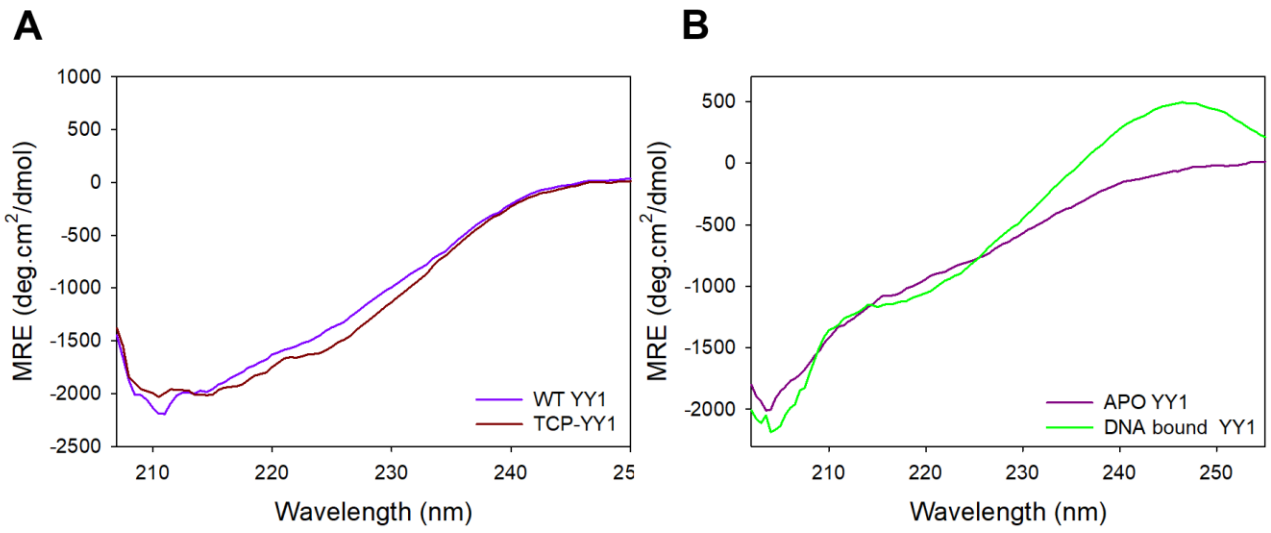


Figure 25: CD spectra of YY1 in various conditions.
A) Comparison of CD spectra of YY1 and its TCP construct, showing overall similarity in secondary structure. Both constructs contained the same secondary structural content. **B)** Comparison of the CD spectra of apo and DNA-bound YY1, which do exhibit a change in structure. All spectra are only shown where the HT voltage was below 600, so as to remove data reflecting high levels of noise. Overall YY1 shows a far UV CD spectrum which is characteristic of large amounts of random coil. Figures were plotted using SigmaPlot 12.0.

The addition of DNA to YY1 did result in a change in the shape of the far UV CD spectrum of YY1. The most obvious change is that at higher wavelengths of about 240-250 nm, the protein's MRE is greatly increased, in the positive range. This might suggest the formation of extended 3_1 or poly (Pro) II helix structures (Kelly *et al.*, 2005), which characteristically have high MRE at longer wavelengths compared to where other structures form peaks. Additionally, the DNA-bound YY1 spectrum shows a distinct trough at 220 nm compared to the apo form, which may suggest an increase in alpha helical content (Kelly *et al.*, 2005).

5.4.1.3. FHD SECONDARY STRUCTURE

The FHD CD spectrum was measured to ensure that it was correctly folded (***Figure 26***). This spectrum displayed a shape characteristic of the FHD, where the troughs at about 222 nm and 208 nm indicated a relatively large amount of alpha helix (Kelly *et al.*, 2005). This is consistent

with previous CD spectra of the FHD (Blane and Fanucchi, 2015), and, therefore, indicates correct secondary structural composition. As such, the FOXP2 FHD was likely folded correctly.

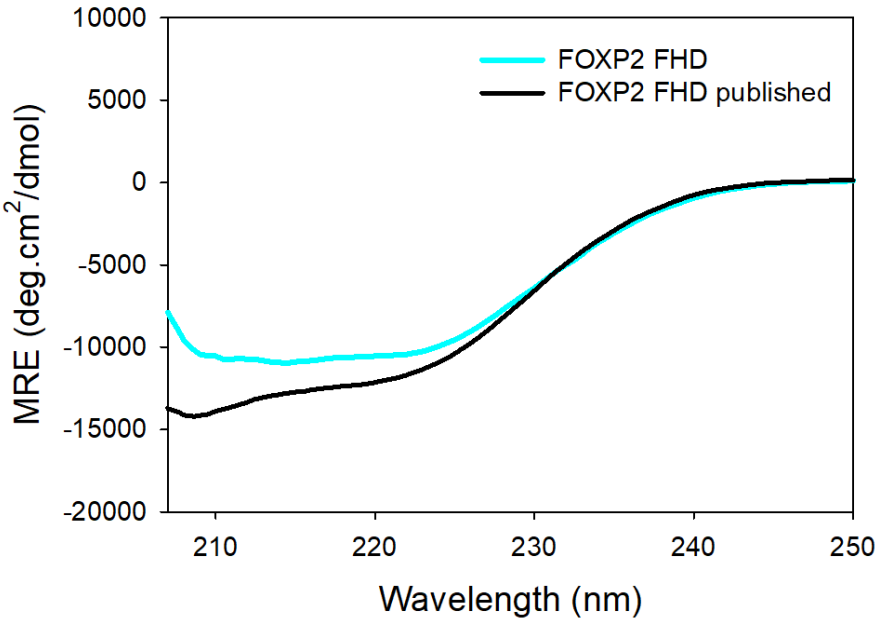


Figure 26: Far UV CD Spectrum of FOXP2 FHD.
FOXP2 FHD was in a buffer of 10 mM HEPES and 59 mM $NaSO_4$, pH 7.5 (blue). The troughs at 208 and 222 nm are indicative of a predominantly alpha helical structure (Kelly and Price, 2000), which matches with the crystal structure of the protein (Stroud *et al.*, 2006). This matches with the troughs seen in the previously published data (black) (Blane *et al.*, 2018), the data for which was kindly provided by Dr Blane. Some differences in the intensity of the troughs was observed between studies. The spectrum is shown from the point at which no noise was seen in both the published data and in this study. Graph was plotted using SigmaPlot 12.0.

5.4.2. YY1 TERTIARY STRUCTURE

Any protein containing tryptophan or tyrosine residues is capable of fluorescence when excited at 280 or 295 nm (Lakowicz, 2006). When tryptophan and tyrosine are excited at these wavelengths, a Stokes shift occurs, which is the phenomenon where the energy of emission is lower than that of absorption. The loss of energy seen in tryptophan fluorescence is sensitive to the local environment (Lakowicz, 2006). Therefore, the emission wavelength of tryptophan varies according to its exposure to both the solvent and surrounding residues which might interact with it. As a result, intrinsic fluorescence of a protein can indirectly provide

information on the areas surrounding the tryptophan residues through the emission wavelength and its changes under different conditions.

Intrinsic fluorescence was measured for YY1 in order to ensure that it was correctly folded after purification, as well as to characterise it under different conditions. To further determine how well folded the protein was, it was necessary to compare its native spectrum with its denatured spectrum (***Figure 27***). Therefore, the protein was denatured using 6 M guanidinium chloride and the spectra were compared. The fluorescence spectrum of the denatured YY1 exhibited a red shift along with increased intensity in both the 280 and 295 nm spectra. The higher intensity increase in the 280 nm spectra was expected, because tyrosine .often fluoresces at high intensities under denaturing conditions, when more exposed to the solution (Lakowicz, 2006). The red shift indicates that the tryptophan residues, of which there are two, experienced a change in environment, and were also likely more exposed to the solution. The shift is not very large, which is also expected, because YY1 is not a highly structured protein (Górecki *et al.*, 2015).

Tryptophan fluorescence of TCP YY1 was compared with that of the wild type to ensure that the tag did not make an apparent difference to the tertiary structure. The spectra were not significantly different, suggesting that the tag was an innocuous addition to the protein sequence.

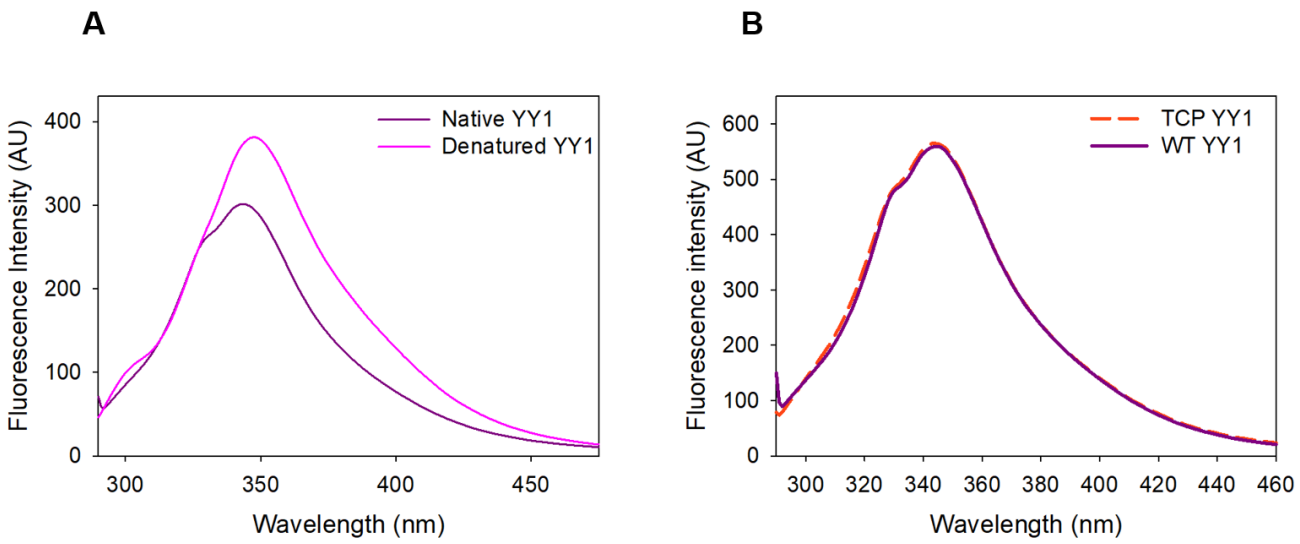


Figure 27: Fluorescence emission spectra of YY1 under different conditions to confirm tertiary structural integrity.
A) The fluorescence spectra (280 nm excitation) of 5 µM YY1 in its native form (purple) and denatured with 6 M guanidinium hydrochloride (pink). Upon denaturation, a red shift was observed, indicating structural change. **B)** Fluorescence spectra (280 nm excitation) of 10 µM YY1 (purple) and 10 µM TCP-YY1 (red). The addition of the TCP tag did not alter the tertiary structure of YY1. All graphs were plotted using SigmaPlot 12.0.

DNA binding can causes structural changes in a protein. For example, the level of structure in a zinc finger motif can increase upon binding (Houbaviy and Burley, 2001; Górecki *et al.*, 2015). YY1 has only two tryptophan residues, both of which occur in the REPO motif upstream of the DNA binding domain. Therefore, the emission spectrum after excitation at 295 nm gives a relatively localised indication of any changes that might occur upon DNA binding (***Figure 28***). When bound to either the AAV P5 or CNTNAP2 DNA sequences, the fluorescence emission intensity is quenched. This is typical of DNA binding (or at least DNA being in the proximity of the fluorophores), and does seem to suggest that the DNA has some level of proximity to the tryptophan residues. No apparent wavelength shift is visible, however, suggesting that large structural changes which result in the tryptophan residues being in a different local environment upon DNA binding are unlikely. Interestingly, the YY1 specific AAV P5 sequence quenches the tryptophan fluorescence less than the sequence it does not specifically bind to, the CNTNAP2 sequence. This may reflect some differences between how the protein makes non-specific contacts with DNA compared to how it binds with sequences it has higher affinity and specificity to.

The difference in quenching is not apparent when exited at 280 nm when both tryptophan and tyrosine residues are excited (***Figure 28***). The tyrosine residues in YY1 occur throughout the protein sequence, and several of them are present within the DNA binding domain. The 280 nm spectra, therefore, indicate that the global tertiary structure is less dependent on the specific DNA sequence bound.

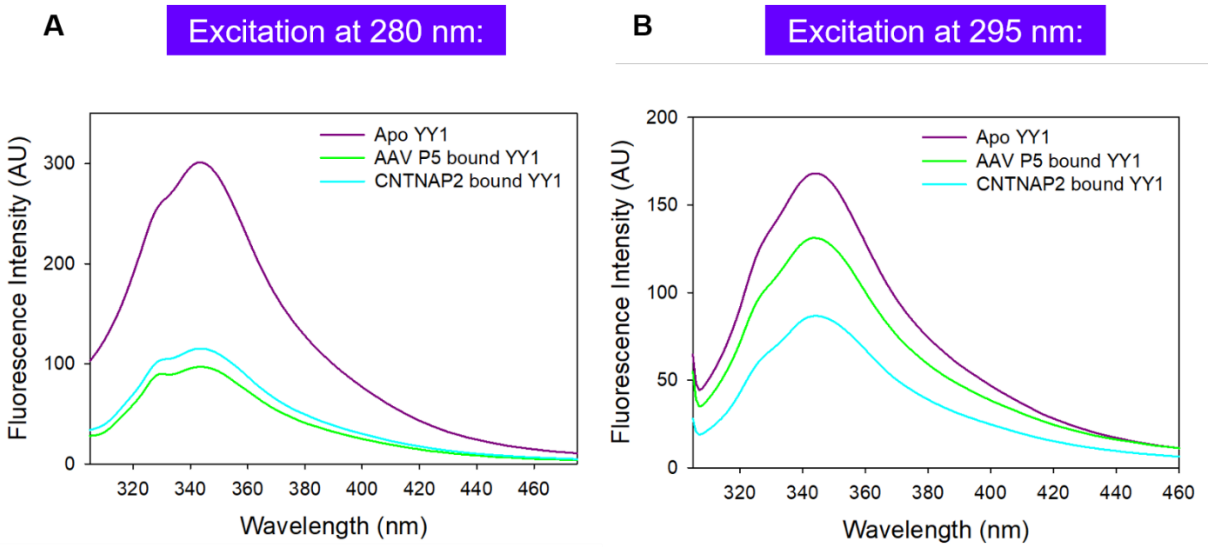


Figure 28: Fluorescence emission spectra of YY1 in the presence and absence of DNA. Emission spectra were recorded with an excitation wavelength of both 280 (**A**) and 295 (**B**) nm in order to obtain both local (295 nm) and more global (280 nm) structural information. A protein concentration of 5 µM was used in addition to 10 µM of the relevant DNA. Both DNA sequences quenched YY1 fluorescence. CNTNAP2 DNA, however, exhibited greater quenching than the AAV P5 promoter DNA sequence for 295 nm. Graphs were plotted using SigmaPlot 12.0.

5.4.3. FHD TERTIARY STRUCTURE

Like with YY1, the structural integrity of the FOXP2 FHD was confirmed using fluorescence spectroscopy. Because the FHD spectrum has been studied in detail before (Blane and Fanucchi, 2015), it was not necessary to do this in various conditions. Rather, the native FHD was compared to previous spectra in the literature. They were found to match, indicating that the protein was likely folded in the same way as previous studies and, therefore, native (***Figure 29***) (Blane and Fanucchi, 2015).

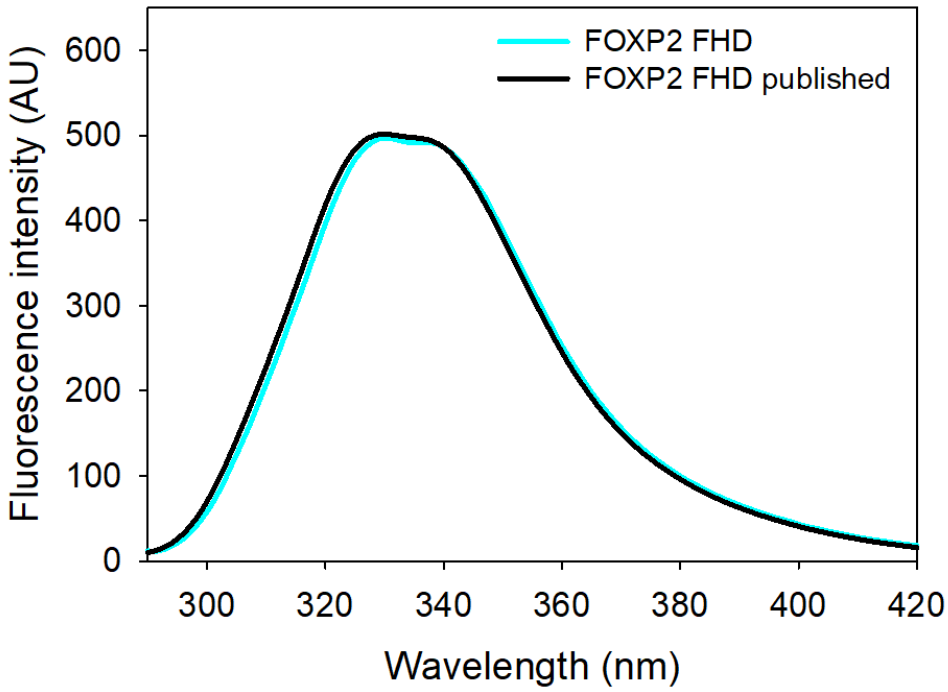


Figure 29: Comparison of the current and previously published FOXP2 FHD fluorescence emission spectra when excited at 280 nm.
The fluorescence spectrum of FOXP2 FHD was compared with the previously published data corresponding to excitation at 280 nm (Blane *et al.*, 2018).FOXP2 contains the expected tertiary structure. The data for the published spectrum was kindly provided by the Dr Blane. Both spectra indicate the same tertiary structure. This Graph was plotted using SigmaPlot 12.0.

5.5. QUATERNARY STRUCTURAL STUDIES

The quaternary structure of a protein is often important for its function. Some proteins are only functional in specific quaternary structures, while others may alter their function as their quaternary structure changes (Weintraub *et al.* 2017). Oligomeric interfaces are frequently hydrophobic, but can also be based more on hydrophilic contacts, while some can involve disulfide bridges (Uversky and Dunker 2010). In the cell, YY1 is associated with multimeric interactions which are important for its role in DNA structural organisation and in the formation of proteinaceous membrane-less organelles (Weintraub *et al.*, 2017). *In vitro*, YY1 has been observed to form a range of oligomers in some studies but not in others (Golebiowski *et al.*, 2011; López-Perrote *et al.*, 2014; Wai *et al.*, 2016). These variable results combined with the physiological relevance of YY1 oligomerisation make this an important aspect that needs

further clarity. This is especially important considering the potential role of oligomerisation in DNA binding, for which very little information is available. The quaternary structure of YY1 was probed using both size exclusion chromatography and fluorescence anisotropy under different conditions so as to attempt to determine if, when and how YY1 forms oligomers.

Size exclusion chromatography can be used to separate different sized multimers from each other in native conditions. The resulting elution profile will show earlier peaks that are representative of larger species and later peaks that represent smaller species (with size, in this case, being reflected by hydrodynamic volume) (Gámez-Valero *et al.* 2016). The SE-FPLC column used in this study had an exclusion range of 10-600 kDa, while the SE-HPLC column had a range of 6-700 kDa. In this study, FPLC and HPLC SEC both demonstrated that YY1 elutes in a large peak in the void volume, followed by one or more smaller peaks that were able to be resolved by the column (***Figure 29*** and ***Figure 30***). These smaller peaks were less visible when performing size exclusion using FPLC compared to HPLC, which has higher sensitivity. FPLC, however, did not require any adjustment of pH and could be done at the higher YY1 concentrations used during purification, making both techniques valuable for comparison of behaviour under the different conditions they allowed.

SE-FPLC was conducted at YY1 concentrations of both 50 and 25 μM in the presence and absence of a reducing agent in order to confirm that the cysteine residues in the zinc finger were not forming disulphide bridges instead of interacting with metal ions to form the zinc fingers. (***Figure 29***). At both concentrations, the elution profiles are the same whether under reducing or non-reducing conditions. Therefore, YY1 does not seem to be forming disulfide bonded oligomers. While the peak in the void volume is very visible, the peak eluted later after around 65 ml is more difficult to see, especially at the higher concentration. The YY1 multimers in the void volume peak may consist of a variety of sizes. Unfortunately, this could not be assessed as columns with a higher separation range were unavailable. The different peaks could not be used separately in downstream experiments, because when the isolated peaks were size excluded again, they eluted with the original profile containing multiple peaks, indicating an equilibrium. Because of the lack of resolution within the void volume, fluorescence anisotropy was used to observe global size changes of the protein upon concentration increases, which is discussed later in this section.

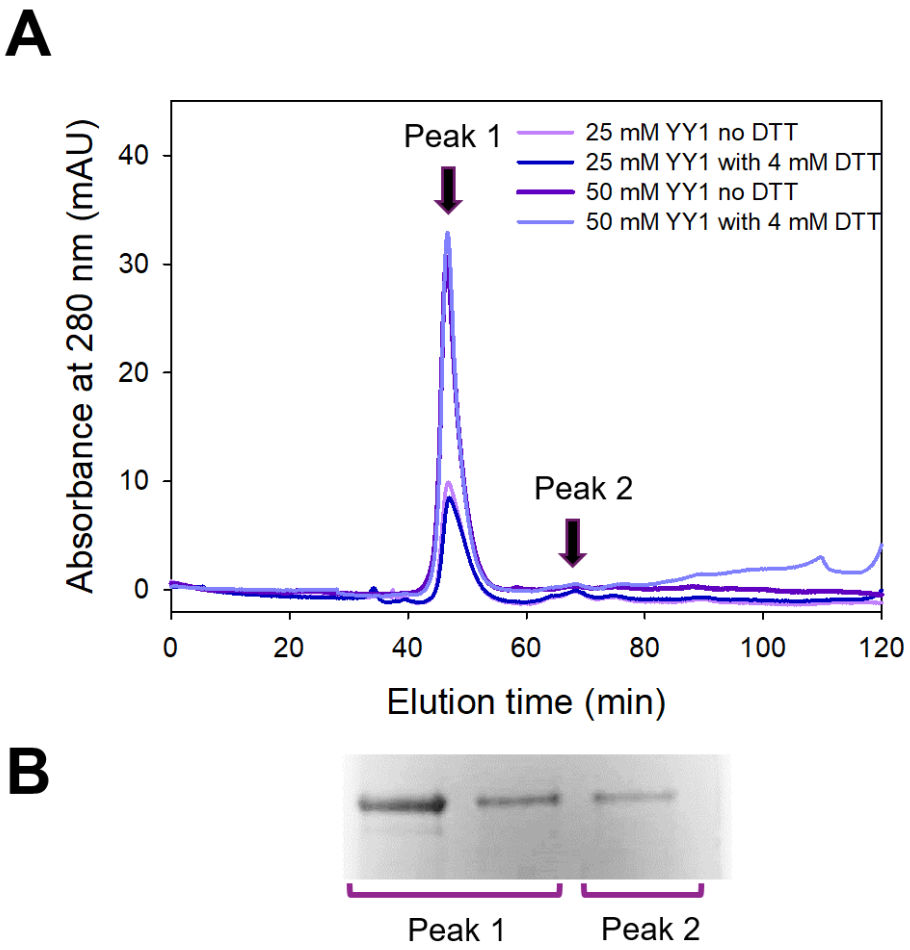


Figure 30: SE-FPLC elution profile of YY1 at 25 and 50 µM, in reducing and non-reducing conditions.
A) Two peaks are detected at about 45 and 68 ml. The former peak is within the void volume, which is at 600 kDa. As such, YY1 could exist in one or many different apparent molecular weights above this. The latter peak elutes at a volume corresponding to 125 kDa relevant to the standards (HMW calibration kit, Cytiva, USA). These sizes are unlikely to be accurate as YY1 is not globular. The protein showed relatively little difference in peaks at these concentrations, and no difference under reducing and non-reducing conditions. Arrows indicate YY1 elution. **B**) SDS PAGE gel corresponding to peaks one and two in SE-FPLC of YY1. The gel indicates that YY1 is eluted in both peaks. Graph was plotted using SigmaPlot 12.0.

SE-HPLC was performed on lower concentrations of YY1 than were used in SE-FPLC, ranging from 2.25 µM to 20 µM (***Figure 30***). These concentrations were too low to use successfully with SE-FPLC but it was necessary to determine the oligomeric state of YY1 at the lower concentrations so as to get a better understanding of how YY1 oligomerisation relates to its function. Separation of YY1 using SE-HPLC resulted in a similar profile to the SE-FPLC

results with a large peak in the void volume followed by several smaller peaks. As with the SE-FPLC profile, this large peak may be a heterologous mixture of YY1 multimers. Because HPLC is a higher resolution technique than FPLC, it was possible to distinguish between the smaller peaks more easily. In a buffer containing 300 mM NaCl, several oligomeric species were observed. The species which eluted after the void volume did not elute in highly resolved peaks. Nevertheless, they did show consistency between concentrations, where the highest protein absorbance was observed at peaks 2 and 3 in ***Figure 30***, which elute at volumes corresponding to about 150 and 44 kDa respectively when compared to the globular standards. The predicted molecular weight of YY1 is about 45 kDa, suggesting that peak 3 may be monomeric. There was some indication of other YY1 complexes in between these main peaks, which is most visible in the 10 μM sample of YY1. The apparent size of peaks 2 and 3, as well as the void volume peak were also seen by Lopez-Perrote (2014), although in different ratios of intensity, which may be due to different conditions.

The HPLC elution profile for YY1 at 500 mM NaCl contained a few differences compared to the lower ionic strength buffer (***Figure 30***). Here, the void volume peak was higher, while the smaller species were less apparent. Therefore, ionic strength does seem to influence the oligomeric state of YY1. The increase in the void volume peak and decrease in the intensity of following peaks resulting from increased ionic strength, suggests that YY1 likely favours oligomerisation with more salt present.

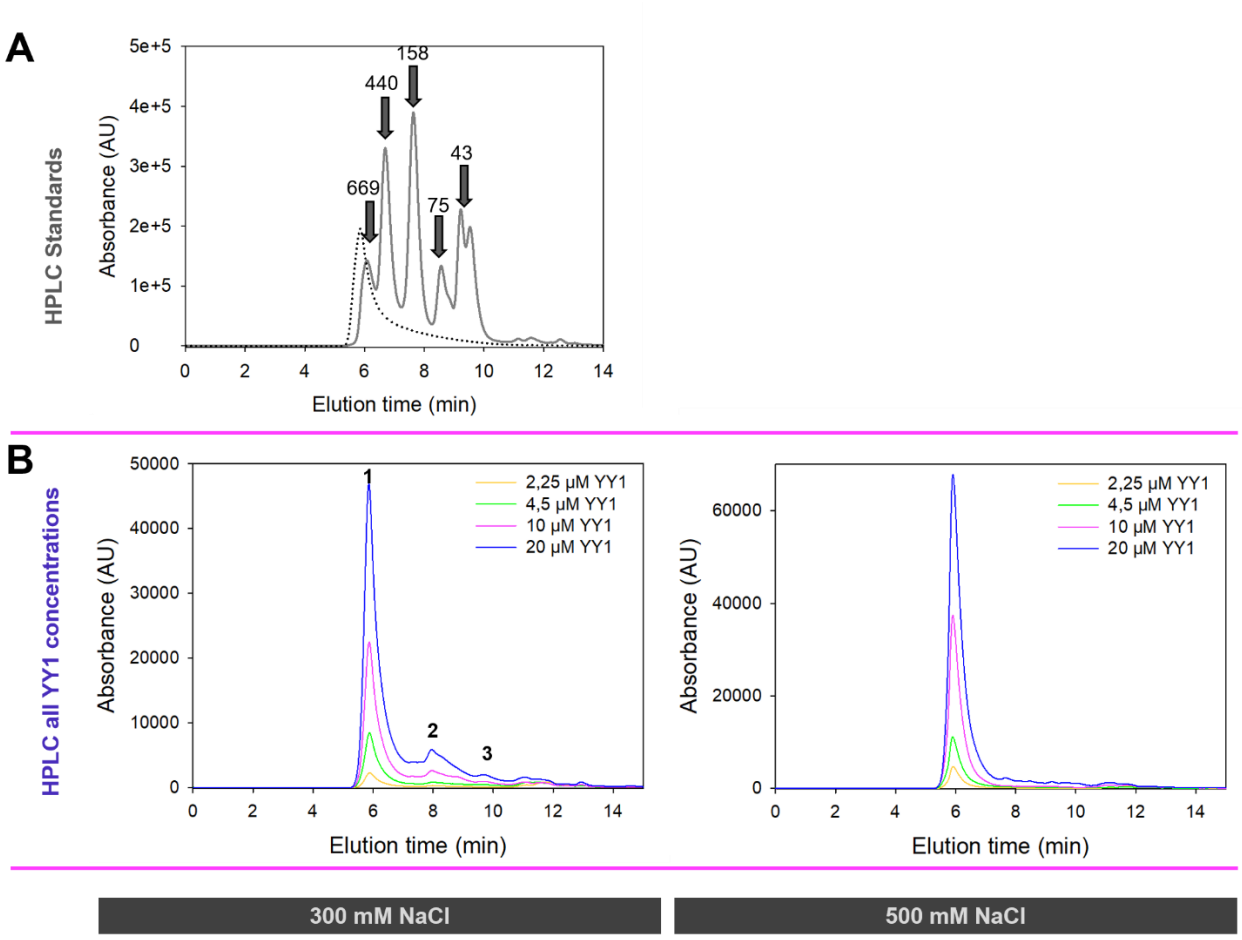


Figure 31: SE-HPLC profiles for YY1 under high (500mM) and low (300 mM) salt conditions.
A) The standard curve profile with the sizes of the peaks (HMW calibration kit, Cytiva, USA). The dotted line represents blue dextran, which elutes in the void volume. **B)** The elution profile of increasing concentrations of YY1 at 300 mM NaCl (left) and 500 mM NaCl (right). Higher salt concentrations increase the formation of large multimers. Several peaks across different sizes are apparent in the samples containing lower salt concentrations, indicating that YY1 can exist in a mixture of oligomeric forms. The resolution of these peaks varies among the different YY1 concentrations, which may indicate potential changes in the exact species involved across different YY1 concentrations. Separate peaks could not be shown in an SDS PAGE gel due to the low concentrations of the smaller peaks and the difficulty in distinguishing fractions corresponding to each peak due to distance from the flow cell and collection tube. Figures were plotted in SigmaPlot 12.0.

The peak in the void volumes were not necessarily homologous – it may have been an amalgamation of different sizes of oligomers above the exclusion limit (600 and 700 kDa for Se-FPLC and SE-HPLC, respectively). It is not possible to make conclusions on the size of any of the species in the different peaks despite the use of standards. This is because these exclusion

ranges and standards only apply to globular proteins. YY1 is an intrinsically disordered protein (Górecki *et al.* 2015). This makes it impossible to accurately determine its size based on SEC, as its hydrodynamic volume is expected to differ from that of a highly structured globular protein – intrinsically disordered proteins can range anywhere from 1.2 to 12 times the hydrodynamic volume of a globular protein the same chain size (Uversky and Dunker 2010).

The size exclusion profile of YY1 confirms that it is capable of forming a variety of oligomers. Because adequate separation of the very large species was not possible due to the limitations of the column, fluorescence anisotropy was employed in order to gain a more global understanding of how factors such as concentration may influence its oligomer formation. Increasing concentrations of unlabelled YY1 were titrated into NTA ATTO 550-labelled YY1 (***Figure 31***). An overall increase in anisotropy would suggest that YY1 complex size is concentration-dependent. This study was done at three salt concentrations: 100, 300 and 500 mM NaCl.

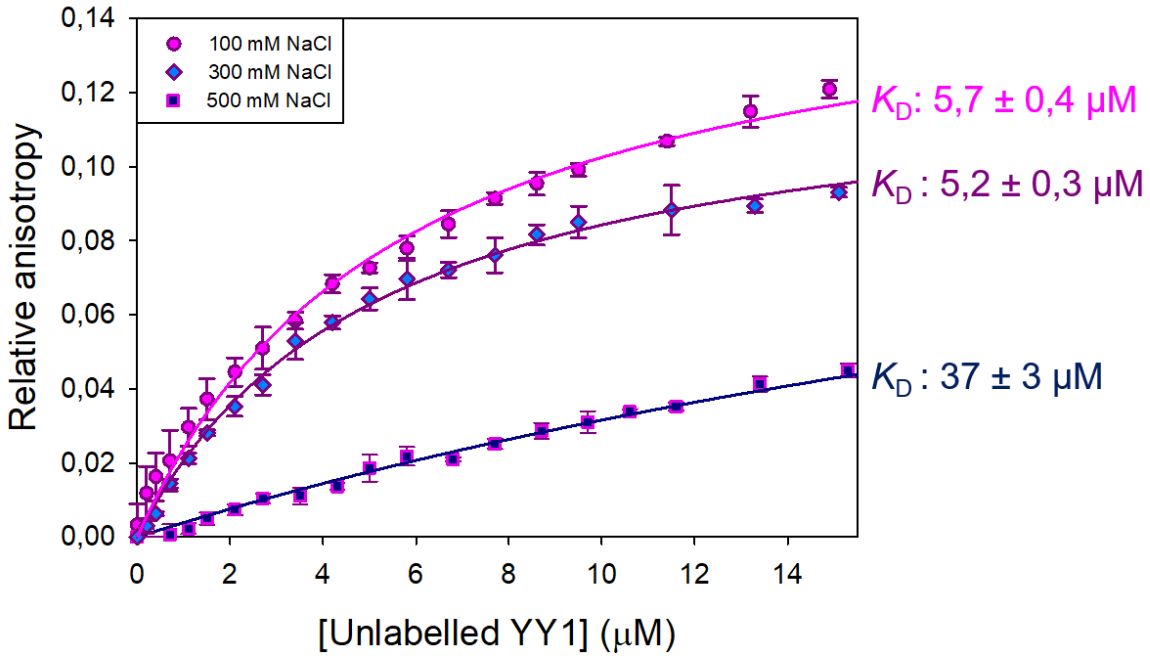


Figure 32: Anisotropy binding curves showing YY1-YY1 interaction at different salt concentrations.
Unlabelled YY1 was titrated into 0.2 µM NTA ATTO 550-labelled YY1. The light pink circles show YY1-YY1 binding in a buffer containing 100 mM, while the diamond and squares represent salt concentrations of 300 and 500 mM, respectively. On the right are the predicted apparent K_D values for each curve. The curve was made and fitted (ligand binding single site model) using SigmaPlot 12.0.

An increase in the overall complex size was observed for all ionic strengths when the YY1 concentration was increased. The total change in size, indicated by the relative anisotropy increase, was similar for YY1 in both 100 and 300 mM NaCl. The apparent K_Ds for both of these ionic strengths were at 5.7 ± 0.4 (100 mM NaCl) and 5.2 ± 0.3 μM (300 mM NaCl) (***Figure 31***). These K_D values are difficult to interpret as measures of affinity for three reasons. Firstly, the actual size and oligomeric state of the protein are unknown. Secondly, it is likely that the oligomeric state differs across different concentrations. Lastly, the composition of oligomeric states may differ across these salt concentrations, as was seen in the SE-HPLC results when using 300 and 500 mM NaCl. However, the K_D values do provide insight into the concentrations at which the oligomeric profile of YY1 shifts. The average complex size of the protein stays more consistent at concentrations higher than twice the K_D, where the asymptote of the curve begins. For YY1 at 300 and 100 mM NaCl, the asymptote begins at about 10 or 11 μM, whereas at 500 mM NaCl no clear asymptote is apparent. This means that at lower concentrations of NaCl, shifts in YY1 concentration have a greater effect on the oligomeric profile of YY1.

In contrast to the 100 and 300 mM NaCl YY1 anisotropy profiles, the 500 mM NaCl curve indicated a smaller net change in complex size (***Figure 31***). Additionally, the K_D for the 500 mM NaCl sample was 37 ± 3 μM, which is nearly an order of magnitude higher than that of the lower ionic strengths. In conjunction with the YY1 SEC profile, the higher K_D for the 500 mM NaCl curve suggests a continued steady increase in average complex size as the salt concentration increases further, as the oligomeric profiles between salt concentrations are different. Therefore, the K_D value cannot be used to compare affinity, as different species are involved across different ionic strengths. This confirms the observation in SE-HPLC that ionic strength does affect the oligomeric profile of YY1.

Together, the anisotropy and HPLC results indicate that YY1 forms concentration dependant oligomers influenced by ionic strength when in the μM range. The protein forms a range of different oligomeric species, which were observed to change with changing ionic strength in the SE-HPLC elution profile.

5.6. YY1 STRUCTURAL DYNAMICS

The structural dynamics of a protein are an intrinsic part of their function, and can be used as an indicator of how they might bind to ligands or even be folded. HDXMS is a technique which probes the dynamics of a protein on the basis of how exposed the backbone is to the solution due to its flexibility (Konermann *et al.* 2011). Little is known about the structure of YY1, especially outside of the DNA binding domain. The structural dynamics of the full length protein were, therefore, probed with and without DNA to obtain a clearer picture of how the protein behaves in solution, and how this behaviour is influenced by DNA binding.

After deuterium uptake was quenched at different time points, the protein was digested into peptides, which were separated and analysed using mass spectrometry to determine the amount of deuterium uptake for each peptide at each time point. This resulted in the creation of deuterium uptake plots for each peptide, which, when overlapping peptides were subtracted from each other, created a heat map for both the apo form of YY1 and the DNA-bound form. The heat maps reflect the overall level of deuterium uptake across the sequence. Some sections of the heat map have higher degrees of resolution and reliability than others. The resolution and the reliability of the heat maps are dependent on peptide coverage, which can be seen by the yellow and green lines in ***Figures 32*** to ***36***. Some areas of the protein have better peptide coverage than others due to the protease's inability to cleave at certain amino acids, as well as due to the availability of the backbone. The more peptides used in the calculations, the more confidence one can have in the data. The higher the degree of sequence overlap between peptides within a region, the better the resolution was because this allowed for the deuterium uptake of smaller regions to be calculated. Regions where there was little to no overlap only indicated the deuterium uptake for that entire peptide stretch, as opposed to individual residues. In this way, the resolution of this method could vary.

YY1 only contains one structural domain, which is the four zinc finger motifs making up the DNA binding domain at the C-terminal end of the protein (Houbaviy *et al.* 1996). Upstream of the DNA binding domain, the protein is generally associated with intrinsic disorder (Houbaviy and Burley 2001; Górecki *et al.* 2015). However, the N-terminal region can be separated into distinct regions based on the patterns of its sequence: it has two acid regions, which contain a large amount of acidic amino acids; a polyhistidine region; a glycine-alanine and glycine-lysine rich region; and a spacer region containing two conserved sections of sequence (Gordon *et al.*

2006). As intrinsic disorder is often defined by certain characteristics of the sequence, including acidic regions and simple sequences with repeating amino acids (Uversky and Dunker 2010), these sections are each likely to have their own distinctive behaviour. Additionally, previous structural studies for the protein have been conducted by removal of these particular sections (Gordon *et al.* 2006). Therefore, each distinctive region was separated in this analysis for easier comparisons.

5.6.1. ACID RICH REGION 1

The first acid rich region of YY1 (amino acids 11-63) (***Figure 32***) has very similar peptide coverage in both the apo and DNA-bound samples. The peptide coverage starts at residue 17 (a tyrosine), and ends at residue 56 (glutamine). This area omits the histidine tag, as well as some of the N and C terminal amino acids in this region. Mostly, however, the area has good coverage. Like with the rest of the protein sequence, most of the residues were found to be 80-100% deuterated. This means that this region is very exposed to the solution, and likely consists of flexible random coils. Certain residues, or small stretches thereof, in the apo YY1, however, are much less deuterated – several of which are acidic residues (***Figure 32***). As such, the first acid rich region can be considered to be disordered with small pockets of potential structure.

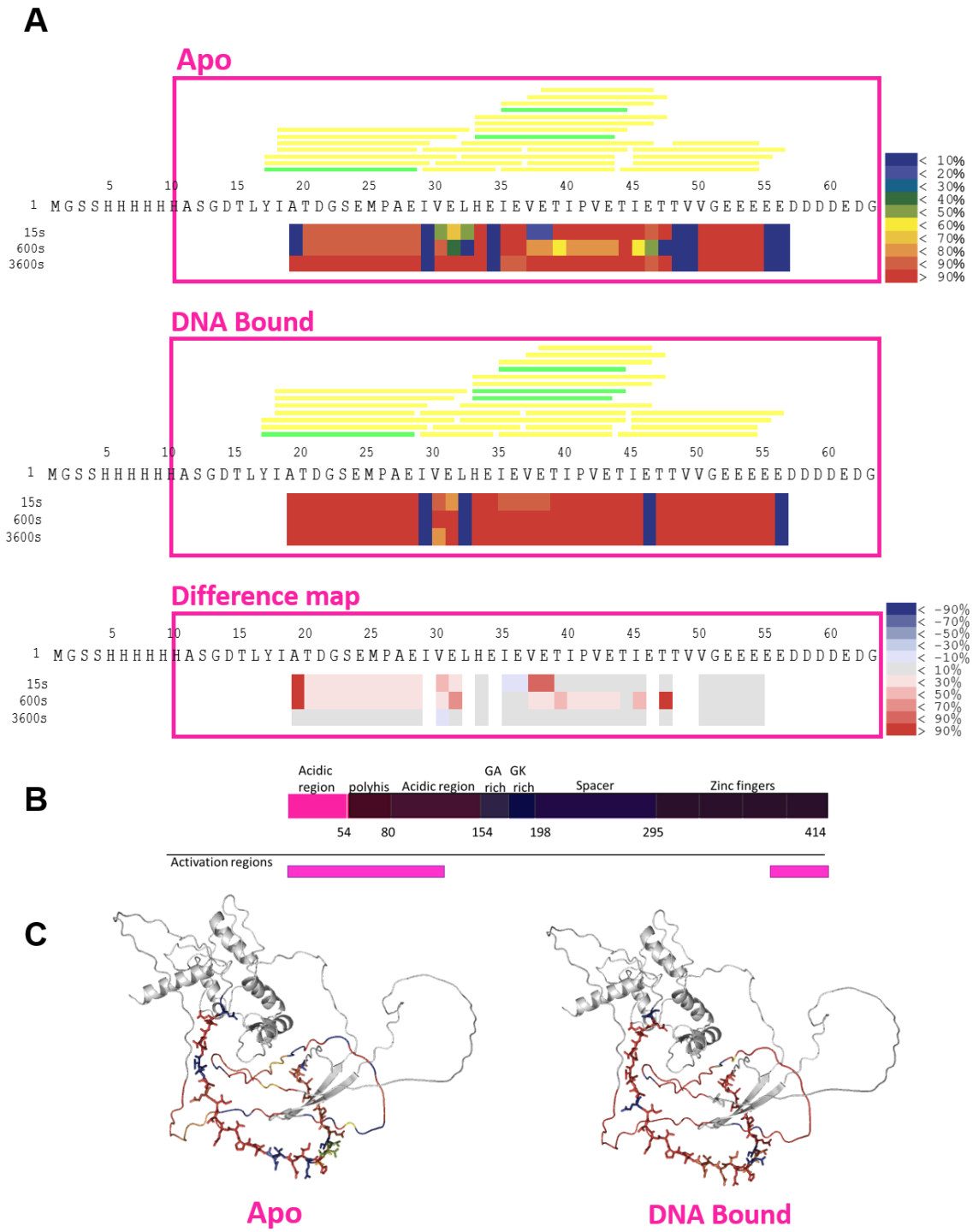


Figure 33: Heat map and difference map of the first acidic region of YY1 in its apo and DNA bound forms.

A) Apo YY1 is at the top while below it is the DNA-bound YY1. The % of deuteration of the backbone for each residue is represented by the colour scale, ranging from red (more than 90% deuteration) to blue (less than 10% deuteration). The less deuterated a residue's backbone, the more likely it is to be buried, whereas more deuterated residue backbones are more exposed to the solvent. The green and yellow bars above the heat map represent the high and medium confidence peptides used to create the heat map. The difference map shows the difference in deuterium uptake between the DNA bound and apo forms. The difference map is coloured on

a scale from red to blue, where red shows a greater uptake of deuterium and blue is a decrease in deuterium uptake upon DNA binding. The heat map shows that acid rich region 1 is predominantly disordered, with a few specific residues which are likely to form pockets of structure. **B)** The location of the first acidic region within the sequence and its associated function are shown below the difference map. **C)** At the bottom is the AlphaFold structure of YY1, with the acidic region coloured according to the apo and DNA-bound heat maps at 15 s. This predicted region is very low confidence on AlphaFold (Jumper *et al.*, 2021; Varadi *et al.*, 2022), which is common for disordered regions, and can only be interpreted as an indication of an ensemble of conformations, as opposed to a location of the backbone within 3D space (Ruff and Pappu, 2021). The propensity of the region to have an exposed backbone under each condition can be seen by the same colours as in the heat map. Heat maps were generated using HDX-aminer (Sierra Analytics, USA). Structures were rendered using PyMOL molecular graphics system version 1.8, Schrödinger, LLC. PDB code used was 2a07 (Alfieri *et al.* 2013).

The first acidic region of DNA-bound YY1 exhibited an overall increase in deuterium uptake compared to the apo form of the protein. This was reflected in the difference map (***Figure 32***), which indicates at least a 30 % increase in deuterium uptake across much of the region for at least one time point. In many cases, this shift results in fully deuterated residues. Furthermore, five residues experienced a 70-100% increase in deuterium uptake, which are extreme changes. The increase in deuterium uptake within the first acidic region indicates that the backbone becomes more exposed to the solution upon DNA binding. This exposure to the solution suggests a decrease in the amount or size of structured pockets of the protein, or may indicate a decrease in protein-protein interactions. Compared to the other protein regions, the first acidic region was one of the areas which showed the greatest amount of change upon DNA binding

5.6.2. THE POLYHISTIDINE REGION

The polyhistidine region of YY1 had no peptide coverage in either the apo or the DNA-bound samples. This is likely due to the large amount of histidine residues – pepsin, the enzyme used to cleave the protein, is inefficient at cleaving sequences at histidine residues (Hamuro *et al.* 2008). Therefore, no peptides were found for this region.

5.6.3. ACIDIC REGION 2

At the C-terminal end of the basic poly-histidine region is the second acid rich region (residues 93-163). Like the first acidic region, it has relatively good peptide coverage overall. Exceptions to this are residues 93-155 in the DNA-bound sample, where only three peptides were available. Both apo and DNA-bound YY1 also had poorer coverage from residues 155-160, where only two peptides were available (***Figure 33***). Those particular sections, therefore, need to be interpreted with caution.

Like with acidic region 1, acidic region 2 in the apo protein had more non-deuterated sections in comparison to much of the overall protein sequence. Despite this, the majority of the second acidic region was over 70% deuterated. Therefore, the second acidic region was still disordered overall, but had more patches within it that were inaccessible to the solution compared to other regions of the protein (***Figure 33***). This is especially true for residues 108-144. As with the first acidic region, several of the backbones of these residues are fully deuterated throughout each time point, or through most of the time points.

For both apo and DNA-bound forms of YY1 in the second acid rich region, there are several residues which have more deuterium uptake in earlier time points than in later time points. This implies that several sections within this region, as well as others within the protein, behave with distinctive dynamic trends. One conformation may last for relatively long times before changing to another, which may also last long enough to result in a difficult to see trend in deuterium uptake over time.

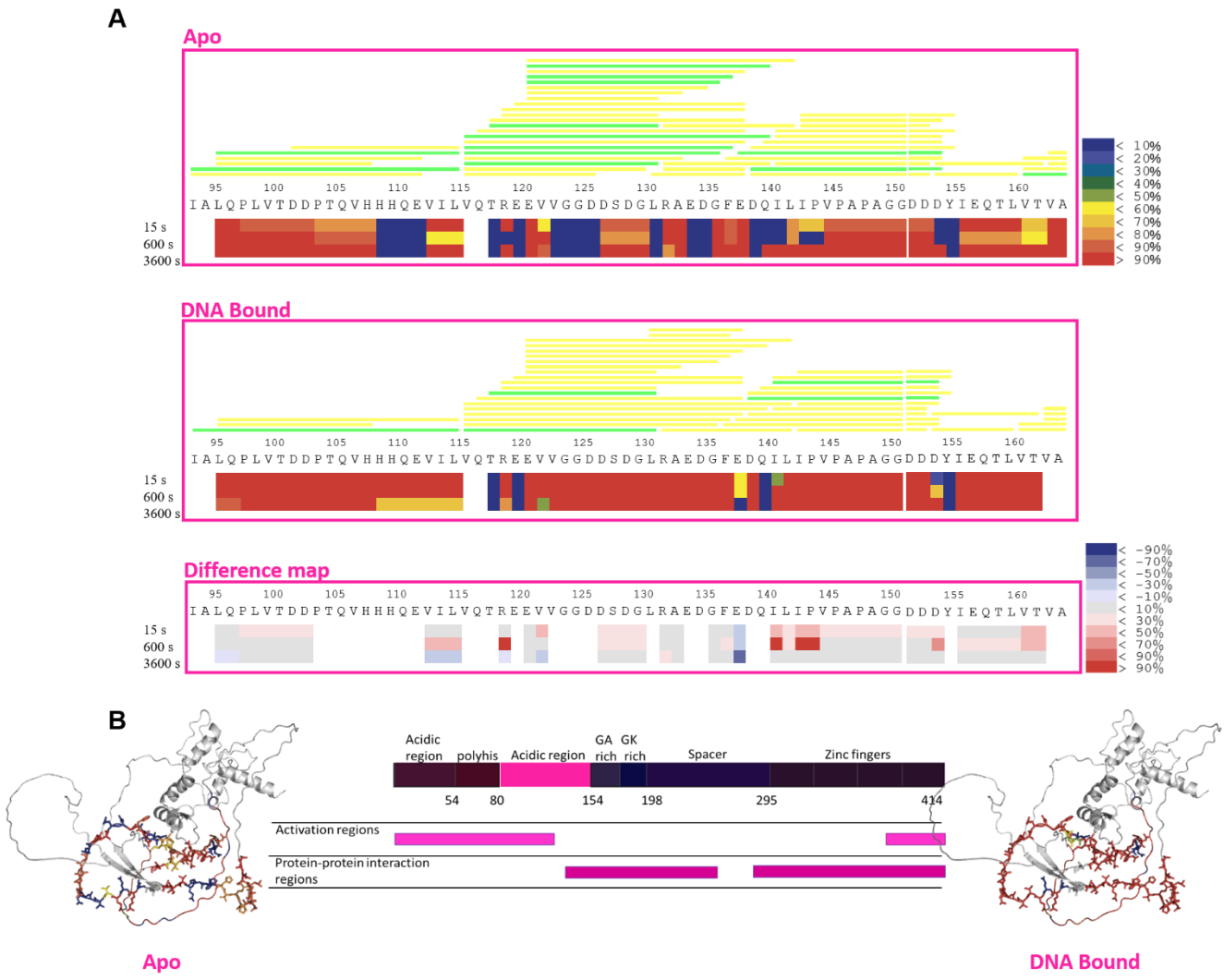


Figure 34: HDX heat maps and difference map of the second acidic region of YY1 when bound and unbound to DNA.

A) Apo YY1 is at the top while below it is the DNA-bound YY1. The % of deuteration of the backbone for each residue is represented by the colour scale, ranging from red (more than 90% deuteration) to blue (less than 10% deuteration). The less deuterated a residue's backbone, the more likely it is to be buried whereas more deuterated residue backbones are more exposed to the solvent. The green and yellow bars above the heat map represent the high and medium confidence peptides used to create the heat map. The difference map shows the difference in deuterium uptake between the DNA bound and apo forms. The difference map is coloured on a scale from red to blue, where red shows a greater uptake of deuterium and blue is a decrease in deuterium uptake upon DNA binding. This region of YY1 contains several pockets of structure. The DNA-bound form of the protein does appear to be more solvent exposed than the apo form. Heat and difference maps were generated by HDX-aminer. **B)** The location of the second acidic region within the sequence and its associated function is shown below the difference map. To the sides bottom are the AlphaFold structures of YY1, with the acidic region coloured according to the apo and DNA-bound heat maps at 15 s. This predicted region is very low confidence on AlphaFold (Jumper *et al.*, 2021; Varadi *et al.*, 2022), which is common for

disordered regions, and can only be interpreted as an indication of an ensemble of conformations, as opposed to a location of the backbone within 3D space (Ruff and Pappu, 2021). The propensity of the region to have an exposed backbone under each condition can be seen by the same colours as in the heat map. Structures were rendered using PyMOL molecular graphics system version 1.8, Schrödinger, LLC. PDB code used was 2a07 (Alfieri *et al.* 2013).

As was seen in the first acidic region, the second acidic region of YY1 experienced an increase in deuterium uptake upon DNA binding, indicating increased exposure to solution in the presence of DNA. The difference map shows that most residues which can be compared experience an increase in deuterium uptake upon DNA binding by at least 30 % for at least one time point. Several residues were also calculated to be areas with a higher degree of change, where the increase in deuterium uptake ranged from 50-90 % across at least two time points. This shows that overall, the second acid rich region in YY1 becomes more exposed to solution, and that several pockets of structure become lost upon DNA binding.

Although most residues in the second acidic region tended to become more flexible upon DNA binding, one residue in this region showed a decrease in deuterium uptake across all time points, which suggests that this residue (Glu137) may become involved in new backbone contacts or become sterically unavailable to the solution in a binding event. Two other residues show a decrease in deuterium uptake in one of the time points, but also an increase in alternative time points. This indicates that a change does occur upon DNA binding, but that there are probably several different conformations for these residues, the shorter-lived of which are probably more exposed to solution, while longer lived conformations may also have existed which were visible in the last time point. Alternatively, this change may have been a result of back exchange. Given that this region has a lot of peptide coverage, however, it is possible that this may be a real phenomenon.

Overall, the second acid rich region of YY1 was shown to contain several pockets of structure, as with the first acidic region. The DNA-bound form of the protein does appear to be more available to the solution overall in these regions, although there are a few residues which also show a decrease in availability, which may indicate either changes in structure, contacts between the protein, or both. Any changes are unlikely to result from actual contact with the DNA; not only is this region not associated with DNA binding, it is also an acidic region with more negatively charged residues than positive, and would, therefore, not be compatible with the phosphate backbone of nucleic acids.

5.6.4. GA AND GK RICH REGION.

The GA and GK rich regions (residues 163-206) are associated with transcriptional repression and also with protein-protein interactions. Besides the residues for which they are referred to, they are also rich in serine. The simplicity of this sequence coupled with the repetitive residues is a typical characteristic of sequences belonging to an intrinsically disordered region (Uversky and Dunker 2010). Both the apo and DNA-bound forms of this region have less peptide coverage in comparison to the acidic regions already discussed (***Figure 34***). There is also relatively little overlap between peptides in this region. Therefore, the heat map reflects the average deuterium uptake for most of the residues covered, as opposed to single residues or small clusters thereof.

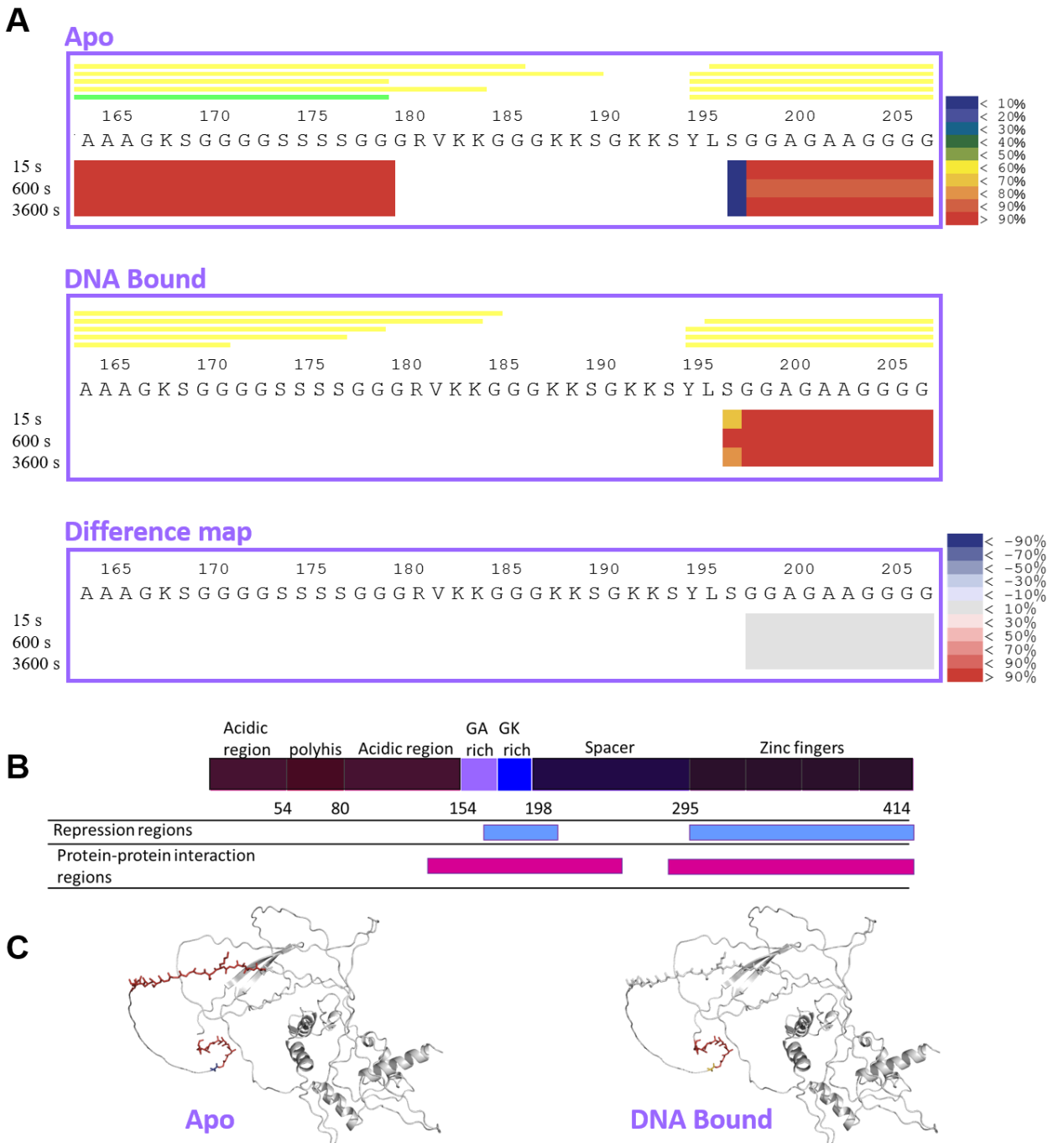


Figure 35: HDX heat and difference maps of GA and GK rich regions of YY1 in apo form and bound to DNA.

A) Top shows the heat map for apo YY1, middle the DNA-bound form and below is the difference map. The % of deuteration of the backbone for each residue is represented by the colour scale, ranging from red (more than 90% deuteration) to blue (less than 10% deuteration). The less deuterated a residue's backbone, the more likely it is to be buried whereas more deuterated residue backbones are more exposed to the solvent. The green and yellow bars above the heat map represent the high and medium confidence peptides used to create the heat map. The difference map shows the difference in deuterium uptake between the DNA bound and apo forms. The difference map is coloured on a scale from red to blue, where red shows a

greater uptake of deuterium and blue is a decrease in deuterium uptake upon DNA binding. For the regions with similar peptide coverage, the deuterium uptake was about the same. Heat and difference maps were generated by HDX-aminer (Sierra Analytics, USA). **B)** The location of the GA and GK rich regions within the sequence and its associated functions are shown below the difference map. **C)** At the bottom is the Alphafold structure of YY1, with the acidic region coloured according to the apo and DNA-bound heat maps at 15 s. This predicted region is very low confidence on alphafold (Jumper *et al.*, 2021; Varadi *et al.*, 2022), which is common for disordered regions, and can only be interpreted as an indication of an ensemble of conformations, as opposed to a location of the backbone within 3D space (Ruff and Pappu, 2021). This region is highly exposed to the solution, indicating few if any pockets of structure. Structures were rendered using PyMOL molecular graphics system version 1.8, Schrödinger, LLC. PDB code used was 2a07 (Alfieri *et al.* 2013).

Overall, this region appears to be very exposed to the solution, as most of it is more than 80% deuterated. Given that the sequence within this region can be associated with intrinsic disorder, this degree of flexibility makes sense. Residues 196 onwards appear to be mostly similar for the apo and DNA-bound protein. Upstream of this, there is unfortunately no usable coverage for the DNA-bound sample, although the apo section does contain nearly fully deuterated peptides from 164-178. Overall, the areas covered in the GAGK rich regions appear to be highly disordered.

5.6.5. SPACER REGION

Despite its dismissive name, the spacer region (residues 207-305) of YY1 seems to be one of the most integral to its function. Besides the DNA binding domain, the only other conserved regions of the protein are located within the spacer region. It has been associated with transcriptional repression, and also with protein-protein interactions (Gordon *et al.* 2006; Alfieri *et al.* 2013). Very little is known about this region, as no homologous structures have been solved. One structure does exist, but only as a thirty residue stretch of the first conserved sequence when bound to another protein. This section of protein was observed to form a beta turn and termed the recruitment of ploycomb (REPO) domain by the authors (Alfieri *et al.* 2013), although, given its short length and its protein-bound state in the crystal structure, there is no guarantee it behaves in the same manner when located within the entire protein and not involved in a specific heterologous interaction.

In this study, the area termed the REPO domain (residues 208-239) was seen to be mostly deuterated for both apo and DNA-bound YY1 (above 80 or 90%), but did have some residues that were not exposed to solution where the peptides overlapped (***Figure 35***). The backbone of the residues which have almost no deuterium uptake at one or more time points mostly correspond to the regions of the REPO motif where beta strands are expected. The regions within the strands do not all consistently show low level of uptake – several timepoints in some of the locations also indicate over 90% deuterium uptake. Similarly, the regions with high levels of deuterium uptake correspond to the expected turns of the REPO motif.

Given the changes in deuterium uptake between time points within the REPO motif of the apo protein, it is likely that this motif can exist in and flip between structured and disordered conformations. The REPO motif is conserved in YY1 (Kim *et al.* 2007). Therefore, the formation of structure and disorder within it is likely to be functionally important.

Relatively little change between the apo and DNA-bound forms of YY1 was seen in the difference map for the REPO domain (not counting the low peptide coverage regions) (***Figure 35***). Only one time point in a small stretch within a potential beta strand region displays an increase in deuterium uptake upon DNA binding at two residues. This may result from a change in the structural motif this region may form.

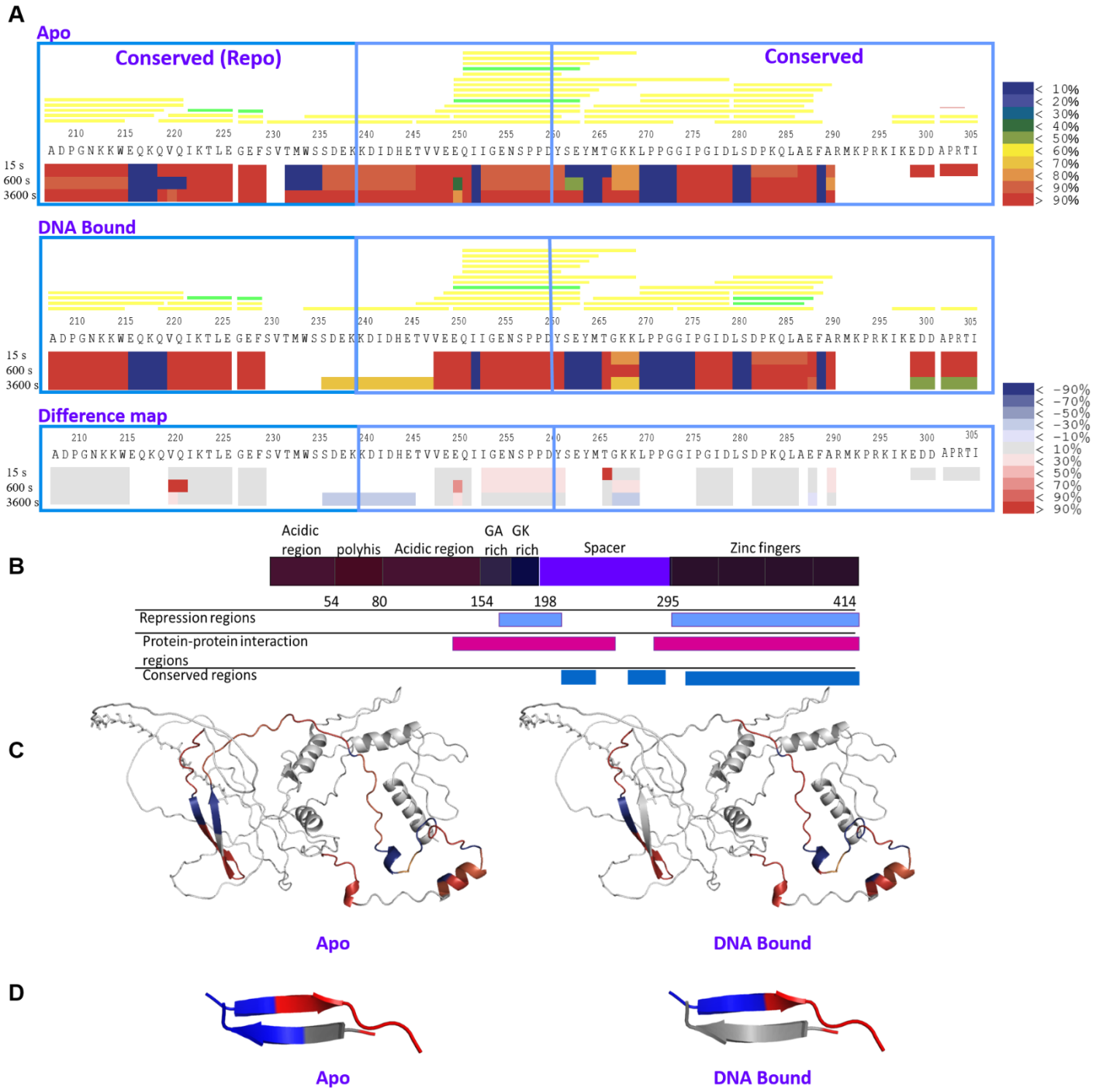


Figure 36: HDXMS heat and difference maps of apo and DNA-bound YY1 in its spacer region.

A) Top shows the heat map for apo YY1, middle the DNA-bound form and below is the difference map. The % of deuteration of the backbone for each residue is represented by the colour scale, ranging from red (more than 90% deuteration) to blue (less than 10% deuteration). The less deuterated a residue's backbone, the more likely it is to be buried whereas more deuterated residue backbones are more exposed to the solvent. The green and yellow bars above the heat map represent the high and medium confidence peptides used to create the heat map. The difference map shows the difference in deuterium uptake between the DNA bound and

apo forms. The difference map is coloured on a scale from red to blue, where red shows a greater uptake of deuterium and blue is a decrease in deuterium uptake upon DNA binding. Small changes in backbone dynamics were seen upon DNA binding. The non-conserved region is largely unstructured, while the second conserved section of the spacer region has distinct pockets of structure. **B)** The location of the GA and GK rich regions within the sequence and its associated functions are shown below the difference map. **C)** The AlphaFold structure of YY1, with the acidic region coloured according to the apo and DNA-bound heat maps at 15 s. This predicted region is generally very low- low confidence on AlphaFold (Jumper *et al.*, 2021; Varadi *et al.*, 2022), which is common for disordered regions, and can only be interpreted as an indication of an ensemble of conformations, as opposed to a location of the backbone within 3D space (Ruff and Pappu, 2021). In this case, some overlap is seen between HDX structure association and the predicted AlphaFold structure. However, the predicted conformations are not fully consistent and show clear differences in the context in which YY1 is being observed. **D)** At the bottom is the crystal structure of the first conserved region (Alfieri *et al.*, 2013), known as the REPO domain, when in complex with a polycomb protein. The structure is coloured according to deuterium uptake for the 15 s time points. Deuterium uptake patterns do support the formation of the REPO motif in apo YY1. Heat and difference maps were generated by HDX-aminer (Sierra Analytics, USA). Structure was rendered using PyMOL molecular graphics system version 1.8, Schrödinger, LLC. PDB code used was 4C5I (Alfieri *et al.* 2013).

Between the two conserved sequences in the spacer lies a less-conserved stretch of residues (residues 240-260). The peptide coverage within this region is varied. Where the coverage is sufficient, however, the majority of this region appears to be highly unstructured in both the apo and the DNA-bound forms of the protein. One residue (Ile251) displayed minimal exposure of its backbone to the solution compared to the rest of this region. As this is a result based on several peptides with the same region of overlap, it suggests that a pocket of structure may exist within this non-conserved region for both apo and DNA-bound YY1. The limited resolution of the section of residues following Ile251 makes the actual size of this pocket difficult to distinguish. However, it does imply that the random coil structure in this section predicted by alphafold is likely not entirely correct (***Figure 35***).The difference map for the non-conserved region with adequate peptide coverage indicates that some increase in deuterium uptake occurred in the non-conserved section of the spacer region upon DNA binding.

The second conserved region of YY1 (residues 261-307) has a mixture of random coil and some helical structure predicted in the alphafold sequence, mostly with low to very low confidence. The HDXMS data produced in this study does not agree with most of these predictions. Indeed, only one short helical motif in the alphafold structure corresponds with the levels of deuterium uptake observed. Other than this, according to the HDXMS data, the second

conserved section of the spacer region contains a mixture of disorder and several of pockets of structure in both the apo and the DNA-bound states of YY1. The pockets of structure are relatively extensive compared to much of the rest of the YY1 sequence, and mostly have minimal deuterium uptake throughout all time points taken. Therefore, this section of the YY1 sequence is likely to contain relatively consistent pockets of structure. Relatively little change upon DNA binding is reflected by the difference map, besides at one residue location for one time point, where an increase in deuterium uptake occurred. Therefore, overall, it does not appear that the dynamics of the second conserved region of the spacer region is greatly affected by DNA binding.

Overall, the spacer region consists of many flexible regions, with some sections of structured pockets, especially in the second conserved area. For the most part, these distinct regions tend to show relatively little change over time, instead remaining either deuterated or non-deuterated for the full time period. This suggests that their conformational ensembles probably only change for very short amounts of time, or very seldom.

5.6.6. DNA BINDING DOMAIN

The YY1 DNA binding domain is where one would expect to see the most changes in deuterium uptake to occur when comparing the apo and DNA-bound forms. This is because DNA binding usually causes some level of structural changes in the domain to which it binds (Górecki *et al.* 2015; Morris *et al.* 2018). Furthermore, the contacts between the DNA and the protein result in the exclusion of water, making the protein backbone less available for hydrogen exchange. Unfortunately, there was very little peptide coverage in the DNA binding domain (***Figure 36***). Pepsin is a relatively non-specific protease, although it does have a preference towards certain amino acids, such as hydrophobic ones (Hamuro *et al.* 2008). Nevertheless, pepsin does require the peptide backbone of the substrate to be accessible in order to cleave it. As such, proteases, including pepsin, cleave disordered, flexible regions of a protein more easily than highly folded regions (Fontana *et al.* 1997). The zinc finger regions of YY1, therefore, are likely to contain more structure compared to the rest of the protein, resulting in less cleavage within this region when under the conditions ideal for proteolysis in the disordered regions. Furthermore, zinc finger motifs have been found to possess a certain amount of resistance to protease activity when bound to metal ions (Berg 1990). Therefore,

although the low peptide coverage does not allow adequate information about the deuterium uptake within this region, it does indicate a clear difference in the structure of the YY1 DNA binding domain compared to the disordered N-terminal region.

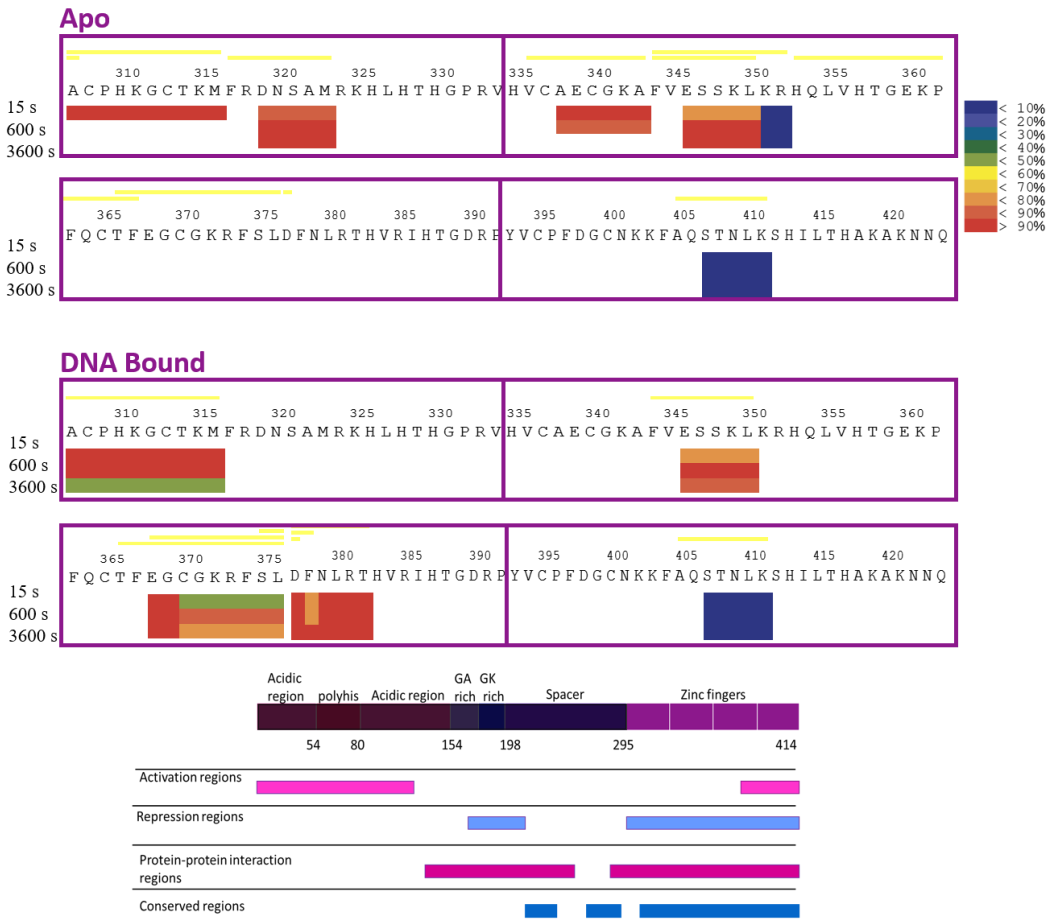


Figure 37: Heat and difference maps of the zinc finger region of apo and DNA-bound YY1.

Top shows the heat map for apo YY, middle the DNA-bound form and below is the difference map. The % of deuteration of the backbone for each residue is represented by the colour scale, ranging from red (more than 90% deuteration) to blue (less than 10% deuteration). The less deuterated a residue, the more likely it is buried whereas more deuterated residues are more exposed to the solvent. The green and yellow bars above the heat map represent the high and medium confidence peptides used to create the heat map. The difference map shows the difference in deuterium uptake between the DNA bound and apo forms. The difference map is coloured on a scale from red to blue, where red shows a greater uptake of deuterium and blue is a decrease in deuterium uptake upon DNA binding. Each zinc finger is shown in a purple box. Due to the lack of peptide coverage no conclusion regarding the dynamics within each zin finger could be made. Therefore, the comparison of the structure with the predicted AlphaFold or crystal structures are not shown. Heat and difference maps were generated by HDX-aminer (Sierra Analytics, USA).

The lack of peptide coverage within this region leaves little confidence in the predicted deuterium uptake, as most regions have only one peptide available, if not two. Furthermore, the length of these peptides results in low resolution. Therefore, no conclusion regarding the dynamics within each zin finger could be made.

5.6.7. OVERALL DYNAMICS

Overall, the HDXMS of YY1 indicates that, firstly, it is a protein with high levels of backbone exposure to the solution. This is not what one would expect from any globular protein, which would consist of many hydrogen bonded peptides and also many residues buried in a hydrophobic interior. Therefore, the HDXMS profile of YY1 strongly supports previous findings that it is not a globular protein (Górecki *et al.* 2015). Little peptide coverage was available in the DNA binding domain, which made it difficult to draw any conclusions about this region which is unfortunate as changes upon DNA binding could therefore not be detected. However, we were able to detect some changes in DNA-bound YY1 in the acid rich regions where, overall, the protein backbone seemed to become more exposed to solution upon DNA binding. This suggests that some kind of structural change is initiated upon DNA binding, which may indicate some involvement between the acid rich region and the DBD in the apo form of the protein.

5.7. YY1-FOXP2 FHD INTERACTION

Heterologous and homologous interactions are important for a transcription factor's ability to function. YY1 has been suggested to interact with a variety of other transcription factors in order to regulate its function and transcriptional activation or repression. One of these potential partners is FOXP2 (Estruch *et al.* 2018), although this has hitherto not been confirmed *in vitro*. In this study, the interaction between YY1 and the FOXP2 FHD was tested in isolation, and the influence of DNA on this interaction was probed in order to ascertain if these two proteins interact and how this might affect their functions.

The YY1-FOXP2 interaction was investigated using fluorescence anisotropy. Increasing concentrations of the FOXP2 FHD were titrated into 0.2 μM NTA ATTO 550-labelled YY1

(**Figure *38***). An increase in anisotropy was observed, implying an increase in the size of the complex due to an interaction being formed between YY1 and the FOXP2 FHD. The nature of a protein-protein interaction can depend on a variety of bonds, including hydrophobic and hydrophilic (Uversky and Dunker, 2010). Ionic strength can be used to probe whether a protein-protein interface may be largely hydrophilic or hydrophobic (Alrosan *et al.*, 2021). Therefore, three different salt concentrations were used in this study: 100, 300 and 500 mM NaCl. The K_Ds calculated for these were 73 ± 5 μM, 28 ± 2 μM and 67 ± 6 μM, respectively. These values are all within the same range of low affinity protein-protein interactions (Kastritis *et al.*, 2011; Jarmoskaite *et al.*, 2020). This suggests that the YY1-FHD interaction is not affected by ionic strength.

Solutions with high ionic strengths tend to disrupt polar interactions between proteins (Alrosan *et al.*, 2021), DNA or other biomolecules. This is because the salt outcompetes the binding partners in high concentrations. Alternatively, a salting out phenomenon may occur where the addition of salt may enhance macromolecular interactions, excluding the molecules from the solvent, which results in precipitation. Previous studies with YY1 at NaCl concentrations exceeding 1 M have been performed and YY1 remained soluble. Since the interaction is relatively unaffected by salt concentration, and no precipitation was observed, it is likely that it is based more on hydrophobic contacts compared to hydrophilic ones. Therefore, the interaction between YY1 and FHD is probably mostly hydrophobic in nature.

In addition to studying the potential interaction between FHD and YY1, the interaction between these proteins in the presence of YY1 DNA was monitored. The FHD was titrated against NTA ATTO 550 labelled YY1 which had background AAV P5 promoter DNA present in saturating concentrations (**Figure *38***). The predicted K_D for this value was similar to the apo samples, indicating that YY1 is capable of binding to the FOXP2 FHD in the presence of DNA. If not, one would expect the affinity between YY1 and the FHD to be significantly affected in some way.

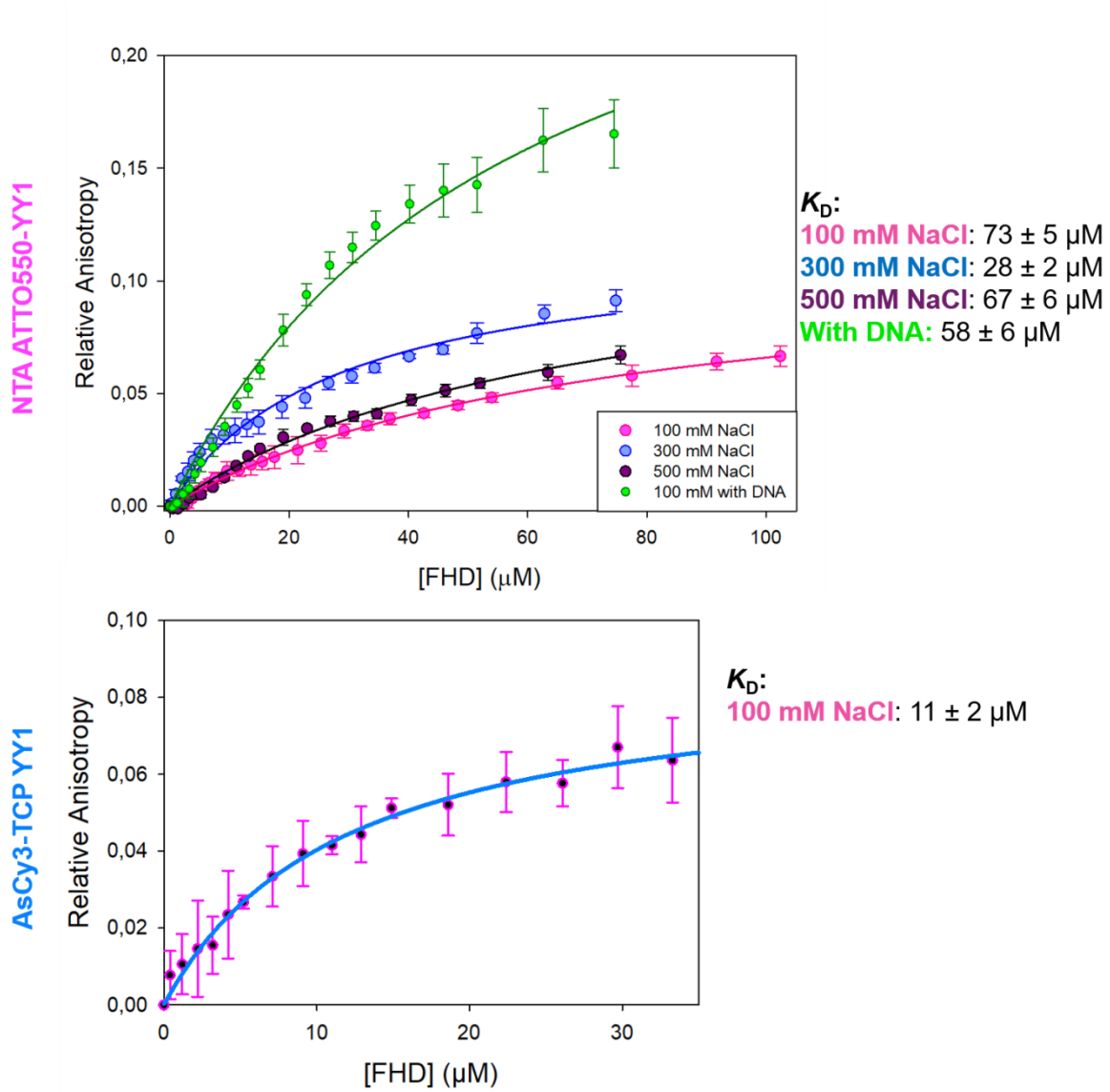


Figure 38: FHD-YY1 binding curves using fluorescence anisotropy.
Top shows 0.2 µM NTA ATTO550 labelled YY1, with increasing concentrations of FHD titrated against it. Each curve shows a different condition, where pink is YY1 in the presence of 100 mM NaCl, blue is YY1 in 300 mM NaCl, and purple is YY1 in 500 mM NaCl. The green is YY1 in 100 mM NaCl and 10 µM AAV P5 DNA. Bottom shows AsCy3 labelled TCP-YY1 with increasing concentrations of FHD titrated in, and at 100 mM NaCl. YY1 does bind to the FOXP2 FHD *in vitro*, and independently of ionic strength and the presence of DNA. Graphs were plotted using SigmaPlot 12.0 and assembled with PowerPoint (2013).

While the affinity between YY1 and the FHD was not greatly altered upon YY1-DNA binding, the relative anisotropy increase was much higher compared to the apo curves. This indicates that there was a greater increase in size when DNA was present. There are several possibilities that might explain this. Firstly, YY1 might bind in a different stoichiometry when DNA is

present. This might include more FHD molecules binding per YY1, or it may involve the bound FHD being able to recruit more YY1 molecules. Secondly, the initial size of YY1 bound to DNA might be less than that of the apo form, if the apo YY1 forms larger oligomers which could be disrupted by DNA. This would result in the change in size upon FHD binding to result in a larger size change, thus the change in anisotropy would increase because the net change in size would also increase.

In order to ensure that the anisotropy increase that is seen for all YY1-FHD curves was not due to an effect from the label (explained in section 4.8.5), a second FHD-YY1 binding curve was done, using the TCP-YY1 construct labelled with AsCy3 instead of NTA ATTO550 (**Figure *38***). This curve, like its ATTO550-YY1 counterpart, did indicate an interaction between FHD and YY1. This confirmed that the trend was relatively reliable. Compared to the ATTO labelled curve, the AsCy3 curve does have some differences: firstly, it suggested higher binding affinity, at a K_D of 11.8 ± 2 μM compared to 73±5 μM. Secondly, the AsCy3 label did not have the same fluorescence lifetime, which resulted in more noise in the data. The lower K_D is still in the 10s of μM range, which means that overall, it is not necessarily a significant difference, given the different label used – the alteration in the technique is capable of this degree of change (Jarmoskaite *et al.*, 2020). Additionally, the greater noise in the AsCy3-TCP-YY1 curve resulted in a poorer fit to the binding curve compared to the ATTO550 labelled YY1. This was likely to have introduced sufficient error for the apparent difference in binding affinity to be insignificant. Overall, this curve does serve to confirm that the binding interaction observed is likely to be real, and is also in the same range of concentrations of FHD as with NTA ATTO550 labelled samples.

5.8. YY1-FHD INTERACTION INFLUENCE ON DNA BINDING

Transcription factors often rely heavily on heterologous interactions to regulate their DNA binding (Li *et al.* 2004). This can include the protein-protein interaction interfering with a protein-DNA interaction, enhancing it, or not affecting it and instead functioning in an alternative manner (Gordon *et al.* 2006). In addition, some protein-protein complexes form ternary structures with the DNA, and some do not (Gordon *et al.* 2006). How one protein affects or does not affect the DNA-binding ability of another protein can indicate why they might interact. Therefore, a supershift EMSA was done with YY1 specific and FHD specific DNA sequences, in the presence of both of the proteins.

Although the YY1-FHD interaction was found to not be affected by DNA in the previous section, the effect on YY1 DNA binding could not be seen, because the technique was used to probe changes in the overall YY1 complex size, instead of looking directly at the YY1-DNA interaction. To observe if increasing FHD concentrations might affect the YY1-AAV P5 DNA interaction, a supershift assay was done. YY1 concentration was increased as in a normal EMSA across four lanes, which were repeated three times. In the first repeat, no FHD was added. In the second repeat, a lower FHD concentration was added, and in the third set, a higher concentration of FHD was present. Whether or not the FHD was present, YY1 was still seen to bind to its DNA in the same pattern of bands (**Figure *39***). Lanes where FHD has been added to the YY1-DNA complex have higher levels of smearing, which might indicate some non-specific binding between the FHD and DNA, or might indicate a slightly altered on and off rate of YY1 to the DNA.

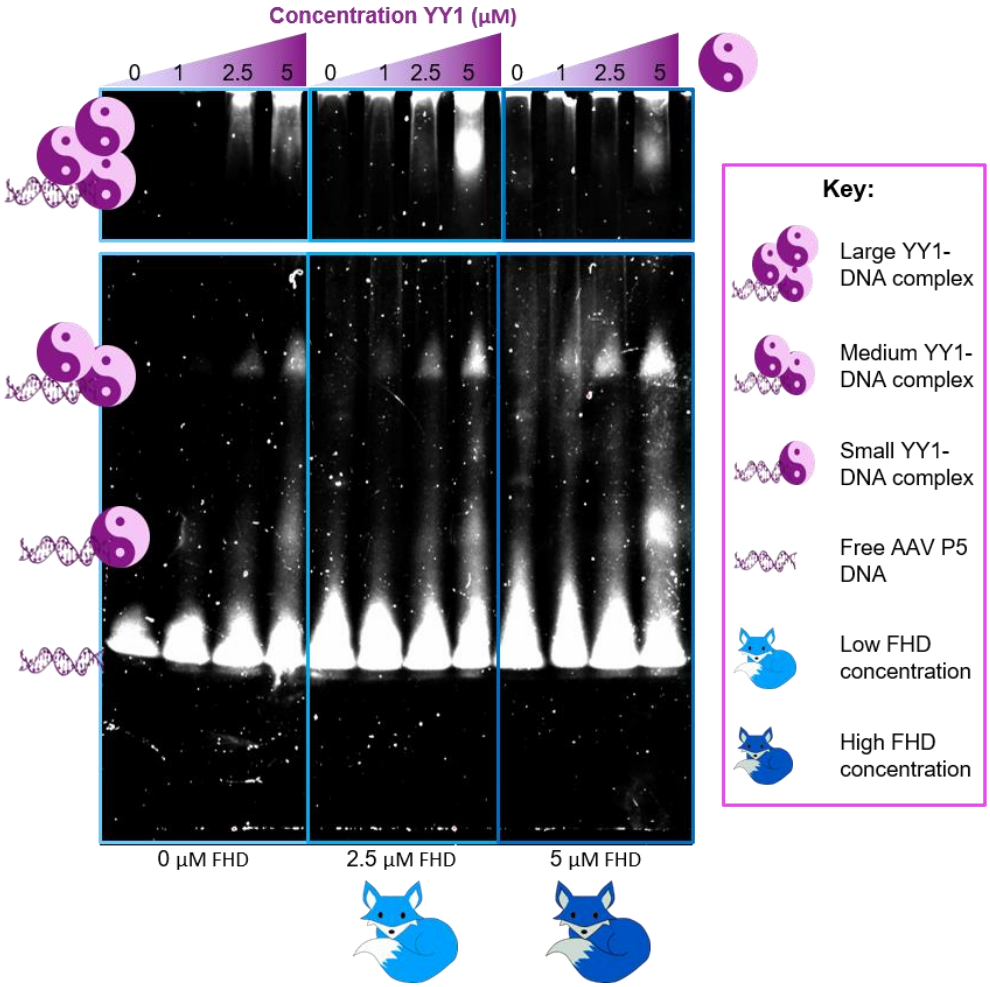


Figure 39: Supershift assay to determine the effects of the FOXP2 FHD on the YY1-AAV P5 DNA interaction.

This figure is made up of two gels which were resolved under exactly the same conditions using the same samples. The top section shows large complexes which required a longer time for resolution, resulting in the free DNA bands no longer having full visibility. The lower section (highlighted in boxes) is the corresponding gel with shorter running times, allowing better visibility of the free-DNA bands. From the left to the right, each box contains a repeat of increasing YY1 concentrations across the lanes. Free DNA, protein-DNA and protein-protein DNA complexes are shown by the symbols on the left where the black yin yang sign represents DNA that is specific to YY1 and the purple yin yang sign represents YY1 protein. Within each box, the FHD concentration is kept constant. The FHD concentration is increased in each box as shown by the darker blues. The concentration of AAV P5 DNA was kept constant throughout. From the gel it can be seen that YY1 forms more than one oligomeric species when bound to DNA and the formation of these oligomeric species is influenced by the presence of FHD. Figure made using InkScape 1.0 and PowerPoint (2013).

The greatest change seen in this assay is the band seen at the top of the gel which represents a very large complex. For this band to be visible, it was necessary for the free DNA to be run off

the gel. The presence of FHD affected both the intensity and position of this band. For both the low and high concentrations of FHD used, the band at the top of the gel migrated further down compared to the band where only YY1 and DNA were present. This slightly lower band in the YY1-FHD-DNA lanes was more intense at low FHD concentrations than high FHD concentrations. This suggests that FHD does bind to YY1, but not does enhance or inhibit DNA binding. Instead, it indicates that low concentrations of FHD might result in a FHD-YY1-DNA complex, and that the YY1 component may have an altered oligomeric structure upon FHD binding. This altered oligomer is, however, speculation based on the indication that a differently sized band is formed when FHD is present.

Additionally, this higher oligomeric species only appears to be present when the concentrations of both the FHD and YY1 are high. This could indicate different things. Firstly, the simplest explanation would be that YY1 binds both FHD and DNA simultaneously when both proteins are present in high concentrations. Additionally, FHD might bind YY1 in multiple places, resulting in the formation of larger complexes. Secondly, at high FHD concentrations, FHD might compete with YY1 for DNA or with the DNA for YY1. This would explain the formation of a slightly smaller sized complex observed in the highest concentration lanes of YY1 in the presence of FHD. Thirdly, it could be a mixture of the factors, especially considering the heterogenous nature of intrinsically disorder proteins.

A second supershift assay was done to explore how the presence of YY1 might affect the FHD-CNTNAP2 DNA interaction. In this case, the FHD was increased over four lanes, which was then repeated in the presence of increasing concentrations of YY1 (low, medium and high). As with the previous assay, much higher bands were only clearly visible when the free DNA was run off the gel. Therefore, two gels were used for full visibility (***Figure 40***).

When no YY1 is present, there is only a single band containing FHD and its specific DNA. Interestingly, when YY1 is added and there is low or no FHD present there is a band visible at the very top the gel. This band disappears when the FHD concentration is increased. This band is likely the result of a non-specific interaction between YY1 oligomers and the CNTNAP2 DNA. The disappearance of this band as the FHD concentration increases is a result of FHD outcompeting YY1 for binding to its cognate sequence, resulting in decreasing amounts of YY1-DNA complexes, and, therefore, less intense bands. This indicates that while YY1 seems to be able to bind both DNA and the FHD simultaneously, the FOXP2 FHD is unable to bind to YY1 and DNA at the same time.

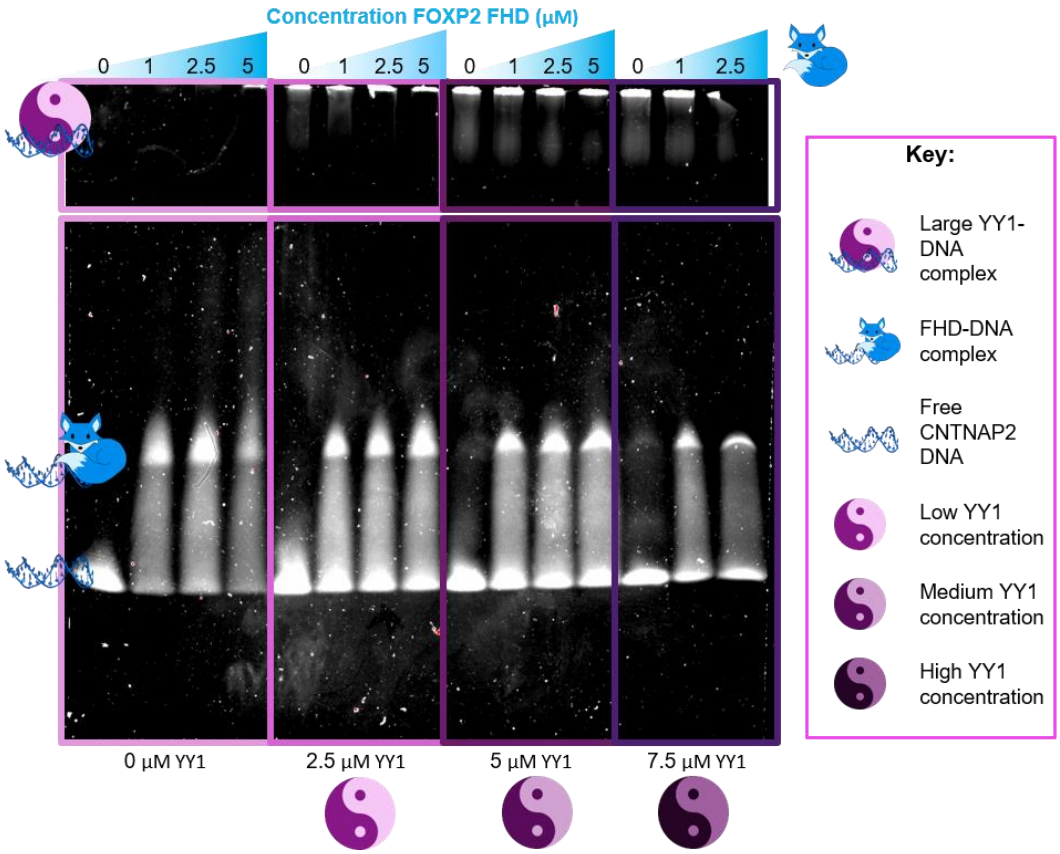


Figure 40: Supershift assay showing the influence of YY1 on the FHD-CNTNAP2 DNA interaction.

This figure is made up of two gels which were resolved under exactly the same conditions using the same samples. The top section shows large complexes which required a longer time for resolution, resulting in the free DNA bands no longer having full visibility. The lower section (highlighted in boxes) is the corresponding gel with shorter running times, allowing better visibility of the free-DNA bands. FHD concentrations are increased every four lanes, starting with no FHD, then repeated. Background YY1 is added in increasing concentrations in each box, with none present in the first, as represented by the yin yang symbols. The concentration of CNTNAP2 DNA is kept consistent throughout. As the higher complexes were too large to enter the gel before the free DNA migrated off the bottom, this image is composed of two gels combined so as to observe both large and small complexes. Free CNTNAP2 DNA is represented by the symbol of the brown fox and DNA. FHD binds CNTNAP2 DNA in a clear distinct band present in the middle of the gel as represented by the symbol of the brown and blue fox and DNA. YY1 interacts non-specifically with DNA, particularly at higher YY1 concentrations, seen as a faint band at the top of the gel represented by the fox, yin yang symbol and DNA. This band disappears as the FHD concentration is increased because FHD likely outcompetes YY1 for the DNA. Figure made using InkScape 1.0 and PowerPoint (2013).

DISCUSSION

YY1 is a transcription factor associated with multiple functions in the regulation of gene expression. The diversity of the problems which can arise from either altered expression of YY1 or mutations within it seem to implicate this transcription factor in a diverse range of functions (Verheul *et al.*, 2020). For example, it can act as either a tumour suppressor or stimulator (Sarvagalla *et al.*, 2019). This function seems to be context specific, although some tissue types seem more frequently associated with one function over the other (Sarvagalla *et al.*, 2019). Another example includes how YY1 is overexpressed in the development of Alzheimer's disease but under-expressed in Parkinson's disease (Aubry *et al.*, 2015; Pal *et al.*, 2016).

Depending on the specific context, these changes in function clearly indicate the underlying complexity with which YY1 controls expression on a molecular level. How and if it binds to DNA, RNA and other proteins is likely tightly linked to the cellular and situational specific interactional partners it has, PTMs and also the level of YY1 expression (Uversky and Dunker, 2010; Weintraub *et al.*, 2017). These interactions and modifications introduce a diverse range of potential assemblies, which are responsible for different functions. In this way, the interactions YY1 makes are very likely what allow YY1 to operate in such a ubiquitous manner, where it is involved in so many functions.

Therefore, to understand how YY1 influences gene expression, it is necessary to try and learn how it behaves by itself, in the presence of DNA and in the presence of its interacting partners. By slowly understanding how it behaves in isolation and with its other interacting partners, the exact mechanism by which it assembles different complexes, causes DNA structural alterations (Weintraub *et al.*, 2017), induces and represses transcription (Wang *et al.*, 2015), and is involved in X-inactivation can be understood (Jeon and Lee, 2011). This is important for knowing how to potentially develop drugs or other therapies to treat the myriad of conditions with which this protein is involved.

6.1. YY1 STRUCTURE AND INTRINSIC DISORDER

YY1 has previously been found to have characteristics of an intrinsically disordered protein, especially within the N-terminal region (Górecki *et al.*, 2015). Intrinsic disorder is a relatively general term, and can occur in a variety of degrees (Uversky and Dunker, 2010). Molten globule structures, pre-molten globule and true random coil all fall under intrinsic disorder (Uversky and Dunker, 2010). The specific behaviour of a disordered region is inferred by its sequence (Dunker *et al.*, 2001). Within a structurally disordered protein, some regions might behave differently from others depending on the sequence characteristics. One of the important aspects of intrinsically disordered proteins is their ability to form pockets of structure, which are frequently transient (Mohan *et al.*, 2006). Molecular recognition features are sections within a disordered region capable of forming structure, often upon a binding event or due to post-translational modification (Mohan *et al.*, 2006). Because of the heterogeneity of structure and disorder in an intrinsically disordered protein, it is important to understand where the order and disorder are located in the relevant conditions.

The CD spectra obtained for YY1 in this study agree with those which have previously been reported (Górecki *et al.*, 2015; Chen *et al.*, 2019), having troughs mostly around 200-210 nm (***Figure 25***). This is typical of intrinsically disordered proteins (Uversky, 2009), indicating that YY1 has a relatively large amount of disorder.

The lack of peptides produced within the DNA binding region during the HDXMS study strongly suggests that the zinc finger region is more structured compared to the rest of the protein, whether bound to DNA or not. Conversely, the N-terminal region is cleaved into many smaller peptides, as seen in the HDXMS data here (***Figure 33*-Figure *37***) and by the lack of SDS PAGE bands after proteolysis in other studies (Houbaviy and Burley, 2001; Górecki *et al.*, 2015).

The degree of disorder throughout a sequence can be variable and is likely to contain pockets of structure (Mohan *et al.*, 2006). Both the intrinsic fluorescence data of YY1 (***Figure 28***) and the HDXMS (***Figure 33*-*Figure 36***) data indicate that YY1 does contain sections in which some structure is present.

The intrinsic fluorescence of YY1 shows a shift in the environment of both tryptophan residues (***Figure 28***), which are located in the REPO domain, upon denaturation. The HDXMS data shows that at least some of the beta strand of the REPO motif is likely to be structured. However,

at some time points, sections of this region become more disordered (***Figure 36***). The intrinsic fluorescence and HDXMS data together, therefore, indicate some degree of structural formation of YY1 within the REPO motif, which shows that it is not entirely dependent on heterologous interactions to form structure. The formation of the REPO motif in the absence of its polycomb group binding partners may facilitate higher affinity binding compared to what is usually associated with intrinsic disorder. As it is likely that the REPO motif is frequently already formed, less energy will be required upon binding than if it were dependent on interacting with a partner to become structured. Therefore, it would be easier for YY1 to recruit the polycomb proteins specifically compared to other, more non-specific interacting partners of YY1.

Besides the REPO motif, the structured regions in the apo form of YY1's acidic regions seem to be most frequent (***Figure 33***, ***Figure 34***), as well as some of the other conserved sections of the spacer region (***Figure 36***). The structure seen in the acidic region of YY1 is unexpected, as this type of sequence is usually strongly associated with disorder (Dunker *et al.*, 2001). The increased structure, therefore, may be contributed by YY1-YY1 interactions, indicating an interface and induced structure upon binding (discussed further below in section 6.3). The pockets of structure which were observed suggest that several molecular recognition features are present in this region, which may, therefore, be more highly involved in protein-protein interactions than previously assumed. Previous CD and *in silico* predictions have suggested that the YY1 sequence has the propensity to form beta sheets (Górecki *et al.*, 2015). Therefore, several of the pockets of structure observed in the HDXMS in this study are likely to form this motif.

The second conserved section of the spacer region contains a relatively large portion of structure in comparison to the rest of the protein (***Figure 36***). This region has previously been implicated in protein-protein interactions (Gordon *et al.*, 2006). For example, TFIIB, part of the transcriptional pre-initiation complex, is involved with binding to this region (Thomas and Seto, 1999). As seen in the REPO motif, the structured residues might facilitate more specific interactions than disordered regions.

Overall, using HDXMS, the specific regions in YY1 which are more structured or more disordered have been identified. This helps to clarify which sections within YY1 may be involved with different types of interactions. The disordered regions likely allow for non-specific, low affinity, heterogeneous interactions, or simply allow protein flexibility (Uversky,

2019). The pockets of structure within the disorder may well be important for more specific binding events. Alternatively, some of the structured pockets may be formed in the apo protein as a result of a YY1-YY1 interface.

6.2. EFFECT OF DNA BINDING ON YY1 STRUCTURE

The structural content of an intrinsically disordered protein is dependent on several factors, which include binding partners (Mohan *et al.*, 2006). For some intrinsically disordered proteins, large structural shifts have been observed upon certain binding events, including a complete shift from disorder to structure (Rice *et al.*, 1997; Permyakov *et al.*, 2003; Uversky and Dunker, 2010). Previous models regarding YY1 transcriptional repression and activation have included possible mechanisms where YY1 acts either by protein-protein interactions only or by binding to DNA (Gordon *et al.*, 2006). Therefore, it is important to know whether YY1 experiences a shift in its structural composition upon DNA binding, and then to further determine how this might play a role in the control of gene expression.

Several of the structural studies in this work indicated that YY1 does indeed experience a shift in structural composition in the presence of DNA. Firstly, the Far UV CD spectra of apo and DNA-bound YY1 indicate a clear shift in secondary structural composition upon DNA binding (***Figure 25***). Secondly, the HDXMS data clearly indicate that there is a shift in the structural composition of the N-terminal region of YY1 upon DNA binding (***Figure 33*****Figure *34*** and **Figure *36***). Specifically, the most notable differences were an increase in the backbone's exposure to solution within the acidic regions of YY1 when bound to DNA (***Figure 33*** and **Figure *34***). The overall negative charge of the acidic regions makes it possible that it interacts with the more positive DNA-binding domain of YY1 in the apo form. Upon DNA binding, the acidic region would be excluded, and its molecular recognition features might be more accessible to other binding partners. Furthermore, molecular recognition features are characteristically context-dependent (Uversky and Dunker, 2010). One region of an intrinsically disordered protein may form these pockets of structure when specific partners are present, but not others (Uversky and Dunker, 2010). Therefore, the changes in the backbone dynamics of YY1 upon DNA binding likely reflect an alteration in the molecular recognition features present in this region.

The spacer region of YY1 showed a few changes in its backbone dynamics upon DNA binding (***Figure 36***). However, the fluorescence spectra of the apo and DNA-bound YY1 indicate no major change in the environment of the tryptophan residues in the spacer region (***Figure 28***). As such, the spacer region experienced some shifts in structure which did not result in large environmental changes. The spacer region has previously been found to be important for protein-protein interactions, and is directly adjacent to the DNA binding domain (Thomas and Seto, 1999). While slight, the changes in backbone dynamics within this region could be important for the recognition of binding partners.

Overall, DNA binding was shown to alter YY1's structure in the N-terminal region outside of the DNA-binding domain. The altered secondary structure and dynamics indicates that the molecular recognition features present in the apo form of the protein likely differ from those in the DNA-bound form. The small pockets of structure which make up molecular recognition features are responsible for many of the intermolecular interactions of intrinsically disordered proteins (Mohan *et al.*, 2006). Therefore, changes in the structural composition of YY1 in the acid rich and spacer regions upon DNA binding indicate a potential shift in binding partners when the protein is by itself and when it is bound to DNA.

6.3. YY1 DNA BINDING

YY1 binds to DNA via its zinc finger domain, which consists of four C2H2 zinc finger motifs (Houbaviy *et al.*, 1996). As a transcription factor, the ability of YY1 to bind to DNA is integral to its function. In particular, the DNA binding domain is vital for the protein's ability to repress transcription (Thomas and Seto, 1999). Extensive research has already been done on YY1-DNA interactions (Houbaviy and Burley, 2001; Golebiowski *et al.*, 2012; Weintraub *et al.*, 2017; Chen *et al.*, 2019). However, there is a wide disparity in the literature regarding the affinity of the YY1-DNA interaction. Furthermore, very little has been established between structural characteristics such as the oligomeric state of YY1 and its DNA binding.

The K_D of the YY1- AAV P5 promoter DNA interaction was found to be 3 μM in this study (***Figure 32***). The K_Ds observed here and in previous studies range quite significantly (from low nanomolar to low micromolar) (Gualberto *et al.*, 1992; Lace *et al.*, 2009; Golebiowski *et al.*, 2012). Some of this may originate from differences in technique. However, the large degree of variation may stem from differences in the oligomeric profile of YY1 between studies.

Unfortunately, not many DNA-binding studies of YY1 have probed YY1 quaternary structure, which could be significant in altering DNA-binding affinity and specificity (see section 6.3).

6.4. YY1 OLIGOMERISATION IN CONJUNCTION WITH DNA BINDING

Evidence in this study from SEC (***Figure 31***) and fluorescence anisotropy (***Figure 32***) clearly indicates that YY1 does form oligomers of various sizes in a concentration dependent manner. In terms of approximate size, these oligomeric species eluted at similar times as was found in previous studies which observed homologous YY1 interactions – however, several other studies have only observed one species (Golebiowski *et al.*, 2011; López-Perrote *et al.*, 2014). One potential reason for this is a difference in post-translational modifications, which is a factor known to heavily influence the behaviour of intrinsically disordered proteins (Fuxreiter, 2012; Hsu *et al.*, 2013). According to our SEC results, there are species present in purified YY1 of about 50, 75 and 150 kDa, and then there are also larger species that eluted within the void volume of the column. Previous speculation has identified this very large peak in void volume to be bodies of aggregate (López-Perrote *et al.*, 2014). However, in this study, the isolated void volume peak re-equilibrated back into the original peaks when size excluded again. Furthermore, the protein that was isolated from the void volume retained its DNA binding activity (***Figure 22***). This functional ability of these large oligomers may suggest that rather than just being amorphous aggregates, they are large, soluble complex oligomeric functional species capable of activity. While these large YY1 complexes could be an artefact of *in vitro* behaviour, the functional role of YY1 and *in vivo* observations elsewhere (Weintraub *et al*., 2017) indicate a strong potential physiological relevance, as discussed in detail in this section.

YY1 is involved in recruiting the polycomb group of proteins, which form polycomb bodies (Alfieri *et al.*, 2013). Additionally, it has been found in nuclear speckles (Salichs *et al.*, 2009). Both of these structures are proteinaceous membrane-less organelles, which are essentially heterogenous, liquid-like functional regions in the cell, containing high concentrations of specific proteins and nucleic acids (Uversky, 2019). Proteinaceous membrane-less organelles are formed via liquid-liquid phase transitions, which are driven by many factors, including the characteristics of their members (Ambadipudi *et al.*, 2017; Uversky, 2019). These members are frequently intrinsically disordered proteins, or are proteins with a mixture of disorder and structure (Banani *et al.*, 2017). Properties which are often associated with these proteins are their RNA-binding capacities, sequences containing repeats of polar and charged amino acids, low complexity domains and weak multivalent interactions (Uversky, 2019). These multivalent

interactions can be between the same molecule and also with other molecules (Banani *et al.*, 2017; Uversky, 2019).

Given that YY1 is involved in at least two of these structures, and has many of the properties of proteins involved in their formation (Alfieri *et al.*, 2013; Górecki *et al.*, 2015; Uversky, 2019), this gives the high levels of YY1 multimerisation some more context. Soluble aggregation is still a possibility, but YY1 is known to be involved in liquid-liquid phase transitions which require weak, heterogeneous and multivalent interactions (Wilkinson *et al.*, 2006; Banani *et al.*, 2017). Simultaneously, the proteins involved in these phase transitions must remain dynamic at high concentrations (Fuxreiter, 2012), which are often 10-100 times higher than outside of a proteinaceous membrane-less organelle (Milo *et al.*, 2010). Indeed, proteins involved in proteinaceous membrane-less organelles can be pathogenic when they lose their dynamic natures within these structures, forming rigid aggregates which are no longer capable of maintaining the liquid or liquid-like state required (Uversky, 2019). These are associated with problems such as neurodegenerative diseases, including Alzheimer's and Parkinson's disease – both of which YY1 is associated with when its expression is mis-regulated (Aubry *et al.*, 2015; Li *et al.*, 2016; Ambadipudi *et al.*, 2017; Liu and Fang, 2019). As such, the heterogeneity of YY1's self-association, as well as its sensitivity to salt, DNA and concentration (discussed below) does make sense.

The void volume peak from the YY1 SEC increased in intensity with increased salt concentrations, while peaks of smaller size decreased (***Figure 31***). Furthermore, there was a distinct difference in YY1 oligomerisation seen in the anisotropy curves studied under low and medium concentrations of NaCl (100 and 300 mM) compared to the 500 mM salt solution (***Figure 32***). The higher salt concentration resulted in a smaller overall change in the average oligomeric size of YY1 in solution. When taken in conjunction with the larger amount of oligmerisation observed when performing SEC under high salt conditions, this can be interpreted not as a lower affinity, but instead as an increased propensity for homologous interactions, where the higher salt concentration increased oligomerisation. If high ionic strength encourages self-association, then the initial low concentrations of YY1 in the anisotropy curve would have greater average oligomeric size compared to the YY1 in low ionic strength buffers. As such, the increased salt concentration would act on the solution in a similar way to increased YY1 concentration. This would lead to a smaller change in average oligomeric size due to a shift having occurred already in the oligomeric equilibrium. Therefore,

concentration and ionic strength both seem to be strong influencing factors on YY1 oligomerisation with essentially the same effect.

This increase in average oligomeric size at high salt concentrations implies that the YY1-YY1 interaction has some dependency on hydrophobic interactions. If it relied on hydrophilic contacts, the increased salt would be more likely to disrupt the oligomeric interface, due to the salt outcompeting YY1 for itself. The similarity in the YY1-YY1 binding curves at either 100 mM or 300 mM NaCl (***Figure 32***) indicates that this effect is limited to large ionic strength alterations and the relationship between salt concentration and oligomeric size is not linear.

While salting out is a possible explanation for the effect of high salt concentrations on YY1 structure, it is unlikely that the salt causes YY1 to form amorphous insoluble aggregates because YY1 did tend to show the ability to retain high levels of activity and stability over time (fewer insoluble aggregates and better DNA binding activity) when stored in buffers with high ionic strength (500 mM NaCl as opposed to 100 mM NaCl). As such, the aggregate-like structures observed by López-Perrote *et al*. (2014) may either be simply reversible according to buffer and concentration conditions, or may be heterogeneous multimeric structures resulting from weak self-association. As mentioned above, these may have a relationship with YY1's function in the formation of proteinaceous membrane-less organelles.

Previous studies have indicated that large YY1 oligomers purified with SEC are capable of binding to DNA but that they seem to have decreased specificity for the nucleic acid (López-Perrote *et al.*, 2014). However, these studies did not take into account various YY1 concentrations, use appropriate gels to resolve very large complexes or investigate the relationship between the oligomeric state of YY1 and its DNA bound state. The supershift assay with FOXP2-specific CNTNAP2 DNA confirms that large YY1 multimers bind to DNA non-specifically. This is seen on the supershift gel in ***Figure 40*** as fainter bands at the very top of the gel, in lanes containing only YY1 and FHD-specific DNA. Furthermore, the free DNA in both the supershift and the EMSA gels is smeared, which is characteristic of a non-specific interaction. The supershift assay allowed for the entrance of the larger YY1-DNA complexes in the gel, unlike the regular EMSA (***Figure 21***). The particular non-specific binding detected here is especially significant because YY1 does interact with the CNTNAP2 promoter at a different binding site (Chiocchetti *et al.*, 2015). It is possible that YY1 oligomers are, therefore, not entirely non-specific in terms of DNA binding to this sequence, but have an altered range.

Additionally, when allowed to bind specifically to its own DNA, several oligomeric species of YY1 are formed (***Figure 39***). Two smaller species are apparent, with the larger one being more consistent, while the smaller at first increases with increasing YY1, then decreases in intensity again at the highest YY1 concentrations. Already, this implies the existence of a range of YY1-DNA stoichiometries. In cases where gels were given adequate time to resolve, a much larger complex was observed in addition to the smaller species. This band in the AAV P5 gels increases with increasing YY1 concentrations (***Figure 39***).

These very large YY1 oligomers seen binding to DNA in the EMSAs explain the unusual linear shape of the anisotropy curve that measured the binding of YY1 to DNA (***Figure 23***). As YY1 is added to DNA in increasing concentrations, there is an initial interaction. After it has reached saturation and little to no free DNA is available, YY1 simply begins to form larger DNA bound oligomers. The result is the transition from a mixture of smaller oligomeric species bound to DNA into larger oligomeric species bound to DNA (***Figure 41***). The formation of larger complexes in the EMSA gel only as the concentration becomes increased suggests that saturation of the DNA is preferred before large complexes begin to form, rather that oligomerisation occurring concurrently with DNA binding (***Figure 39***).

This large oligomeric formation, however, seems to be disrupted when more DNA becomes available, where by about 6 μM of DNA the majority of the YY1 oligomers seemed to have experienced the decrease in overall complex size (***Figure 23***). This would explain the shape of the anisotropy curve which was constructed in the opposite way: titrating DNA titrated into labelled YY1. This curve showed a decrease in relative anisotropy with increasing DNA concentration. Therefore, the overall size of the originally labelled YY1 decreases with increasing DNA concentration. From the SEC profile, YY1 is seen to be made up of a variety of oligomers, some of them very large (***Figure 31***). Therefore, even at the lower concentration of YY1 used in anisotropy (0.2 μM), YY1 is likely to still contain a mixture of oligomers. This is supported by the observation of YY1 complexes by Weintraub *et al*. (2017) at low concentrations. The decrease in anisotropy strongly implies that DNA breaks up the YY1 oligomer, as the protein prefers to bind to DNA with lower stoichiometry (***Figure 41***).

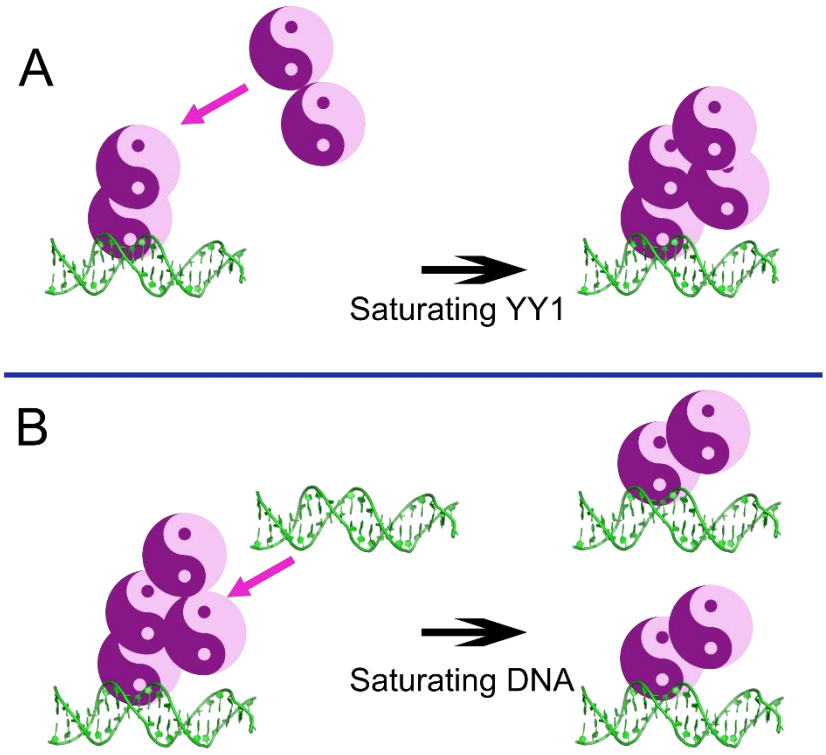


Figure 41: Summary of YY1 oligomerisation and DNA binding behaviour.
A) When YY1 saturates DNA, it continues to form larger DNA-bound oligomers. **B)** When DNA saturates YY1, the complex size gets smaller overall, indicating a loss in oligomeric state or a decrease in oligomeric size. Please note that the YY1 oligomers are representations of the overall situation, not an estimate of the actual oligomeric species involved. This is a phenomenon observed with YY1 and its specific DNA. Non-specific DNA may differ. Image made using Inscape 1.0.

To summarise, when YY1 saturates DNA, it forms large DNA-bound oligomeric species (***Figure 41***). On the other hand, when DNA is saturating, YY1 complexes dissociate in order to bind to the free DNA. This was observed primarily with YY1-DNA fluorescence anisotropy (***Figure 23***), EMSAs (***Figure 21*** and **Figure *39***) and is supported by the HDXMS profile (***Figure 33*** and**Figure *34***) (discussed further below). The interplay between YY1 oligomerisation and DNA binding raises the question of where the YY1 oligomeric interface is located.

The dissociation of YY1 oligomers upon DNA binding suggests that the DNA binding domain of YY1 is important for these YY1-YY1 contacts. When looking at functional regions of YY1, the C-terminal zinc finger domain is important to YY1 transcriptional repression activity, while the N-terminal around the acidic regions is associated with its transcriptional activation

function (Thomas and Seto, 1999; Gordon *et al.*, 2006). It has been suggested that some of YY1's regulatory mechanisms may involve the N-terminal activation region masking the DBD, and opening up under conditions where YY1 needs to bind DNA (Gordon *et al.*, 2006). This model was proposed by Thomas and Seto (2006) when the transcriptional activation activity of YY1 was eliminated upon removal of the acidic regions, and YY1 could bind to DNA as a repressor.

The difference in deuteration in the acidic regions upon YY1 DNA binding seen in HDXMS (***Figure 33*** and ***Figure 34***), increases the plausibility of this hypothesis, as very little evidence has previously supported it. The acidic regions in the DNA-bound form of YY1 generally displayed an increase in deuterium uptake compared to the apo protein. This implies a decrease in structure, and could also be a result of this region becoming free from other inter- or intra-molecular interactions. Given the decrease in complex size upon DNA binding based on the anisotropy results (***Figure 23***), the same behaviour should be occurring in the HDXMS, where the DNA-bound form should consist of either fewer or smaller YY1 complexes. Therefore, the changes seen between the two forms of YY1 are very likely to result from the loss of an oligomeric interface. As such, the increase in availability of the acidic regions to the solution results in the dissociation of YY1 oligomers (***Figure 42***).

The complementarity of the overall charges of the acidic region (being negative) and the DBD (being mostly positive) supports this. If the acidic region of one YY1 molecule binds to the zinc finger region of another YY1 molecule, then much of the relationship between YY1 oligomers and DNA can be explained. When bound to DNA, the zinc finger of one YY1 molecule becomes freed from the acidic region due to better affinity for DNA. Therefore, oligomeric complexes resulting from this interaction are dispersed. When YY1 is saturating the DNA, however, it can still bind to itself. The DNA-bound form of the protein has a free acidic region (***Figure 42***). This acidic region is now available to bind the zinc finger domain of another YY1 molecule, and that molecule's acidic region can bind to another's zinc finger region. As such, very large oligomeric complexes could be formed, which would form relatively easily into a range of large sizes due to the flexibility of the protein originating from its intrinsic disorder.

Weintraub *et al*. (2017) observed that YY1 does self-associate around enhancer-promoter loops of DNA. However, the exact size of these oligomers was not ascertained. López-Perrote *et al*. (2014) also observed YY1 multimerisation in the cell, but not the effect thereof, nor the exact

sizes. They did observe what was likely to be a bell shaped dimer which could further associate into a tetramer using electron microscopy (López-Perrote *et al.*, 2014). This specific elongated orientation of the protein and the pairwise increase in oligomerisation might be explained by the acidic region's interaction with the DBD, which would require pairs to start with, and may well further associate. As such, the specific structural changes seen in YY1 upon DNA binding provides valuable insight into how YY1 oligomerisation and DNA binding regulate its function.

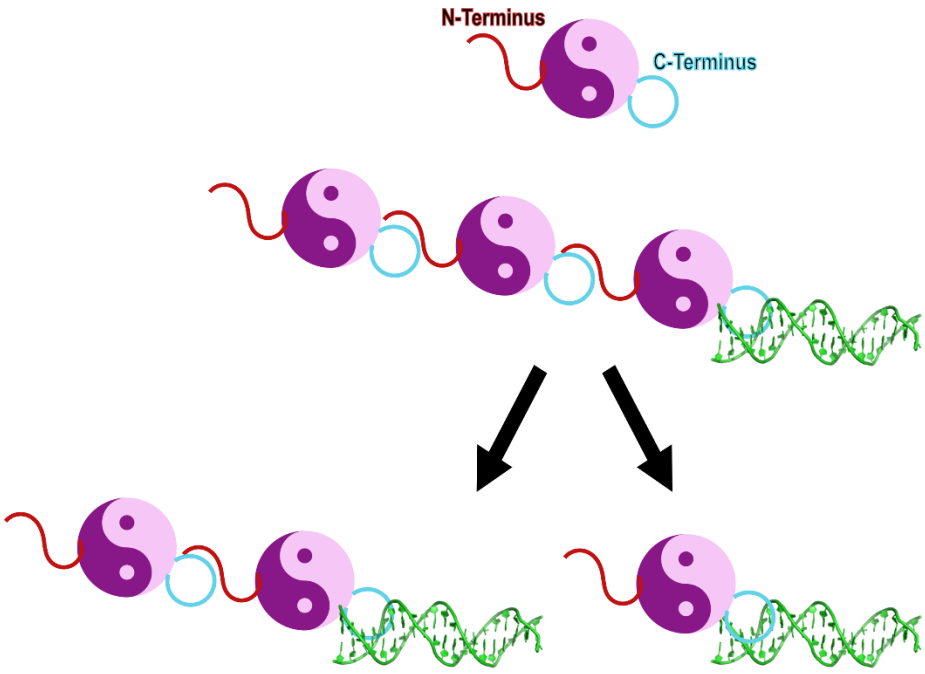


Figure 42: Model of how YY1 might oligomerise and how the presence of DNA may make it dissociate.
The acidic N-terminal region binds to the zinc finger DBD on the C-terminal end of the protein to form oligomers. DNA also binds to the C-terminal end and can outcompete the acidic region for it. Image created using InkScape 1.0.

If YY1 oligomerisation results in DNA structuring and in turn changes in its DNA binding specificity, this implies that the concentration dependent nature of these oligomers is probably important for regulating these functions. Very low concentrations of YY1 in the cell might, therefore, lead to a difference in the functions of the molecule compared to when it is at much higher concentrations. This would help to explain one of the mechanisms by which YY1 carries out such diverse functions within different tissues and at different developmental stages. It may also help to explain one reason for its uncontrolled expression resulting in dementia.

This altered DNA binding behaviour is also potentially of interest in terms of proteinaceous membrane-less organelle formation. Within these structures, YY1's higher concentration is likely to result in increased lack of specificity for DNA binding, which might help to increase the number of multivalent interactions YY1 is capable of making and stabilise the organelles with which it is involved. Additionally, RNA has been observed to play a role in stabilising YY1 oligomerisation (Weintraub *et al.*, 2017). Given that proteinaceous membrane-less organelles consist of a mixture of proteins, RNA and sometimes DNA (Uversky, 2019), the changes in oligomerisation related to binding these molecules is likely to be significant in its role in nuclear speckles and polycomb bodies.

While these links need to be confirmed, and the purpose of YY1 oligomerisation further explored, there is clearly a relationship between YY1 concentration, oligomerisation and DNA binding which does suggest that this is an important aspect of YY1 functioning.

6.5. YY1 AND THE FOXP2 FHD INTERACTION

The self-association of YY1 is clearly an important aspect of its DNA binding behaviour, where changes in oligomerisation seem to affect DNA binding and also seem to be influenced by DNA. This is very likely to form part of the manner in which YY1 controls gene expression. Besides its ability to bind to DNA, heterologous interactions with other transcription factors is a vital part of YY1's activity (Gordon *et al.*, 2006; Verheul *et al.*, 2020). In some cases, the presence of other proteins has been found to enhance YY1's DNA binding activity (López-Perrote *et al.*, 2014). In other cases, YY1 is responsible for recruiting large complexes to DNA, such as the polycomb group which forms polycomb bodies (Alfieri *et al.*, 2013). Alternatively, some of its interactions with other transcription factors are likely to inhibit YY1's DNA binding activity, where a protein may mask the DBD of YY1 (Gordon *et al.*, 2006). Not all of YY1's interacting partners have been identified. Those which have been, have often not been studied in terms of their effect on YY1's function.

FOXP2 is one such transcription factor. Both FOXP2 and YY1 function in neural development, and both are associated with disorders such as ASD, schizophrenia and speech developmental problems (Fisher and Scharff, 2009; Bowers and Konopka, 2012; Gabriele *et al.*, 2017). They have also been found to interact *in vivo* using BRET assays (Estruch *et al.*, 2018). Whether this interaction is formed as part of a complex, however, or whether it can occur in isolation, is

unknown (Estruch *et al.*, 2018). Additionally, how an interaction between YY1 and FOXP2 might affect either protein's function has not been assessed.

Therefore, the FOXP2 FHD-YY1 interaction was investigated in this study both in terms of binding ability and in terms of the effect each protein had on the other's DNA binding capacity. This was to attempt to better understand how each protein works in terms of the heterologous interactions they form. In turn, this would help to expand understanding around how both of these proteins are involved in neurodevelopmental processes.

The FOXP2 FHD and YY1 do interact in isolation *in vitro*, as was seen using fluorescence anisotropy (***Figure 38***). This association is relatively low affinity, in the tens of micoromolar range. Different affinities of protein-protein interactions often serve specific purposes. Low affinity interactions typically have K_Ds above 10^{-6} M while high affinity interactions are generally above 10^{-10} M (Kastritis *et al.*, 2011). Intrinsically disordered proteins' heterologous interactions often are low affinity, due to the energetic changes required for the formation of structure within molecular recognition factors (Uversky and Dunker, 2010). This can function to help with tight regulational control, where low affinity coupled with high on and off rates allow for necessary interactions to take place and also to be stopped with relatively low effort (Uversky, 2019). Transcriptional regulation is complex and requires an available off switch to regulatory processes as much as it requires an efficient on switch (Uversky and Dunker, 2010). Furthermore, the formation of proteinaceous membrane-less organelles via liquid-liquid phase transitions requires the low affinity, multivalent heterologous interactions between intrinsically disordered proteins to retain a dynamic and reversible state (Banani *et al.*, 2017; Uversky, 2019).

FHD might interact with YY1 via its N-terminal or its zinc finger regions. However, given that YY1 binds to both the FHD and DNA simultaneously (***Figure 38*** and ***Figure 39***), it is more likely that this interaction is located within the intrinsically disordered N-terminal regions of YY1 (Górecki *et al.*, 2015). Nevertheless, this would need to be confirmed, as proteins with multiple zinc finger motifs have been known to be able to bind more than one macromolecule simultaneously (Iuchi, 2001) While it is not clear which part of this N-terminal region of YY1 may be involved in the YY1-FHD interaction, the lack of effect of ionic strength on binding suggests an importance of hydrophobic contacts within the interacting region or regions. The YY1 N-terminal region has several small patches of hydrophobic residues, but the area within

the protein with the overall longest stretches of these is the GA and GK rich regions, followed by the spacer region.

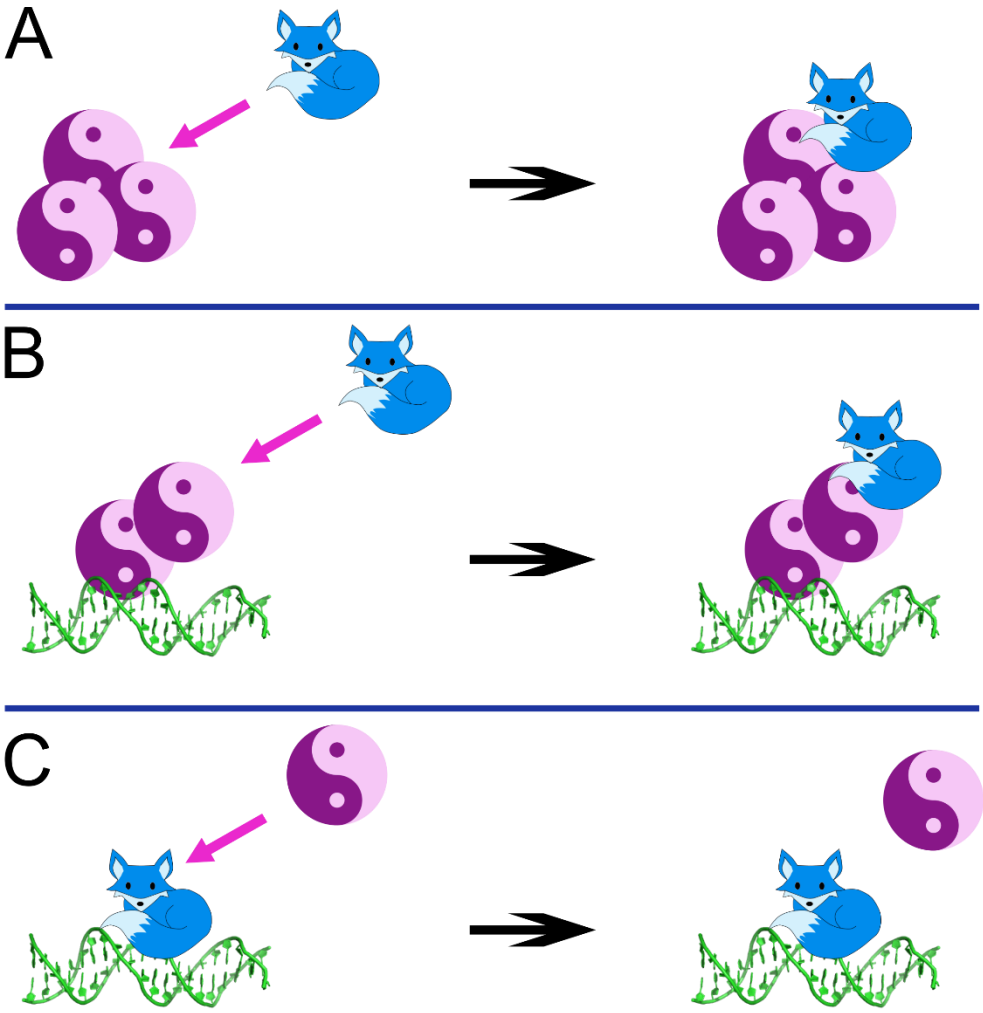


Figure 43: Model of how the FOXP2 FHD binds to YY1 in the presence and absence of DNA.
A) The FOXP2 FHD and YY1 form an interaction in isolation. **B)** YY1 binds to the FOXP2 FHD and to DNA simultaneously. **C)** The FOXP2 FHD cannot bind to its specific DNA and to YY1 at the same time. The FOXP2 FHD-DNA interaction has a lower K_D than the FOXP2 FHD-YY1 interaction (Morris and Fanucchi, 2016). This means that FOXP2 specific DNA likely outcompetes YY1 for FHD. Image created using InkScape 1.0.

Given that the YY1-FOXP2 FHD interaction is most likely located within the intrinsically disordered N-terminal region, the low affinity of the interaction does make sense as it is

characteristic of intrinsically disordered proteins (Uversky and Dunker, 2010). It is also possible that this might change for the full length FOXP2, or that the interaction between the two is partially mediated by other interacting partners – however DNA does not seem to have a significant effect on binding affinity.

In terms of the functional capacity of this interaction, YY1 can bind to FOXP2 and DNA simultaneously, as seen in the supershift assays (***Figure 43***) and fluorescence anisotropy (***Figure 38***). While the presence of FOXP2 did not seem to enhance YY1-DNA binding, as seen with members of the INO80 complex (López-Perrote *et al.*, 2014), it did cause a change in the electrophoretic mobility shift of one of the YY1-DNA bands (***Figure 39***). This indicates that FHD binds at the same time as the DNA and may also alter some of YY1's oligomeric interactions. In contrast, FOXP2 FHD could not interact with both YY1 and the DNA simultaneously (***Figure 40*** and ***Figure 43***). Therefore, if the ternary complex is physiologically relevant, YY1 is likely to act as a mediator between FOXP2 and DNA.

Additionally, this implies that YY1 may function to inhibit FOXP2's DNA binding function, thereby repressing it. Adequate repression of FOXP2 DNA binding would probably require a higher FOXP2-YY1 binding affinity than what is currently observed. In a case where only a slight alteration in FOXP2 DNA binding activity is required for subtle control of gene expression, a low affinity YY1-FHD interaction could be appropriate. Intrinsically disordered proteins are typically involved in recruitment, formation and localisation of large macromolecular complexes (Dunker *et al.*, 2005). As such it is possible that YY1 interacts with apo FOXP2 in order to recruit or localise it as part of a larger complex. In terms of localisation, it is possible that both proteins are present in a particular proteinaceous membrane-less organelle, and interact within one of these organelles as part of a shared functional pathway. In this case, the low affinity of their interaction would serve this localisation function well, and be easily reversible. Therefore, the FOXP2-YY1 interaction might be relatively non-specific, which is characteristic of the multivalent contacts made in liquid-liquid phase transitions (Uversky, 2019).

In contrast, if FOXP2 is used to regulate YY1 in any way, it may be for recruitment of YY1 to certain regions of DNA or to certain proteins. Additionally, FOXP2 may affect YY1's other protein-protein interactions, as opposed to its DNA regulatory ability. This may change according to whether FOXP2 only makes contacts with YY1 via its FHD or if it engages in contacts via its N-terminal domains. However, not much evidence other than the predicted

disorder of many of the full length FOXP2's regions implicates it in proteinaceous membrane-less organelles.

The interaction between YY1 oligomers and the FOXP2 specific section from the CNTNAP2 promoter region adds significance to the FOXP2-YY1 interaction. The CNTNAP2 promoter is not only regulated by FOXP2: YY1 is also known to bind to the CNTNAP2 promoter sequence, where mutations within this region of the promoter leads to the development of ASD (Chiocchetti *et al.*, 2015). Given that YY1 oligomers seem to be able to bind to the same regions as FOXP2 in the CNTNAP2 promoter with low affinity (***Figure 40***), this protein-protein interaction does seem to be potentially important in neural developmental pathways.

Regardless of the larger implications of this interaction, the fact that it does exist *in vitro* confirms the *in vivo* studies which were done (Estruch *et al.*, 2018), and together these indicate that this interaction is likely significant. Given the similarity between some of the functional pathways that both proteins regulate in neural development, it is very possible that the disruption of their interaction could be involved in developmental disorders, including ASD and schizophrenia.

CONCLUSION

In this study a number of novel and relevant discoveries were made and several factors of YY1's control over gene expression have been elucidated here. The manner in which YY1 binds to DNA has been clearly shown to be influenced by its oligomeric structure, which is in turn influenced by saturating concentrations of DNA. This suggests an important role that the concentration of YY1 plays relative to the concentration of DNA. The YY1:DNA ratio influences its DNA binding activity, which importantly includes specificity. At least some of the YY1 oligomers studied in this work seem to rely on the YY1 zinc finger domain and possibly the acidic rich regions of YY1 in order to form, where the latter becomes further intrinsically disordered upon DNA binding.

This mechanism of regulational control performed by YY1 is likely further complicated by the presence of other transcription factors which might interact with YY1. The FHD of FOXP2 does so *in vitro* and in isolation, confirming that these proteins likely interact directly. The binding capacity of both proteins to a FOXP2 binding site in the CNTNAP2 promoter region indicates that the YY1-FOXP2 interaction should be further investigated with relation to neural development. This is because a better understanding of how both FOXP2 and YY1 are implicated in ASD and speech disorders is vital for the development of suitable treatments (Poot *et al.*, 2010; Bowers and Konopka, 2012; Centanni *et al.*, 2015; Nabais Sá *et al.*, 2019). YY1 is capable of binding to both FOXP2 and DNA simultaneously, while the FOXP2 FHD binds either YY1 or DNA. This suggests that YY1 may play a part in FOXP2's DNA binding activity, or that both proteins may be involved in the formation of a much larger complex of transcription factors.

Overall, these findings help to explain how YY1 homo- and hetero-association work in conjunction with its DNA binding activity. This helps to understand how the protein influences its regulatory pathways. Transcription initiation, activation, repression, X-inactivation and chromatin remodelling are a wide array of functions for one protein (Usheva and Shenk, 1996; Jeon and Lee, 2011; Weintraub *et al.*, 2017). The dynamic changes allowed to YY1 by its intrinsic disorder, coupled with the versatility of its function when self-associating helps to explain the wide range of processes with which YY1 is involved, as well as how its function can differ so vastly depending on the context. As such, determining how its behaviour changes

under different oligomeric states and in the presence of its different binding partners is essential to understanding how the protein performs all of its functions. Furthermore, the heterogeneous behaviour of YY1 in solution as well as its disorder can be associated with its function in assembling proteinaceous membrane-less organelles such as ploycomb bodies and possibly nuclear speckles (Salichs *et al.*, 2009; Alfieri *et al.*, 2013; Uversky, 2019). The interactions YY1 forms with itself, DNA, and FOXP2, therefore, are part of a complex regulatory function which is essential for proper neural development. By better understanding the mechanisms discussed, it can be better understood how each transcription factor plays a role in conditions such as ASD, schizophrenia and dementia.

REFERENCES

Adam, I., Mendoza, E., Kobalz, U., Wohlgemuth, S. and Scharff, C. (2017). CNTNAP2 is a direct FoxP2 target in vitro and in vivo in zebra finches: complex regulation by age and activity. *Genes Brain Behav*, **16**(6): 635-642.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. (2002). DNA-binding motifs in gene regulatory proteins. *Mol Biol Cell*. New York, Garland Science.

Alexander, S. C. and Schepartz, A. (2014). Interactions of AsCy3 with cysteine-rich peptides. *Org Lett*, **16**(14): 3824-3827.

Alfieri, C., Gambetta, M. C., Matos, R., Glatt, S., Sehr, P., Fraterman, S., Wilm, M., Müller, J. and Müller, C. W. (2013). Structural basis for targeting the chromatin repressor Sfmbt to Polycomb response elements. *Genes Dev*, **27**(21): 2367-2379.

Allan, J., Mitchell, T., Harborne, N., Bohm, L. and Crane-Robinson, C. (1986). Roles of H1 domains in determining higher order chromatin structure and H1 location. *J Mol Biol*, **187**(4): 591-601.

Alrosan, M., Tan, T. C., Easa, A. M., Gammoh, S. and Alu'datt, M. H. (2021). Molecular forces governing protein-protein interaction: Structure-function relationship of complexes protein in the food industry. *Crit Rev Food Sci Nutr*: 1-17.

Ambadipudi, S., Biernat, J., Riedel, D., Mandelkow, E. and Zweckstetter, M. (2017). Liquid-liquid phase separation of the microtubule-binding repeats of the Alzheimer-related protein Tau. *Nat Commun*, **8**(1): 275.

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

ATTO-Tech. (ND). ATTO 550. 2021, from https://www.atto-tec.com/ATTO-550.html?language=en.

Aubry, S., Shin, W., Crary, J. F., Lefort, R., Qureshi, Y. H., Lefebvre, C., Califano, A. and Shelanski, M. L. (2015). Assembly and interrogation of Alzheimer's disease genetic networks reveal novel regulators of progression. *PLoS One*, **10**(3): e0120352.

Banani, S. F., Lee, H. O., Hyman, A. A. and Rosen, M. K. (2017). Biomolecular condensates: organizers of cellular biochemistry. *Nat Rev Mol Cell Biol*, **18**(5): 285-298.

Belak, Z. R. and Ovsenek, N. (2007). Assembly of the Yin Yang 1 transcription factor into messenger ribonucleoprotein particles requires direct RNA binding activity. *J Biol Chem*, **282**(52): 37913-37920.

Blane, A., Dirr, H. W. and Fanucchi, S. (2018). A phosphomimetic study implicates Ser557 in regulation of FOXP2 DNA binding. *Protein J*, **37**(4): 311-323.

Blane, A. and Fanucchi, S. (2015). Effect of pH on the structure and DNA binding of the FOXP2 Forkhead Domain. *Biochemistry*, **54**(25): 4001-4007.

Booth, W. T., Schlachter, C. R., Pote, S., Ussin, N., Mank, N. J., Klapper, V., Offermann, L. R., Tang, C., Hurlburt, B. K. and Chruszcz, M. (2018). Impact of an N-terminal polyhistidine tag on protein thermal stability. *ACS Omega*, **3**(1): 760-768.

Bowers, J. M. and Konopka, G. (2012). The role of the FOXP family of transcription factors in ASD. *Dis Markers*, **33**(5): 251-260.

Brockdorff, N. (2017). Polycomb complexes in X chromosome inactivation. *Philos Trans R Soc Lond B Biol Sci*, **372**(1733): 20170021.

Cao, H., Xiong, Y., Wang, T., Chen, B., Squier, T. C. and Mayer, M. U. (2007). A red Cy3-Based biarsenical fluorescent probe targeted to a complementary binding peptide. *J Am Chem Soc*, **129**(28): 8672-8673.

Cassandri, M., Smirnov, A., Novelli, F., Pitolli, C., Agostini, M., Malewicz, M., Melino, G. and Raschella, G. (2017). Zinc-finger proteins in health and disease. *Cell Death Discov*, **3**: 17071.

Centanni, T. M., Sanmann, J. N., Green, J. R., Iuzzini-Seigel, J., Bartlett, C., Sanger, W. G. and Hogan, T. P. (2015). The role of candidate-gene CNTNAP2 in childhood apraxia of speech and specific language impairment. *Am J Med Genet B Neuropsychiatr Genet*, **168**(7): 536-543.

Chan, D. S., Kavanagh, M. E., McLean, K. J., Munro, A. W., Matak-Vinković, D., Coyne, A. G. and Abell, C. (2017). Effect of DMSO on protein structure and interactions assessed by collision-induced dissociation and unfolding. *Anal Chem*, **89**(18): 9976-9983.

Chen, K., Lu, Y., Shi, K., Stovall, D. B., Li, D. and Sui, G. (2019). Functional analysis of YY1 zinc fingers through cysteine mutagenesis. *FEBS Lett*, **593**(12): 1392-1402.

Chiocchetti, A. G., Kopp, M., Waltes, R., Haslinger, D., Duketis, E., Jarczok, T. A., Poustka, F., Voran, A., Graab, U., Meyer, J., Klauck, S. M., Fulda, S. and Freitag, C. M. (2015). Variants of the CNTNAP2 5' promoter as risk factors for autism spectrum disorders: a genetic and functional approach. *Mol Psychiatry*, **20**(7): 839-849.

Chokas, A. L., Trivedi, C. M., Lu, M. M., Tucker, P. W., Li, S., Epstein, J. A. and Morrisey, E. E. (2010). Foxp1/2/4-NuRD interactions regulate gene expression and epithelial injury response in the lung via regulation of interleukin-6. *J. Biol. Chem.*, **285**(17): 13304-13313.

Choo, Y. and Klug, A. (1994). Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. *Proc Natl Acad Sci U S A*, **91**(23): 11168-11172.

Chu, Y. P., Chang, C. H., Shiu, J. H., Chang, Y. T., Chen, C. Y. and 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.

Colak, R., Kim, T., Michaut, M., Sun, M., Irimia, M., Bellay, J., Myers, C. L., Blencowe, B. J. and Kim, P. M. (2013). Distinct types of disorder in the human proteome: functional implications for alternative splicing. *PLoS Comput Biol*, **9**(4): e1003030.

Craggs, T. D., Hutton, R. D., Brenlla, A., White, M. F. and Penedo, J. C. (2014). Single-molecule characterization of Fen1 and Fen1/PCNA complexes acting on flap substrates. *Nucleic Acids Res*, **42**(3): 1857-1872.

Crick, S. L., Jayaraman, M., Frieden, C., Wetzel, R. and Pappu, R. V. (2006). Fluorescence correlation spectroscopy shows that monomeric polyglutamine molecules form collapsed structures in aqueous solutions. *Proc Natl Acad Sci U S A*, **103**(45): 16764-16769.

Cromer, M. K., Choi, M., Nelson-Williams, C., Fonseca, A. L., Kunstman, J. W., Korah, R. M., Overton, J. O., Mane, S., Kenney, B., Malchoff, C. D., Stalberg, P., Akerström, G., Westin, G., Hellman, P., Carling, T., Björklund, P. and Lifton, R. P. (2015). Neomorphic effects of recurrent somatic mutations in *Yin Yang 1* in insulin-producing adenomas. *PNAS*, **112**(13): 4062-4067.

Deng, Z., Cao, P., Wan, M. M. and Sui, G. (2010). Yin Yang 1: a multifaceted protein beyond a transcription factor. *Transcription*, **1**(2): 81-84.

Deriziotis, P., Graham, S. A., Estruch, S. B. and Fisher, S. E. (2014). Investigating protein-protein interactions in live cells using bioluminescence resonance energy transfer. *J Vis Exp*, (87).

Dion, M. F., Altschuler, S. J., Wu, L. F. and Rando, O. J. (2005). Genomic characterization reveals a simple histone H4 acetylation code. *Proc Natl Acad Sci U S A*, **102**(15): 5501-5506.

Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M. and Uversky, V. N. (2005). Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J*, **272**(20): 5129-5148.

Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. M., Hipps, K. W., Ausio, J., Nissen, M. S., Reeves, R., Kang, C., Kissinger, C. R., Bailey, R. W., Griswold, M. D., Chiu, W., Garner, E. C. and Obradovic, Z. (2001). Intrinsically disordered protein. *J Mol Graph Model*, **19**(1): 26-59.

Ellenberger, T. (1994). Getting a grip on DNA recognition: structures of the basic region leucine zipper, and the basic region helix-loop-helix DNA-binding domains. *Current Opinion in Structural Biology*, **4**(1): 12-21.

Englander, S. W. and Kallenbach, N. R. (1983). Hydrogen exchange and structural dynamics of proteins and nucleic acids. *Q Rev Biophys*, **16**(4): 521-655.
Estruch, S. B., Graham, S. A., Chinnappa, S. M., Deriziotis, P. and Fisher, S. E. (2016). Functional characterization of rare FOXP2 variants in neurodevelopmental disorder. *J Neurodev Disord*, **8**: 44.
Estruch, S. B., Graham, S. A., Quevedo, M., Vino, A., Dekkers, D. H. W., Deriziotis, P., Sollis, E., Demmers, J., Poot, R. A. and Fisher, S. E. (2018). Proteomic analysis of FOXP proteins reveals interactions between cortical transcription factors associated with neurodevelopmental disorders. *Hum Mol Genet*, **27**(7): 1212-1227.
Fisher, S. E. and Scharff, C. (2009). FOXP2 as a molecular window into speech and language. *Trends Genet*, **25**(4): 166-177.
Fried, M. G. (1989). Measurement of protein-DNA interaction parameters by electrophoresis mobility shift assay. *Electrophoresis*, **10**(5-6): 366-376.
Fuxreiter, M. (2012). Fuzziness: linking regulation to protein dynamics. *Mol Biosyst*, **8**(1): 168-177.
Gabriele, M., Vulto-van Silfhout, A. T., Germain, P. L., Vitriolo, A., Kumar, R., Douglas, E., Haan, E., Kosaki, K., Takenouchi, T., Rauch, A., Steindl, K., Frengen, E., Misceo, D., Pedurupillay, C. R. J., Stromme, P., Rosenfeld, J. A., Shao, Y., Craigen, W. J., Schaaf, C. P., Rodriguez-Buritica, D., Farach, L., Friedman, J., Thulin, P., McLean, S. D., Nugent, K. M., Morton, J., Nicholl, J., Andrieux, J., Stray-Pedersen, A., Chambon, P., Patrier, S., Lynch, S. A., Kjaergaard, S., Torring, P. M., Brasch-Andersen, C., Ronan, A., van Haeringen, A., Anderson, P. J., Powis, Z., Brunner, H. G., Pfundt, R., Schuurs-Hoeijmakers, J. H. M., van Bon, B. W. M., Lelieveld, S., Gilissen, C., Nillesen, W. M., Vissers, L., Gecz, J., Koolen, D. A., Testa, G. and de Vries, B. B. A. (2017). YY1 haploinsufficiency causes an intellectual disability syndrome featuring transcriptional and chromatin dysfunction. *Am J Hum Genet*, **100**(6): 907-925.
Galganski, L., Martyna, O., Urbanek, M. O. and Krzyzosiak, W. J. (2017). Nuclear speckles: molecular organization, biological function and role in disease. *Nucleic Acids Res*, **45**(18): 10350-10368.
Galvin, K. M. and Shi, Y. (1997). Multiple mechanisms of transcriptional repression by YY1. *Mol Cell Biol*, **17**(7): 3723-3732.
Gámez-Valero, A., Monguió-Tortajada, M., Carreras-Planella, L., Franquesa, M., Beyer, K. and Borràs, F. E. (2016). Size-Exclusion Chromatography-based isolation minimally alters Extracellular Vesicles' characteristics compared to precipitating agents. *Sci Rep*, **6**: 33641.
Gao, Y., Sun, L., Wu, Z., Xuan, C., Zhang, J., You, Y. and Chen, X. (2018). miR218 inhibits the proliferation of human glioma cells through downregulation of Yin Yang 1. *Mol Med Rep*, **17**(1): 1926-1932.
Garner, M. M. and Revzin, A. (1986). The use of gel electrophoresis to detect and study nucleic acid-protein interactions. *Trends in biochemical sciences*, **11**(10): 395-396.
Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D. and Bairoch, A. (2005). Protein identification and analysis tools on the ExPASy server. *The proteomics handbook*. Walker, J. M., Humana press**:** 571-607.
Gemayel, R., Chavali, S., Pougach, K., Legendre, M., Zhu, B., Boeynaems, S., van der Zande, E., Gevaert, K., Rousseau, F., Schymkowitz, J., Babu, M. M. and Verstrepen, K. J. (2015). Variable glutamine-rich repeats modulate transcription factor activity. *Mol Cell*, **59**(4): 615-627.
Ghisaidoobe, A. B. and Chung, S. J. (2014). Intrinsic tryptophan fluorescence in the detection and analysis of proteins: a focus on Forster resonance energy transfer techniques. *Int J Mol Sci*, **15**(12): 22518-22538.
Golebiowski, F. M., Górecki, A., Bonarek, P. and Dziedzicka-Wasylewska, M. (2011). Efficient overexpression and purification of active full-length human transcription factor Yin Yang 1 in Escherichia coli. *Protein Expr Purif*, **77**(2): 198-206.
Golebiowski, F. M., Górecki, A., Bonarek, P., Rapala-Kozik, M., Kozik, A. and Dziedzicka-Wasylewska, M. (2012). An investigation of the affinities, specificity and kinetics involved in the interaction between the Yin Yang 1 transcription factor and DNA. *FEBS J*, **279**(17): 3147-3158.

Gordon, S., Akopyan, G., Garban, H. and Bonavida, B. (2006). Transcription factor YY1: structure, function, and therapeutic implications in cancer biology. *Oncogene*, **25**(8): 1125-1142.
Górecki, A., Bonarek, P., Górka, A. K., Figiel, M., Wilamowski, M. and Dziedzicka-Wasylewska, M. (2015). Intrinsic disorder of human Yin Yang 1 protein. *Proteins*, **83**(7): 1284-1296.
Gualberto, A., LePage, D., Pons, G., Mader, S. L., Park, K., Atchison, M. L. and Walsh, K. (1992). Functional antagonism between YY1 and the serum response factor. *Mol Cell Biol*, **12**(9): 4209-4214.
Haberle, V. and Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. *Nat Rev Mol Cell Biol*, **19**(10): 621-637.
Hashimoto, H., Wang, D., Horton, J. R., Zhang, X., Corces, V. G. and Cheng, X. (2017). Structural Basis for the Versatile and Methylation-Dependent Binding of CTCF to DNA. *Mol Cell*, **66**(5): 711-720.
He, Y., Dupree, J., Wang, J., Sandoval, J., Li, J., Liu, H., Shi, Y., Nave, K. A. and Casaccia-Bonnefil, P. (2007). The transcription factor Yin Yang 1 is essential for oligodendrocyte progenitor differentiation. *Neuron*, **55**(2): 217-230.
Hecht, K. A., Xiong, Y., Barrack, D. A., Ford, N. R., Roesijadi, G. and Squier, T. C. (2018). Distance-matched tagging sequence optimizes live-cell protein labeling by a biarsenical fluorescent reagent AsCy3_E. *ACS Omega*, **3**(2): 2104-2110.
Holmes, K. C. (1983). Flexibility in tobacco mosaic virus. *Ciba Found Symp*, **93**: 116-138.
Houbaviy, H. B. and Burley, S. K. (2001). Thermodynamic analysis of the interaction between YY1 and the AAV P5 promoter initiator element. *Chemistry & Biology*, **8**: 179-187.
Houbaviy, H. B., Usheva, A., Shenk, T. and Burley, S. K. (1996). Cocrystal structure of YY1 bound to the adeno-associated virus P5 initiator. *Proc Natl Acad Sci U S A*, **93**: 13577-13582.
Hsu, W. L., Oldfield, C. J., Xue, B., Meng, J., Huang, F., Romero, P., Uversky, V. N. and Dunker, A. K. (2013). Exploring the binding diversity of intrinsically disordered proteins involved in one-to-many binding. *Protein Sci*, **22**(3): 258-273.
Iuchi, S. (2001). Three classes of C2H2 zinc finger proteins. *Cell Mol Life Sci*, **58**(4): 625-635.
Jarmoskaite, I., AlSadhan, I., Vaidyanathan, P. P. and Herschlag, D. (2020). How to measure and evaluate binding affinities. *Elife*, **9**.
Jensen, P. F. and Rand, K. D. (2016). Hydrogen exchange: a sensitive analytical window into protein conformation and dynamics. *Hydrogen exchange mass spectrometry of proteins: fundamentals, methods, and applications*. Weis, D. D. United Kingdom, John Wiley & Sons, Ltd**:** 1-15.
Jeon, Y. and Lee, J. T. (2011). YY1 tethers Xist RNA to the inactive X nucleation center. *Cell*, **146**(1): 119-133.
Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. and Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**(7873): 583-589.
Kastritis, P. L., Moal, I. H., Hwang, H., Weng, Z., Bates, P. A., Bonvin, A. M. J. J. and Janin, J. (2011). A structure-based benchmark for protein-protein binding affinity. *Protein Sci*, **20**(3): 482-491.
Katoh, M., Igarashi, M., Fukuda, H. and Nakagama, H. (2013). Cancer genetics and genomics of human FOX family genes. *Cancer Lett*, **328**(2): 198-206.
Kelly, S. M., Jess, T. J. and Price, N. C. (2005). How to study proteins by circular dichroism. *Biochim Biophys Acta*, **1751**(2): 119-139.
Kelly, S. M. and Price, N. C. (2000). The use of circular dichroism in the investigation of protein structure and function. *Curr Protein Pept Sci*, **1**(4): 349-384.
Kim, J. D., Faulk, C. and Kim, J. (2007). Retroposition and evolution of the DNA-binding motifs of YY1, YY2 and REX1. *Nucleic Acids Res*, **35**(10): 3442-3452.
Kim, J. D. and Kim, J. (2009). YY1's longer DNA-binding motifs. *Genomics*, **93**(2): 152-158.

Knezetic, J. A. and Luse, D. S. (1986). The presence of nucleosomes on a DNA template prevents initiation by RNA polymerase II in vitro. *Cell*, **45**(1): 95-104.
Konermann, L., Pan, J. and Liu, Y. H. (2011). Hydrogen exchange mass spectrometry for studying protein structure and dynamics. *Chem Soc Rev*, **40**(3): 1224-1234.
Konopka, G., Bomar, J. M., Winden, K., Coppola, G., Jonsson, Z. O., Gao, F., Peng, S., Preuss, T. M., Wohlschlegel, J. A. and Geschwind, D. H. (2009). Human-specific transcriptional regulation of CNS development genes by FOXP2. *Nature*, **462**(7270): 213-217.
Lace, M. J., Yamakawa, Y., Ushikai, M., Anson, J. R., Haugen, T. H. and Turek, L. P. (2009). Cellular factor YY1 downregulates the human papillomavirus 16 E6/E7 promoter, P97, in vivo and in vitro from a negative element overlapping the transcription-initiation site. *J Gen Virol*, **90**(Pt 10): 2402-2412.
Laity, J. H., Lee, B. M. and Wright, P. E. (2001). Zinc finger proteins: new insights into structural and functional diversity. *Curr Opin Struct Biol*, **11**(1): 39-46.
Lakowicz, J. R. (2006). *Principles of fluorescence spectroscopy*. New York, Springer.
Lakowicz, J. R., Maliwal, B. P., Cherek, H. and Balter, A. (1983). Rotational freedom of tryptophan residues in proteins and peptides. *Biochemistry*, **22**(8): 1741-1752.
Lalmansingh, A. S., Karmakar, S., Jin, Y. and Nagaich, A. K. (2012). Multiple modes of chromatin remodeling by Forkhead box proteins. *Biochim Biophys Acta*, **1819**(7): 707-715.
Lamond, A. I. and Spector, D. L. (2003). Nuclear speckles: a model for nuclear organelles. *Nat Rev Mol Cell Biol*, **4**(8): 605-612.
Li, B., Carey, M. and Workman, J. L. (2007). The role of chromatin during transcription. *Cell*, **128**(4): 707-719.
Li, S., Morley, M., Lu, M., Zhou, S., Stewart, K., French, C. A., Tucker, H. O., Fisher, S. E. and Morrisey, E. E. (2016). Foxp transcription factors suppress a non-pulmonary gene expression program to permit proper lung development. *Dev Biol*, **416**(2): 338-346.
Li, S., Weidenfeld, J. and Morrisey, E. E. (2004). Transcriptional and DNA binding activity of the Foxp1/2/4 family is modulated by heterotypic and homotypic protein interactions. *Mol Cell Biol*, **24**(2): 809-822.
Li, X., Wang, W., Wang, J., Malovannaya, A., Xi, Y., Li, W., Guerra, R., Hawke, D. H., Qin, J. and Chen, J. (2015). Proteomic analyses reveal distinct chromatin-associated and soluble transcription factor complexes. *Mol Syst Biol*, **11**(1): 775.
Liu, C. and Fang, Y. (2019). New insights of poly(ADP-ribosylation) in neurodegenerative diseases: A focus on protein phase separation and pathologic aggregation. *Biochem Pharmacol*, **167**: 58-63.
López-Perrote, A., Alatwi, H. E., Torreira, E., Ismail, A., Ayora, S., Downs, J. A. and Llorca, O. (2014). Structure of Yin Yang 1 oligomers that cooperate with RuvBL1-RuvBL2 ATPases. *J Biol Chem*, **289**(33): 22614-22629.
Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. and Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature*, **389**(6648): 251-260.
Luse, D. S. (2014). The RNA polymerase II preinitiation complex. Through what pathway is the complex assembled? *Transcription*, **5**(1): e27050.
Masson, G. R., Burke, J. E., Ahn, N. G., Anand, G. S., Borchers, C., Brier, S., Bou-Assaf, G. M., Engen, J. R., Englander, S. W., Faber, J., Garlish, R., Griffin, P. R., Gross, M. L., Guttman, M., Hamuro, Y., Heck, A. J. R., Houde, D., Iacob, R. E., Jorgensen, T. J. D., Kaltashov, I. A., Klinman, J. P., Konermann, L., Man, P., Mayne, L., Pascal, B. D., Reichmann, D., Skehel, M., Snijder, J., Strutzenberg, T. S., Underbakke, E. S., Wagner, C., Wales, T. E., Walters, B. T., Weis, D. D., Wilson, D. J., Wintrode, P. L., Zhang, Z., Zheng, J., Schriemer, D. C. and Rand, K. D. (2019). Recommendations for performing, interpreting and reporting hydrogen deuterium exchange mass spectrometry (HDX-MS) experiments. *Nat Methods*, **16**(7): 595-602.
Medina, E., Córdova, C., Villalobos, P., Reyes, J., Komives, E. A., Ramírez-Sarmiento, C. A. and Babul, J. (2016). Three-Dimensional Domain Swapping Changes the Folding Mechanism of the Forkhead Domain of FoxP1. *Biophys J*, **110**(11): 2349-2360.
Merten, J. A., Schultz, K. M. and Klug, C. S. (2012). Concentration-dependent oligomerization and oligomeric arrangement of LptA. *Protein Sci*, **21**(2): 211-218.

Milo, R., Jorgensen, P., Moran, U., Weber, G. and Springer, M. (2010). BioNumbers--the database of key numbers in molecular and cell biology. *Nucleic Acids Res*, **38**(Database issue): D750-753.
Mohan, A., Oldfield, C. J., Radivojac, P., Vacic, V., Cortese, M. S., Dunker, A. K. and Uversky, V. N. (2006). Analysis of molecular recognition features (MoRFs). *J Mol Biol*, **362**(5): 1043-1059.
Morris, G. and Fanucchi, S. (2016). A key evolutionary mutation enhances DNA binding of the FOXP2 forkhead domain. *Biochemistry*, **55**(13): 1959-1967.
Morris, G., Pahad, N., Dirr, H. W. and Fanucchi, S. (2018). A conserved cation binding site in the DNA binding domain of forkhead box transcription factors regulates DNA binding by FOXP2. *Arch Biochem Biophys*, **657**: 56-64.
Müller, C. and Herrmann, B. (1997). Crystallographic structure of the T domain–DNA complex of the Brachyury transcription factor. *Nature*, **389**: 884-888.
Nabais Sá, M. J., Gabriele, M., Testa, G. and de Vries, B. B. A. (2019). Gabriele-de Vries syndrome. Adam, M. P., Ardinger, H. H. and Pagon, R. A. Seattle, GeneReviews®.
Nelson, C. S., Fuller, C. K., Fordyce, P. M., Greninger, A. L., Li, H. and 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 Res*, **41**(12): 5991-6004.
Nieborak, A. and Górecki, A. (2016). Significance of the pathogenic mutation T372R in the Yin Yang 1 protein interaction with DNA--thermodynamic studies. *FEBS Lett*, **590**(6): 838-847.
Oldfield, C. J. and Dunker, A. K. (2014). Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu Rev Biochem*, **83**: 553-584.
Otto, M. R., Lillo, M. P. and Beechem, J. M. (1994). Resolution of multiphasic reactions by the combination of fluorescence total-intensity and anisotropy stopped-flow kinetic experiments. *Biophys J*, **67**(6): 2511-2521.
Pal, R., Tiwari, P. C., Nath, R. and Pant, K. K. (2016). Role of neuroinflammation and latent transcription factors in pathogenesis of Parkinson's disease. *Neurol Res*, **38**(12): 1111-1122.
Perez-Howard, G. M., Weil, P. A. and Beechem, J. M. (1995). Yeast TATA binding protein interaction with DNA: fluorescence determination of oligomeric state, equilibrium binding, on-rate, and dissociation kinetics. *Biochemistry*, **34**(25): 8005-8017.
Permyakov, S. E., Millett, I. S., Doniach, S., Permyakov, E. A. and Uversky, V. N. (2003). Natively unfolded C-terminal domain of caldesmon remains substantially unstructured after the effective binding to calmodulin. *Proteins*, **53**(4): 855-862.
Petty, K. J. (2001). Metal-chelate affinity chromatography. *Curr Protoc Protein Sci*, **Volume 4 Issue 1**: Unit 9 4.
Pignata, C. and Romano, R. (2014). In this issue: FOX genes and the immune response. *Int Rev Immunol*, **33**(2): 81-82.
Pirrotta, V. and Li, H. B. (2012). A view of nuclear Polycomb bodies. *Curr Opin Genet Dev*, **22**(2): 101-109.
Pollard, T. D., Blanchoin, L. and Mullins, R. D. (2000). Molecular mechanisms controlling actin filament dynamics in nonmuscle cells. *Annu Rev Biophys Biomol Struct*, **29**(545-76).
Poot, M., Beyer, V., Schwaab, I., Damatova, N., Van't Slot, R., Prothero, J., Holder, S. E. and Haaf, T. (2010). Disruption of CNTNAP2 and additional structural genome changes in a boy with speech delay and autism spectrum disorder. *Neurogenetics*, **11**(1): 81-89.
Rezai-Zadeh, N., Zhang, X., Namour, F., Fejer, G., Wen, Y. D., Yao, Y. L., Gyory, I., Wright, K. and Seto, E. (2003). Targeted recruitment of a histone H4-specific methyltransferase by the transcription factor YY1. *Genes Dev*, **17**(8): 1019-1029.
Rice, L. M., Brennwald, P. and Brünger, A. T. (1997). Formation of a yeast SNARE complex is accompanied by significant structural changes. *FEBS Lett*, **415**: 49-55.
Ruff, K. M. and Pappu, R. V. (2021). AlphaFold and implications for intrinsically disordered proteins. *J Mol Biol*, **433**(20): 167208.
Ruppert, J. M., Kinzler, K. W., Wong, A. J., Bigner, S. H., Kao, F. T., Law, M. L., Seuanez, H. N., O'Brien, S. J. and Vogelstein, B. (1988). The GLI-Kruppel family of human genes. *Mol Cell Biol*, **8**(8): 3104-3113.

Salichs, E., Ledda, A., Mularoni, L., Alba, M. M. and de la Luna, S. (2009). Genome-wide analysis of histidine repeats reveals their role in the localization of human proteins to the nuclear speckles compartment. *PLoS Genet*, **5**(3): e1000397.

Sarvagalla, S., Kolapalli, S. P. and Vallabhapurapu, S. (2019). The two sides of YY1 in cancer: A friend and a foe. *Front Oncol*, **9**: 1230.

Schägger, H. and von Jagow, G. (1987). Tricine-sodium dodecyl sulfate-polyacrylamide gel electrophoresis for the separation of proteins in the range from 1 to 100 kDa. *Anal Biochem*, **166**(2): 368-379.

Scharff, C. and Haesler, S. (2005). An evolutionary perspective on FoxP2: strictly for the birds? *Curr Opin Neurobiol*, **15**(6): 694-703.

Schuettengruber, B., Chourrout, D., Vervoort, M., Leblanc, B. and Cavalli, G. (2007). Genome regulation by polycomb and trithorax proteins. *Cell*, **128**(4): 735-745.

Shav-Tal, Y., Blechman, J., Darzacq, X., Montagna, C., Dye, B. T., Patton, J. G., Singer, R. H. and Zipori, D. (2005). Dynamic sorting of nuclear components into distinct nucleolar caps during transcriptional inhibition. *Mol Biol Cell*, **16**(5): 2395-2413.

Shi, Y., Lee, J. and Galvin, K. M. (1997). Everything you have ever wanted to know about Yin Yang 1. *Biochim Biophys Acta*, **1332**(2): F49-F66.

Shi, Y., Seto, E., Chang, L. S. and Shenk, T. (1991). Transcriptional repression by YY1, a human GLI-Krüppel-related protein, and relief of repression by adenovirus E1A protein. *Cell*, **67**(2): 377-388.

Shu, W., Yang, H., Zhang, L., Lu, M. M. and 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. *J Biol Chem*, **276**(29): 27488-27497.

Sigova, A. A., Abraham, B. J., Ji, X., Molinie, B., Hannett, N. M., Guo, Y. E., Jangi, M., Giallourakis, C. C., Sharp, P. A. and Young, R. A. (2015). Transcription factor trapping by RNA in gene regulatory elements. *Science*, **350**(6263): 978-981.

Sin, C., Li, H. and Crawford, D. A. (2014). Transcriptional regulation by FOXP1, FOXP2, and FOXP4 dimerization. *J Mol Neurosci*, **55**(2): 437-448.

Stroud, J. C., Wu, Y., Bates, D. L., Han, A., Nowick, K., Paabo, S., Tong, H. and Chen, L. (2006). Structure of the forkhead domain of FOXP2 bound to DNA. *Structure*, **14**(1): 159-166.

Struhl, K. (1989). Helix-turn-helix, zinc-finger, and leucine-zipper motifs for eukaryotic transcriptional regulatory proteins. *Trends in biochemical sciences*, **14**(4): 137-140.

Sui, G., Affar el, B., Shi, Y., Brignone, C., Wall, N. R., Yin, P., Donohoe, M., Luke, M. P., Calvo, D., Grossman, S. R. and Shi, Y. (2004). Yin Yang 1 is a negative regulator of p53. *Cell*, **117**(7): 859-872.

Sultan, N. A., Rao, R. N., Nadimpalli, S. K. and Swamy, M. J. (2006). Tryptophan environment, secondary structure and thermal unfolding of the galactose-specific seed lectin from Dolichos lablab: fluorescence and circular dichroism spectroscopic studies. *Biochim Biophys Acta*, **1760**(7): 1001-1008.

Thomas, M. J. and Seto, E. (1999). Unlocking the mechanisms of transcription factor YY1: are chromatin modifying enzymes the key? *Gene*, **236**: 197-208.

Thulo, M., Rabie, M. A., Pahad, N., Donald, H. L., Blane, A. A., Perumal, C. M., Penedo, J. C. and Fanucchi, S. (2021). The influence of various regions of the FOXP2 sequence on its structure and DNA-binding function. *Biosci Rep*, **41**(1).

Usheva, A. and Shenk, T. (1996). YY1 transcriptional initiator: protein interactions and association with a DNA site containing unpaired strands. *Proc Natl Acad Sci U S A*, **26;93**(24): 13571-13576.

Usui, N., Co, M., Harper, M., Rieger, M. A., Dougherty, J. D. and Konopka, G. (2017). Sumoylation of FOXP2 regulates motor function and vocal communication through Purkinje cell development. *Biol Psychiatry*, **81**(3): 220-230.

Uversky, V. N. (2009). Intrinsically disordered proteins and their environment: effects of strong denaturants, temperature, pH, counter ions, membranes, binding partners, osmolytes, and macromolecular crowding. *Protein J*, **28**(7-8): 305-325.

Uversky, V. N. (2019). Intrinsically disordered proteins and their "mysterious" (meta)physics. *Frontiers in Physics*, **7**.

Uversky, V. N. (2019). Supramolecular fuzziness of intracellular liquid droplets: Liquid-liquid phase transitions, membrane-less organelles, and intrinsic disorder. *Molecules*, **24**(18).
Uversky, V. N. and Dunker, A. K. (2010). Understanding protein non-folding. *Biochim Biophys Acta*, **1804**(6): 1231-1264.
Uversky, V. N., Gillespie, J. R. and Fink, A. L. (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins*, **41**(3): 415-427.
Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Zidek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., Figurnov, M., Cowie, A., Hobbs, N., Kohli, P., Kleywegt, G., Birney, E., Hassabis, D. and Velankar, S. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res*, **50**(D1): D439-D444.
Varadi, M., Zsolyomi, F., Guharoy, M. and Tompa, P. (2015). Functional advantages of conserved intrinsic disorder in RNA-binding proteins. *PLoS One*, **10**(10): e0139731.
Verheul, T. C. J., van Hijfte, L., Perenthaler, E. and Barakat, T. S. (2020). The why of YY1: Mechanisms of transcriptional regulation by Yin Yang 1. *Front Cell Dev Biol*, **8**: 592164.
Vitalis, A., Wang, X. and Pappu, R. V. (2007). Quantitative characterization of intrinsic disorder in polyglutamine: insights from analysis based on polymer theories. *Biophys J*, **93**(6): 1923-1937.
Wai, D. C. C., Shihab, M., Low, J. K. K. and Mackay, J. P. (2016). The zinc fingers of YY1 bind single-stranded RNA with low sequence specificity. *Nucleic Acids Res*, **44**(19): 9153-9165.
Wang, B., Lin, D., Li, C. and Tucker, P. (2003). Multiple domains define the expression and regulatory properties of Foxp1 forkhead transcriptional repressors. *J Biol Chem*, **278**(27): 24259-24268.
Wang, H., Garzon, R., Sun, H., Ladner, K. J., Singh, R., Dahlman, J., Cheng, A., Hall, B. M., Qualman, S. J., Chandler, D. S., Croce, C. M. and Guttridge, D. C. (2008). NF-kappaB-YY1-miR-29 regulatory circuitry in skeletal myogenesis and rhabdomyosarcoma. *Cancer Cell*, **14**(5): 369-381.
Wang, J., Zhou, L., Li, Z., Zhang, T., Liu, W., Liu, Z., Yuan, Y.-C., Su, F., Xu, L., Wang, Y., Zhou, X., Xu, H., Hua, Y., Wang, Y.-J., Zheng, L., Teng, Y.-E. and Shen, B. (2015). YY1 suppresses FEN1 over-expression and drug resistance in breast cancer. *BMC Cancer*, **15**(1).
Webb, H., Steeb, O., Blane, A., Rotherham, L., Aron, S., Machanick, P., Dirr, H. and 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.
Weintraub, A. S., Li, C. H., Zamudio, A. V., Sigova, A. A., Hannett, N. M., Day, D. S., Abraham, B. J., Cohen, M. A., Nabet, B., Buckley, D. L., Guo, Y. E., Hnisz, D., Jaenisch, R., Bradner, J. E., Gray, N. S. and Young, R. A. (2017). YY1 Is a structural regulator of enhancer-promoter loops. *Cell*, **171**(7): 1573-1588 e1528.
Weis, D. D. (2016). Analysis of disordered proteins by hydrogen exchange mass spectrometry. *Hydrogen exchange mass spectrometry of proteins: fundamentals, methods, and applications*. Weis, D. D. United Kingdom, John Wiley & Sons, Ltd**:** 295-322.
Whitmore, L. and Wallace, B. A. (2004). DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data. *Nucleic Acids Res*, **32**(Web Server issue): W668-673.
Wilkinson, F., Pratt, H. and Atchison, M. L. (2010). PcG recruitment by the YY1 REPO domain can be mediated by Yaf2. *J Cell Biochem*, **109**(3): 478-486.
Wilkinson, F. H., Park, K. and Atchison, M. L. (2006). Polycomb recruitment to DNA in vivo by the YY1 REPO domain. *Proc Natl Acad Sci U S A*, **103**(51): 19296-19301.
Wu, S., Shi, Y., Mulligan, P., Gay, F., Landry, J., Liu, H., Lu, J., Qi, H. H., Wang, W., Nickoloff, J. A., Wu, C. and Shi, Y. (2007). A YY1-INO80 complex regulates genomic stability through homologous recombination-based repair. *Nat Struct Mol Biol*, **14**(12): 1165-1172.
Wu, Y., Borde, M., Heissmeyer, V., Feuerer, M., Lapan, A. D., Stroud, J. C., Bates, D. L., Guo, L., Han, A., Ziegler, S. F., Mathis, D., Benoist, C., Chen, L. and Rao, A. (2006). FOXP3 controls regulatory T cell function through cooperation with NFAT. *Cell*, **126**(2): 375-387.

Yang, L. Q., Sang, P., Tao, Y., Fu, Y. X., Zhang, K. Q., Xie, Y. H. and Liu, S. Q. (2014). Protein dynamics and motions in relation to their functions: several case studies and the underlying mechanisms. *J Biomol Struct Dyn*, **32**(3): 372-393.

Yang, W.-M., Yao, Y.-L., Sun, J.-M., Davie, J. R. and Seto, E. (1997). Isolation and Characterization of cDNAs Corresponding to an Additional Member of the Human Histone Deacetylase Gene Family. *Journal of Biological Chemistry*, **272**(44): 28001-28007.

Yang, W. M., Inouye, C., Bearss, D. and Seto, E. (1996). Transcriptional repression by YY1 is mediated by interaction with a mammalian homolog of the yeast global regulator RPD3. *Proc. Natl. Acad. Sci. USA*, **93**: 12845-12850.

Yang, W. M., Yao, Y. L., Sun, J. M., Davie, J. R. and Seto, E. (1997). Isolation and characterization of cDNAs corresponding to an additional member of the human histone deacetylase gene family. *J Biol Chem*, **272**(44): 28001-28007.

Yue, R., Kang, J., Zhao, C., Hu, W., Tang, Y., Liu, X. and Pei, G. (2009). Beta-arrestin1 regulates zebrafish hematopoiesis through binding to YY1 and relieving polycomb group repression. *Cell*, **139**(3): 535-546.

Zhang, J. J., Zhu, Y., Zhang, X. F., Liu, D. F., Wang, Y., Yang, C., Shi, G. D., Peng, Y. P., Zhang, K., Tian, L., Miao, Y. and Jiang, K. R. (2017). Yin Yang-1 suppresses pancreatic ductal adenocarcinoma cell proliferation and tumor growth by regulating SOX2OT-SOX2 axis. *Cancer Lett*, **408**: 144-154.

Zhang, Z. and Smith, D. L. (1993). Determination of amide hydrogen exchange by mass spectrometry: a new tool for protein structure elucidation. *Protein Sci*, **2**(4): 522-531.

Zhao, C., Hellman, L. M., Zhan, X., Bowman, W. S., Whiteheart, S. W. and Fried, M. G. (2010). Hexahistidine-tag-specific optical probes for analyses of proteins and their interactions. *Anal Biochem*, **399**(2): 237-245.