

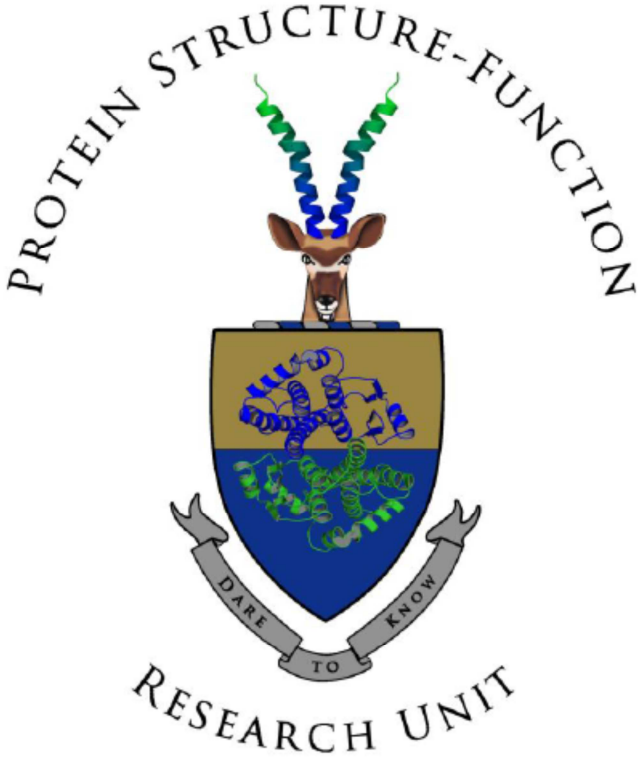
https://dspace.org/

Binding studies of the FOXP2 forkhead domain and its cognate DNA sequences

Webb, Helen Susannah

2016
http://hdl.handle.net/10539/19296

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



Binding studies of the FOXP2 forkhead domain and its cognate DNA sequences

Helen Susannah Webb 9810273f

A thesis submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Doctor of Philosophy (PhD)

Johannesburg, 2015

Declaration

I declare that this dissertation 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 towards any degree or examination at any other university.

Helen Susannah Webb



25 May, 2015

Research Outputs
2011 - 20114

Conference presentations Name	**Date**	**Location**	**Title**	**Authors**
SASBMB 2012	January, 2012	Drakensberg, South Africa	Binding of the FOXP2 forkhead domain to DNA is electrostatically driven	H.S. Letseka S. Fanucchi
EMBO Modern biophysical methods in protein-ligand interactions	October, 2013	Oulu, Finland	The FOXP2 forkhead domain binds various distinct DNA sequences with varied rates and affinities	H.S. Letseka S. Fanucchi
Molecular Biosciences Research Thrust	December 2013	University of the Witwatersrand, South Africa	Identification of novel binding motifs of the FOXP2 forkhead domain	H.S .Letseka S. Fanucchi
6th Cross-Faculty Postgraduate Symposium	28 Oct, 2014	University of the Witwatersrand, South Africa	Binding studies of the FOXP2 forkhead domain and various DNA sequences	H.S. Letseka S. Fanucchi

This work is dedicated to my son, Elliot

And Bradley Peter

You are my sunshine, my only sunshine

You make me happy when skies are grey

Acknowledgements

I would like to thank my family (the Webbs and the Letsekas) for emotional and financial support during my studies. Also the NRF and Wits Post Graduate Merit Award and Wits staff Bursary for funding. And to Prof. Heinrich Dirr for allowing me the opportunity to work in his lab.

I wish to express an enormous amount of gratitude towards my supervisor, Dr Sylvia Fanucchi who has allowed me the freedom to peruse this project and given me the guidance to finish it. I have grown a great deal as a scientist, academic and human being under her mentorship.

I wish to acknowledge the following people's invaluable contribution to this work: the Uniteers of the PSFRU, especially the FOXy crowd. Philip Machanick of Rhodes University's Computer Science Department and Shaun Aron of the Wits Bioinformatics Department for assistance with motif analysis. Lia Rotherham of the Council for Scientific and Industrial Research for assistance with systematic evolution of ligands by exponential enrichment and surface plasmon resonance. Dario Fanucchi for assistance with data fitting. Stoyan Stoychev of the Council for Scientific and Industrial Research for mass spectrometry assistance and Prof. Yasien Sayed for assistance with isothermal titration calorimetry.

Abstract

FOXP2 is the gene product of the so-called “language gene” and is the only protein known to be involved in a monogenetic autosomally inherited language disorder. This disorder has been termed Speech-Language Disorder 1. In addition to the role it plays in language, FOXP2 is thought to be involved in cancer, autism and schizophrenia. FOXP2 is a member of the P subfamily of FOX transcription factors, the DNA-binding domain of which is the forkhead domain. The aim of this work was to investigate the binding mechanism of the FOXP2 forkhead domain and various DNA sequences in order to assess affinity and specificity. It was shown by surface plasmon resonance that the FOXP2 forkhead domain can recognise a variety of DNA sequences, including a novel sequence, identified by systematic evolution of ligands by exponential enrichment. This motif has not previously been reported as a binding motif of the FOXP2 forkhead domain. Kinetic analysis by surface plasmon resonance showed that the novel sequence, as well as other published cognate sequences, each binds to the FOXP2 forkhead domain with different rates and affinities. Molecular docking of the DNA sequences to the FOXP2 forkhead domain revealed that electrostatic interactions between positively charged amino acids and the DNA backbone, as well as base-specific interactions between His554 and the DNA appear to be key in determining rates and affinities of binding interactions of the FOXP2 forkhead domain and DNA. Based on these findings, three types of DNA-binding are proposed for the FOXP2 forkhead domain. These types are: low affinity, non-functional binding; moderate affinity, non-functional binding and high affinity, functional binding. It is probable that each type of binding serves to control the

spatial location of the protein within the nucleus, as well as the local concentration of protein. The proposed mechanism of binding for the forkhead domain of FOXP2 may have a future impact on the binding and function of full length FOXP2.

Table of Contents

1 Background and problem identification 1
1.1 Transcription factors 1
1.2 Protein-DNA recognition 2
1.3 Structures of DNA binding proteins 3
1.3.1 Helix-turn-helix motifs 5
1.3.2 Winged helix motif 7
1.4 DNA structure 8
1.5 Mechanism of protein-DNA recognition 12
1.5.1 Specificity of protein-DNA recognition 13
1.5.2 Base readout 13
1.5.3 Shape readout 18
1.5.5 Cooperativity of protein-DNA recognition 19
1.5.6 Facilitated diffusion 20
1.5.7 Low affinity DNA binding by transcription factors 21
1.6 Forkhead proteins 22
1.6.1 Interaction of FOX proteins with DNA 23
1.6.2 The forkhead box P family 24
1.6.3 The function of FOXP2 27
1.6.4 The structure of FOXP2 32
1.6.5 Domain swapping in the FOXP2 forkhead domain 33
1.6.6 The interaction of the FOXP2 forkhead domain with DNA 36
1.7 Problem identification 40
2. Aims 41
2.1 Overall aim: 41
2.2 Specific objectives: 41
3 Experimental Procedures 42
3.1 Creation of monomeric mutant of the FOXP2 forkhead domain 42
3.1.1 The pET-30 expression system 42
3.1.2 Plasmid extraction 42
3.1.4 Site-directed mutagenesis 42
3.1.5 Transformation of competent cells with plasmid DNA 43

3.2 Overexpression and purification of the FOXP2 forkhead domain and the A539P mutant 44
3.2.1 Expression trials 44
3.2.2 Immobilised metal ion chromatography 45
3.2.3 Determination of the FOXP2 forkhead domain and A539P mutant purity and concentration 45
3.2.3.1 Sodium dodecyl sulphate polyacrylamide gel electrophoresis 45
3.3 Confirmation of the identity of the wild type FOXP2 forkhead domain and the A539P mutant 46
3.4 Confirmation of the structural integrity of the FOXP2 forkhead domain A539P mutant 47
3.5 Identification of novel FOXP2 cognate DNA sequences 47
3.5.1 Systematic evolution of ligands by exponential enrichment 47
3.5.2 Motif identification 49
3.5.3 DNA preparation 50
3.6 Determination of the kinetic and thermodynamic parameters and binding affinity of the FOXP2 forkhead domain binding various DNA sequences 52
3.6.1 Surface plasmon resonance 52
3.6.2 Isothermal titration calorimetry 55
3.7 *In silico* prediction of the residues and bases involved in the interaction between the FOXP2 forkhead domain and various DNA sequences 55
3.7.1 Modelling of DNA sequences 55
3.7.2 Molecular docking 56
3.8 Structural alignment 56
4. Results 57
4.1 Construction of the FOXP2 forkhead domain A539P mutant 57
4.2 Overexpression and purification of the FOXP2 forkhead domain and the A539P mutant 59
4.3 Confirmation of the identity of the wild type FOXP2 FHD and the A539P mutant by mass spectroscopy 62
4.4 Confirmation of the structural integrity of the FOXP2 forkhead domain 64
4.5 Determination of the oligomeric state of the FOXP2 forkhead domain 67
4.6 Identification of novel FOXP2 forkhead domain binding motifs 69
4.7 Rates and affinities of the FOXP2 forkhead domain and various DNA sequences 76
4.8 Isothermal titration calorimetry 82

4.9 *In silico* prediction of the bonds formed between the FOXP2 forkhead domain and various DNA sequences 86

4.10 Structural alignment of the predicted structures of the FOXP2 forkhead domain and various DNA sequences 96

5 Discussion..... 98

5.1 The FOXP2 FHD can bind a variety of distinct sequences..... 98

5.2 The FOXP2 FHD binds distinct sequences with varied affinity and rates 99

5.3 The FOXP2 FHD is monomeric in solution 103

5.4 Proposed types of binding between the FOXP2 forkhead domain and DNA 105

5.5 Possible relevance of proposed mechanism to full length FOXP2 110

6. Conclusion..... 111

7. References: 113

List of figures

Figure 1.1 Schematic representation of DNA binding motifs..... 4
Figure 1.2 Structural variations of the helix-turn-helix motif..... 6
Figure 1.3: Illustration of global and local DNA conformation changes..... 9
Figure 1.4: Hydrogen bond donors and acceptors of DNA base pairs..... 15
Figure 1.5: The interactions formed between specific amino acids and bases. 17
Figure 1.6: Canonical topology of a helix-turn-helix motif, a winged helix motif and a FOXP forkhead domain..... 26
Figure 1.7: Domain architecture of FOXP2..... 33
Figure 1.8: Schematic representation of the FOXP2 domain swapped dimer..... 34
Figure 1.9: Ribbon representation of the FOXP2 forkhead domain bound to DNA (PDB 2A07). 37
Figure 1.10: DNA contacts made by the FOXP2 forkhead domain 38
Figure 3.1: DNA sequences used to determine the kinetic parameters of the binding of the FOXP2 FHD and DNA..... 51
Figure 4.1 Sequence of the FOXP2 forkhead domain A539P mutant insert in pET-30..... 58
Figure 4.2 Expression trials of the wild type FOXP2 forkhead domain..... 60
Figure 4.3 Purification of the FOXP2 forkhead domain and the FOXP2 forkhead domain A539P mutant 61
Figure 4.4 Mass spectra of the FOXP2 forkhead domain and the FOXP2 Forkhead domain A539P mutant. 63
Figure 4.5 Intrinsic Fluorescence of the FOXP2 forkhead domain and the FOXP2 forkhead domain A539P mutant 66
Figure 4.6 Confirmation of the oligomeric state of the wild type FOXP2 forkhead domain and the A539P mutant..... 68
Figure 4.7 Percentage enrichment of systematic evolution of ligands by exponential enrichment rounds 74
Figure 4.8 Screening of identified DNA motifs for FOXP2 forkhead domain binding 75

Figure 4.9 The kinetics of binding of the FOXP2 forkhead domain A539P mutant and various DNA sequences 80
Figure 4.10 Rates and affinities of the FOXP2 forkhead domain and various DNA sequences..... 81
Figure 4.11.: Binding isotherm of the FOXP2 FHD A539P with the Nelson DNA sequence. 85
Figure 4.12 In silico prediction of the interaction between the FOXP2 forkhead domain and the Nelson DNA sequence. 88
Figure 4.13 In silico prediction of the interaction between the FOXP2 forkhead domain and the Wang DNA sequence. 89
Figure 4.14 In silico prediction of the interaction between the FOXP2 forkhead domain and the Webb sequence. 90
Figure 4.15 In silico prediction of the interaction between the FOXP2 forkhead domain and the Zhu sequence. 91
Figure 4.16 The relationship between rates and affinities of various 95
Figure 4.17 Structural alignment of the backbone of the FOXP2 forkhead domain bound to various DNA sequences. 97
Figure 5.2 Proposed mechanism of the FOXP2 forkhead domain DNA-binding. 106

List of tables

Table 4.1 DNA motifs identified by systematic evolution of ligands by exponential enrichment……..72

Table 4.2: DNA sequences used to determine the kinetic parameters of the binding of the FOXP2 forkhead domain and DNA……..78

Table 4.3 Affinities and rates of the FOXP2 forkhead domain and various DNA sequences……82

Table 4.4 Hydrogen Bonds formed between the FOXP2 forkhead domain and various DNA sequences…….93

Frequently used abbreviations

AME	Analysis of motif enrichment
Bp	Base pair
CSIR	Council for Scientific and Industrial Research
DNA	Deoxyribonucleic acid
dNTPs	Deoxynucleotide triphosphates
DREME	Discriminative regular expression motif elicitation
DTT	Dithiothreitol
EDC	1-Ethyl-3-[3-dimethylaminopropyl]carbodiimide hydrochloride
EDTA	Ethylenediaminetetraacetic acid
FHD	Forkhead domain
FOX	Forkhead box protein
HADDOCK	High ambiguity driven protein-protein docking
HEPES	S 4-(2-Hydroxyethyl)-1-piperazineethanesulfonic acid
His-tag	Histidine-tagged
HTH	Helix-turn helix
IDs	Intrinsically disordered regions
IMAC	Immobilised metal affinity chromatography
IPTG	Isopropyl-β-D-thiogalactoside
ITC	Isothermal titration calorimetry
LB	Luria-Bertani Broth
NHS	Sulfo-N-hydroxysuccinimide
NGS	Next generation sequencing
PAGE	Polyacrylamide gel electrophoresis
PCR	Polymerase chain reaction
RNA	Ribonucleic acid
Rpm	Revolutions per minute
SDS	Sodium dodecyl sulphate
SEC	Size exclusion chromatography
SELEX	Systematic evolution of ligands by exponential enrichment
SPCH1	Speech-language disorder 1
SPR	Surface plasmon resonance
TB	Tuberculosis
TF	Transcription factor
Tris	Tris(hydroxymethyl)-aminomethane
UV	Ultra violet

1 Background and problem identification

1.1 Transcription factors

Transcription factors are crucial in gene regulation. It has been estimated that as much as 10% of genes within the human genome encode transcription factors (Lander *et al*., 2001; Vaquerizas *et al*., 2009; Wilson *et al*., 2008). Transcription factors can be divided into two broad classes: basal transcription factors which interact with RNA polymerase and are required for baseline transcription and specific transcription factors which regulate transcription in response to specific biological signals. Transcription factors control gene regulation by either recruiting transcription machinery or by preventing the binding of machinery to promoter sequence elements such as the TATA box. Thus, transcription factors can either up-regulate (activate) or down-regulate (repress) target genes (Triezenberg, 1995).

The study of transcription factors is important as many human pathologies are caused by their dysregulation (Lee and Young, 2013). These diseases and syndromes include cancers (Castillo Bosch *et al*., 2014); (Littlewood *et al*., 2012), neurological (Rump et al., 2011) and autoimmune (Kyewski and Klein, 2006); (Hayden and Ghosh, 2012) disorders, cardiovascular disease (Kathiresan and Srivastava, 2012) and metabolic defects such as diabetes (Maurano *et al*., 2012) and obesity (Miyata *et al*., 2013).

Another reason for the necessity of studies into transcription factor binding is its evolutionary importance. The rapid evolution of transcription factors and the

resulting changes has played a major role in the recent molecular evolution of *Homo sapiens* (Bustamante *et al*., 2005). About 13% of transcription factors are unique to humans and do not occur in other primates compared to only 2% of enzymes (Vaquerizas *et al*., 2009).

The DNA sequences to which transcription factors bind in a sequence specific manner are known as consensus sequences. The length of consensus sequences within an organism is inversely proportional to the size of its genome such that small genomes are able to accommodate long binding sites while larger genomes require clustered short sites (Mirny *et al*., 2009). This is because regulation in larger genomes is more complex, requiring the binding of numerous factors to elements within a promoter.

1.2 Protein-DNA recognition

Protein-DNA interactions are vital to many biological processes including DNA replication, transcription and recombination. It is thus of great importance to study the structure and mechanism of DNA-protein recognition. DNA-protein interactions can be categorised by two factors: whether the interaction is specific to the DNA sequence at which it occurs and whether a functional event takes place as a consequence of the interaction. In the case of transcription factors, a functional event can be defined as a change in the transcription level of the gene to which the factor is bound. Not all DNA-protein binding events lead to a functional event in the cell (Gao *et al*., 2004).

Three types of binding interactions have been identified: functional specific binding, non-functional specific binding and nonspecific non-functional binding

(Elf *et al*., 2007; Kao-Huang and Revzin, 1977; Lin and Riggs, 1975; Phair *et al*., 2004). Non-functional binding (both specific and nonspecific) may play an indirect role in gene regulation by titrating the level of freely accessible transcription factor available to bind functional sites (Todeschini *et al*., 2014). Control of the concentration of transcription factors involved in functional DNA interactions is important because cellular concentrations of transcription factors are high relative to those of other proteins. Non-functional binding occurs because transcription factors are almost never free in solution, but rather associate with DNA and chromatin. (Elf *et al*., 2007; Kao-Huang and Revzin, 1977; Lin and Riggs, 1975; Phair *et al*., 2004). This constant non-specific interaction does not give rise to gene regulation.

Three factors are important in DNA-protein interactions: the structure of the DNA-binding proteins, the local and global structure of DNA within the cell and the mechanisms by which proteins recognise DNA sequences. These topics will be discussed in detail in the following sections.

1.3 Structures of DNA binding proteins

The most abundant structural class of DNA binding motifs are those containing α helices which interact with the major groove of the DNA. This includes helix turn helix motifs, leucine zippers and zinc fingers (Xiong and Sundaralingam, 2001). These proteins all share similar mechanisms of binding but differ in how the DNA-interacting helices are supported by the rest of the protein (Figure 1.1).

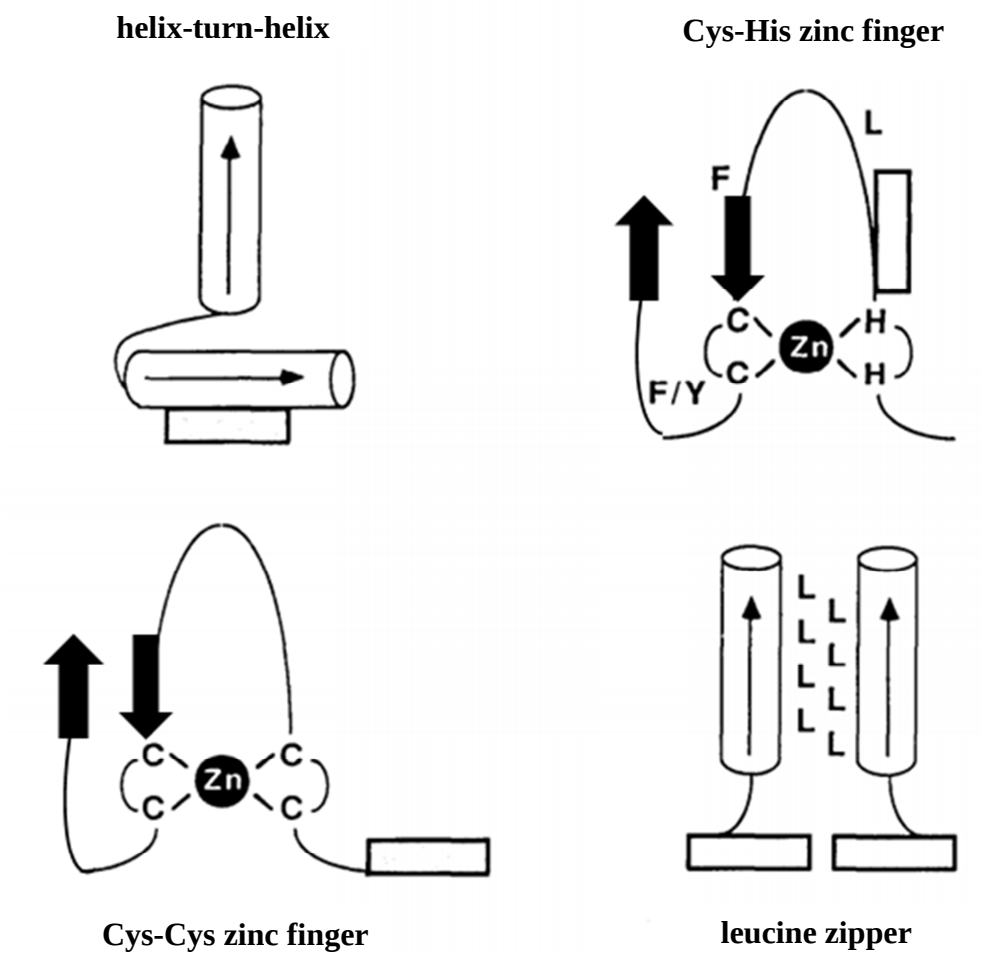


Figure 1.1 Schematic representation of DNA binding motifs. Schematics of the three most common DNA binding motifs are indicated: the helix-turn-helix, Cys-His and Cys-Cys zinc fingers and the leucine zipper. Bound DNA is indicated by bold rectangles, helices are represented as cylinders and important amino acids are represented by the single letter code Adapted from (Struhl, 1989).

Leucine zippers are dimeric motifs made up of α helices of about 60 amino acids which form a parallel coiled coil arrangement and dimerise via leucines at the hydrophobic dimer interface. This interface is typically made up of leucines spaced at 7 residue intervals (O'Shea *et al*., 1991). This interface occurs C-terminally while the N-terminals of the dimer splay out and insert into the major groove on opposite sides of the same DNA molecule.

Proteins that contain the zinc finger motif are the most abundant DNA-binding proteins in the human genome (Lander *et al*., 2001). The zinc finger domain is made up of a short α helix, two antiparallel strands of β sheets and a core zinc ion which is coordinated by two cysteines and two histidines or two pairs of cysteines (Pavletich and Pabo, 1991). DNA recognition occurs by insertion of the helix into the major groove (Pavletich and Pabo, 1991, 1993). Zinc finger-containing proteins often have multiple copies of this domain interspaced by short linker regions (Wolfe *et al*., 2001).

1.3.1 Helix-turn-helix motifs

The general structure of a helix-turn-helix (HTH) motif is a three helix bundle of two α helices separated by a three or four residue turn and a third "recognition" helix which embeds into the major groove of the DNA upon binding. The helices not in contact with the DNA are typically offset by a 120° angle (Brennan and Matthews, 1989). There are a number of structural variants of the HTH motifs (Figure 1.2; Aravind and Anantharaman, 2005). These variations are the inclusion of a fourth helix (tetrahelical HTHs), co-ordination of metal ions and one or two wings (winged helix HTHs).

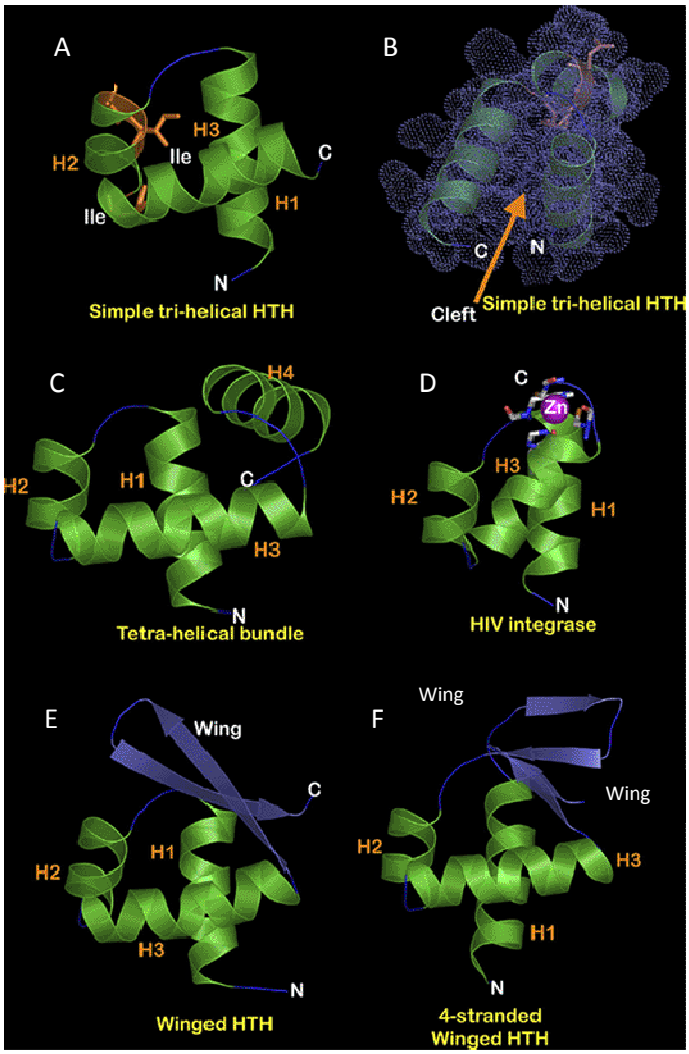


Figure 1.2 Structural variations of the helix-turn-helix motif. Helices are shown in green and strands in blue. A: simple tri-helical HTH domain (PDB: 1k78). Residues that are strongly conserved across all HTH domains are shown in this structure. B: a surface view of the same domain illustrating the shallow cleft. C: tetra-helical bundle (PDB: 1a.4). D: metal-chelating HTH domain of the retroviral integrases with Zn^{+} co-ordinated (HIV integrase, PDB: 1k6y). E: simple 2-stranded winged HTH (PDB: 1smt). F: a 4-stranded winged HTH (PDB: 1cgp). Taken from Aravind and Anantharaman, 2005.

HTH motifs make both specific and nonspecific contacts with DNA. Specific contacts are formed by hydrogen bonds, van der Waals contacts and salt bridges involving the protein and the major groove of the DNA (Brennan and Roderick, 1990; Ohlendorf and Matthew, 1985; Weber and Steitz, 1987). Nonspecific contacts are formed by electrostatic interactions between the negatively charged phosphodiester backbone of the minor groove and positively charged protein residues (Ohlendorf and Matthew, 1985). These nonspecific interactions are often made by flexible arms or loops in the protein adjacent to the HTH motif (Jordan and Pabo, 1988). Proteins containing the HTH motif can be monomeric, heterodimeric or homodimeric (Harrison and Aggarwal, 1990).

1.3.2 Winged helix motif

The winged helix motif is a type of HTH motif which has the general topology of H1-S1-H2-H3-S2-W1-S3-W2, where H designates a helix, S a strand and W a wing (Figure 1.2 E-F; Gajiwala and Burley, 2000). The so called “wings” are large flexible loops. Winged helix proteins [Structural Classification of Proteins (SCOP) classification n 846785] tend to have longer turns in their HTH motifs than other HTH proteins, this results in a greater variety of angles between H1 and H2, ranging from 100° to 150° as opposed to 120° seen in other HTH proteins (Wilson, 1992; Zheng and Fraenkel, 1999).

Winged helix domains can serve a variety of functions within DNA-binding proteins. These include double stranded B-DNA recognition (Lai and Clark, 1993; Passner and Steitz, 1997; Sharrocks, 2001), single stranded RNA recognition (Teplova *et al*., 2006; Wolin and Cedervall, 2002), Z-DNA recognition and

stabilisation (Kim and Khayrutdinov, 2011; Rothenburg and Schwartz, 2002; Takaoka *et al*., 2007) and DNA strand junction and branch recognition (Kitano et al., 2010; Putnam *et al*., 2001; Yamada *et al*., 2004). These functions relate to processes as varied as transcription, DNA repair and RNA splicing (Harami *et al*., 2013 and references therein).

Winged helix proteins typically recognise their target DNA sequences by presentation of helix 3, which is known as the recognition helix, within the major groove of the DNA. Most of the specific DNA contacts are made by polar side chains and the major groove (Gajiwala and Burley, 2000). Nonspecific interactions typically take place between W1 and the minor groove (Harami *et al.*, 2013). Although more rare, specific recognition of the major groove may be mediated by W1 of winged helix motifs (Gajiwala and Burley, 2000).

1.4 DNA structure

DNA structure is affected by and affects DNA-protein recognition. DNA can adopt a number of local and global shape conformations. The global conformations which DNA can adopt are A-, B- and Z- DNA (Figure 1.3). Each of these conformations is differentiated by the handedness of the helices, the pitch (the distance it takes for the helix to make a 360° turn) and the number of bases per turn.

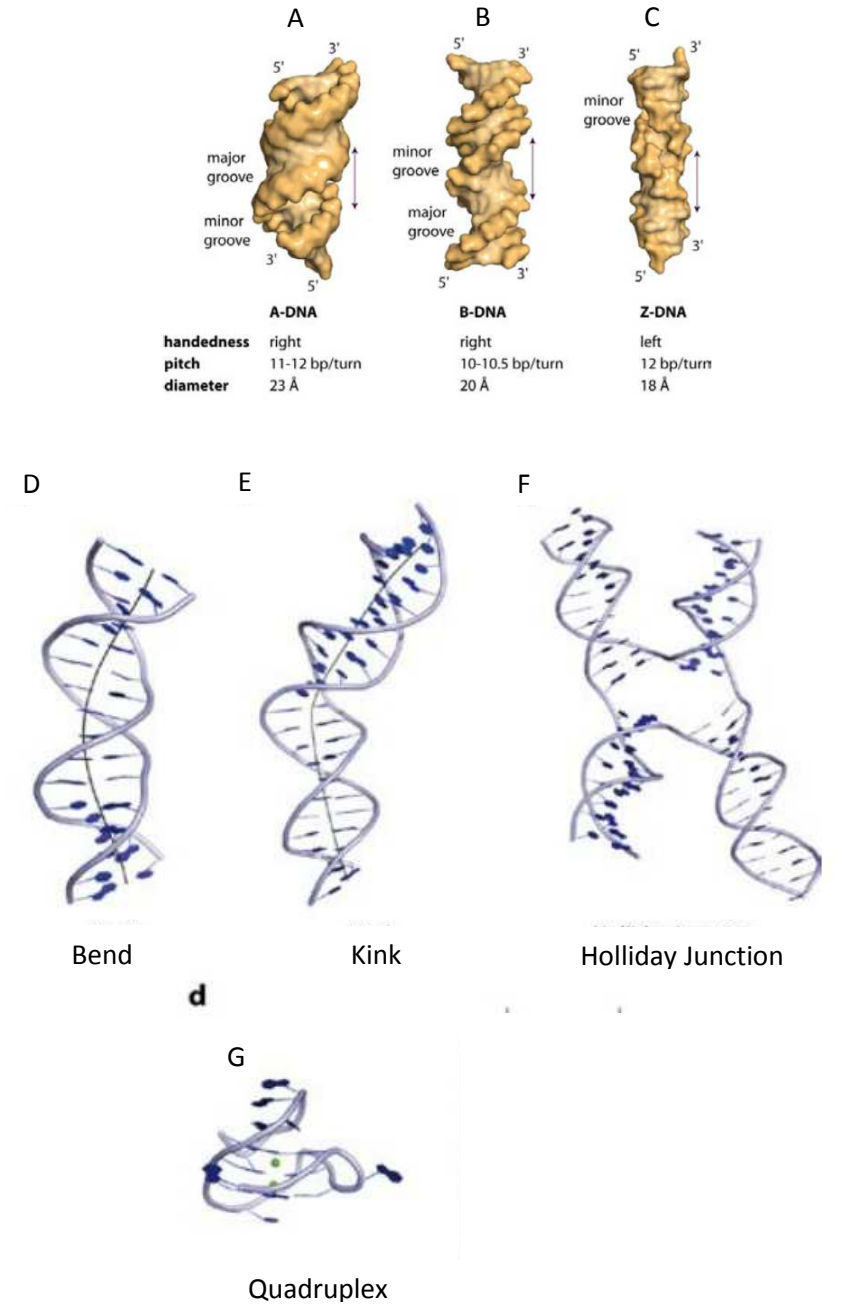


Figure 1.3: Illustration of global and local DNA conformation changes. A: A-DNA; B: B-DNA; C: Z-DNA; D: DNA bend (PDB code 1JJ4); E: DNA kink (PDB code 2KEI); F: Holliday junction (PDB code 2QNC); G: Quadruplex DNA, a four-stranded structure, consisting of guanine-rich sequences, harbouring metal ions (highlighted in green) in the centre (PDB code 3QXR). Adapted from Harteis and Schneider, 2014.

B-DNA is the most common DNA conformation and is the typical structure as suggested by Watson and Crick. It is a right handed helix, the axis of which is through the centre of the helix with 10 bases per pitch and a pitch length of 34 Å.

A-DNA (Figure 1.3 A) is the conformation which dehydrated DNA takes on with 11 bases per pitch and a pitch length of 28.2 Å (Jacobo-Molina and Ding, 1993). It is a right handed helix, the axis of which is shifted from the centre of the helix by approximately 4.5 Å. The distance between the stacked bases is smaller in A-DNA than in B-DNA (2.6 Å and 3.3 Å respectively). Another major difference between A- and B- DNA are the widths and depths of the grooves of the helices. (Ban *et al.*, 1994). B-DNA has a major and a minor groove which are different widths but identical depths whereas the major groove of A-DNA is deep but very narrow and only able to accommodate metal ions and water while the minor groove is wide and shallow (Arnott and Hukins, 1972). The variation in grooves between A- and B- DNA is caused by the shift of the helix axis off centre in A-DNA. A-DNA type helices are probably only found *in vivo* in DNA-RNA hybrid helices and in double stranded RNA helices (Dickerson, 1983).

Z-DNA (Figure 1.3 C) gets its name from the zigzag like pattern of its phosphate backbone. It is a left handed helix with 12 bases per pitch and a pitch length of 43 Å. Z-DNA forms when GC repeats occur at high salt concentrations. In Z-DNA there is only a single uniform groove which corresponds to the minor groove in B-DNA (Wang *et al.*, 1981). Bases recognised in the major groove of B-DNA will be recognised in Z-DNA because they present a similar convex surface. Because negatively charged phosphate groups are forced close together in this configuration, it is not energetically favourable under physiological conditions.

High salt concentrations provide charge shielding of the phosphates resulting in less electrostatic repulsion Thus, the backbone of Z-DNA is only stable at high salt concentrations.

Triple helix DNA structures are formed when a duplex of one polypurine strand and one polypyrimadine strand associates with another ploypyrimadine strand (Felsenfeld *et al*., 1957). The additional polypyrimadine strand associates in the major groove of the duplex and forms non-canonical Hoogsteen base pairs with the polypurine strand (Arnott *et al*., 1976). Triplex DNA is formed during DNA replication, transcription and recombination (Frank-Kamenetskii and Mirkin, 1995).

Local DNA conformation can be altered by bends, kinks, Holliday junctions and the formation of quadruplex DNA (Figure 1.3). DNA kinks are caused when bases become unstacked causing an abrupt disruption to the linearity of the helix and occur most frequently between A and T base steps (Dickerson, 1998). This happens when two adjacent bases lean toward each other on the axis perpendicular to the axis of the helix (Dickerson, 1998). The lean takes place along the edge of the base pair. Kinks change the direction of the helix (Spiriti and van der Vaart, 2012). When more than one kink occurs in a sequence, the effect is cumulative and the result is extensive curvature known as bending (Dickerson, 1998). Bending often occurs as a consequence of protein interaction with DNA.

Holliday junctions (Figure 1.3 F) are formed when four helices of two DNA duplexes exchange strands to make X-shaped structures, also known as cruciform structures (Duckett *et al*., 1988). Holliday junctions are found during DNA

recombination and double strand break repairs as well as being involved in transcriptional regulation (Matos and West, 2014).

Quadruplex DNA (Figure1.3 G) is a four stranded structure facilitated by guanine rich sequence. These structures are transient, unstable and prone to strand breaks as well as chromosomal rearrangements (Lopes *et al*., 2011). Quadruplexes form under physiological conditions during DNA replication (Biffi and Tannahill, 2013) at telomeres but are tightly regulated by the cell. Quadruplex DNA is recognised by specific DNA-binding proteins such as those involved in telomere formation (Paeschke *et al*., 2005).

1.5 Mechanism of protein-DNA recognition

Initially it was thought that certain amino acids would always recognise specific bases and that this "code" was the basis for transcription factor specificity. In 1976, based on the very few structures of DNA-protein complexes available and the chemistry of the atoms near the edge of base pairs, Seeman *et al*. proposed such a code (Seeman *et al*., 1976). However, as the amount of structural data available has increased it has become clear that no such one to one relationship between nucleic acids and amino acids exists (Benos *et al*., 2002).

Jones *et al*. identified three mechanisms of protein–DNA recognition. These mechanisms were termed single headed, double headed and enveloping (Jones *et al., 1999)*. Single headed interactions are characterised by a single cluster of amino acids interacting with DNA, double headed interactions have two such clusters and enveloping interactions have a large surface area of interaction between the protein and DNA. They also found that the vast majority of DNA

binding proteins make contacts in the major groove alone while a much smaller proportion of proteins make contact with both grooves and a very small percentage make contacts with the minor groove alone (Jones *et al*., 1999). This is because the minor groove is too narrow to accommodate the insertion of protein structures and can only do so when the DNA is severely distorted.

1.5.1 Specificity of protein-DNA recognition

By meta-analysis of available structural data Rohs *et al*. (2010) found that all transcription factors in the Protein Data Bank use one of two methods for sequence specific DNA recognition. The first they termed "base readout" while the second they termed "shape readout". Base readout occurs when the transcription factor recognises specific chemical signatures of each base in a DNA sequence. Shape readout involves recognition through sequence specific 3D structures within a DNA sequence. This was the first study to give clear definitions to these types of readout.

1.5.2 Base readout

Base readout occurs when residues within the protein recognise specific bases in the major or minor groove of DNA. Residues can form direct or water mediated hydrogen bonds with bases. A single hydrogen bond very rarely provides specificity but double hydrogen bonds confer a large degree of specificity to a protein-DNA interaction (Coulocheri and Pigis, 2007).

Double hydrogen bonds can be either bidentate (separate donors and acceptors) or bifurcate (shared acceptor with separate donors). Although there is no universal DNA-protein binding code, it has been shown that because of the available donor

and acceptor atoms on base pairs (Figure 1.4), certain amino acids show clear preference for hydrogen bonding with certain bases because they are able to form bifurcated or bidentated bonds involving more than one atom (Luscombe, 2001). It is important to note that in terms of hydrogen bond donors and acceptors, the minor groove of A-T and T-A or G-C and C-G base pairs are indistinguishable allowing less specificity. For this reason only nonspecific binding can take place in the minor groove, while all specific interactions occur in the major groove.

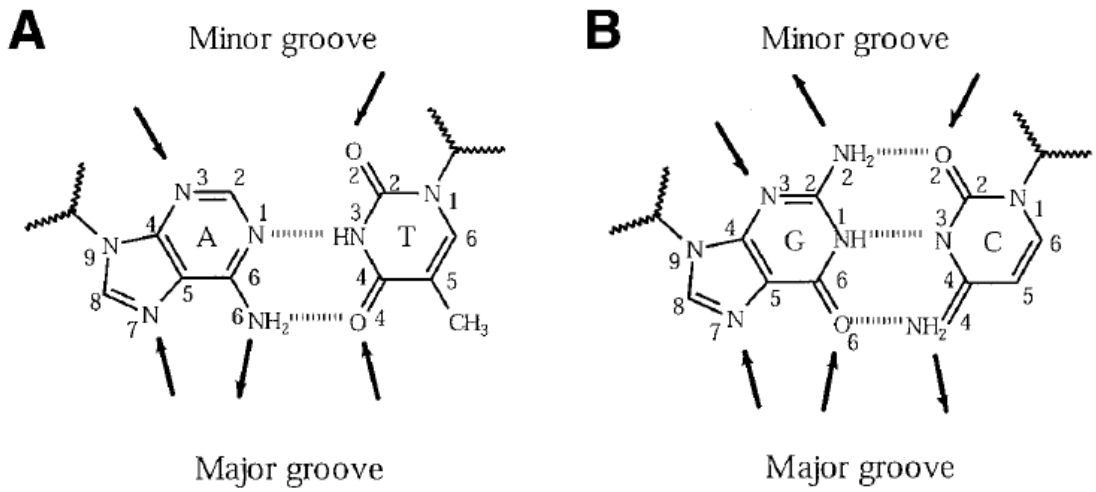


Figure 1.4: Hydrogen bond donors and acceptors of DNA base pairs. Arrows point towards acceptors and away from donors. A – Hydrogen bonds possible with an AT base pair; B- hydrogen bonds possible with a GC base pair. Figure taken from (Luscombe, 2001)

In the major groove arginine, lysine, serine and histidine preferentially bind to guanine while asparagine and glutamine bind to adenine (Luscombe, 2001). Bidentate bonds are responsible for the specificity of arginine and lysine for guanine and the specificity of asparagine and glutamine for adenine (Luscombe, 2001). Examples of these interactions are shown in Figure 1.5.

In the available structures of DNA-protein complexes, arginine is involved in more base interactions than any other amino acid. Luscombe, 2001 suggests that there are three reasons for this: firstly, the length of the side chains which allow them to reach bases within the DNA, secondly, the ability to interact in different conformations and lastly, the ability to produce ideal hydrogen bond geometries. Lysine and glutamine also have long side chains but fail to meet the other criteria. In addition to Hydrogen bonding other forces play a role in sequence specificity. Hydrophobic interactions between the protein and methyl groups of thiamines and cytosines are involved in base readout (Berg and Hippel, 1988). Van der Waals contacts have also been reported to play a role in base readout although most van der Waals contacts are merely stabilising rather than providing sequence specificity (Luscombe, 2001).

Water mediated interactions between DNA and proteins play an important role in the binding process. Water molecules facilitate binding in one of two ways, either by acting as an electrostatic buffer between the negatively charged phosphate backbone of the DNA and negatively charged protein side chains or by allowing protein side chains which are unable to reach DNA because of packing and steric hindrance to make contact with the DNA via water mediated hydrogen bonding (Reddy *et al*., 2001). Divalent cations such as Mg^{2+} are also involved in protein-

DNA interactions as they stabilise the negatively charged phosphates of the DNA backbone (Hartwig, 2001).

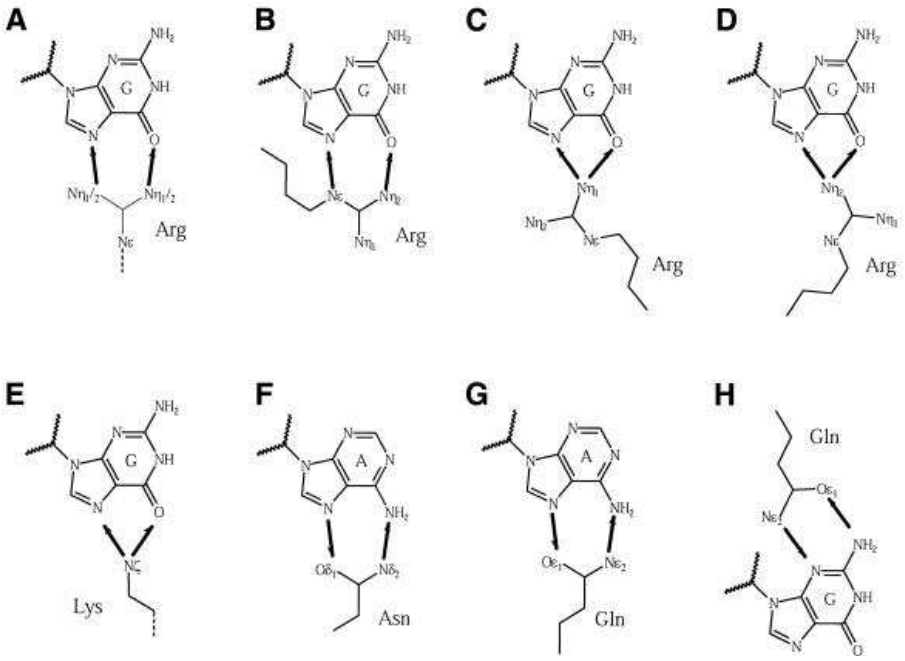


Figure 1.5: The interactions formed between specific amino acids and bases. A-D interactions between arginine and guanine; E interaction between lysine and guanine; F interaction between asparagine and adenine; G interaction between glutamine and guanine; H interaction between glutamine and adenine (taken from (Luscombe, 2001).

1.5.3 Shape readout

Shape readout is the process of DNA-binding which occurs when a transcription factor recognises the shape of the double helix rather than specific bases. This can occur at either a local or global level. Global DNA shape is determined by the environment such as salt concentration and cell cycle stage. Z-DNA and A-DNA have different global structure from typical B-DNA and such structures can be recognised by transcription factors (Takaoka *et al*., 2007; Travers, 1989).

Local shape readout refers to the recognition by the protein of the structure of either the minor or major groove. The most prevalent form of local shape readout occurs at the minor groove. A-T rich DNA has a narrow minor groove because the width of the G-C base pair is greater than that of the A-T base pair (Rohs *et al*., 2009). Because of this, the negatively charged phosphate backbones of the two strands within an A-T rich helix are closer to one another. This leads to a greater overall negative charge in the minor groove of A-T rich regions when compared to G-C rich regions (Jayaram *et al*., 1989). Many transcription factors can recognise the narrow groove and the negative electrostatic potential of A-T rich regions adjacent to the bases with which the protein will form base specific contacts within the major groove. This recognition is usually facilitated by contacts between the DNA phosphate backbone and arginines (Rohs *et al*., 2009).

1.5.4 Dynamic DNA readout

Many DNA binding proteins contain intrinsically disordered regions (IDs) which contribute to DNA recognition by facilitating diffusion or by modulating specificity through base interactions (Billeter, 1997; Mark *et al*., 2005). Most of

these IDs cannot be seen because their inherent flexibility makes them invisible to structure determining techniques such as X-ray crystallography and as such have not been dealt with extensively in DNA-protein interactions as it has not been clear whether such regions are in contact with DNA. Fuxreiter *et al*., 2011 propose that these IDs provide a third type of readout (beyond base and shape readout) which they term as dynamic DNA readout, whereby dynamic regions of highly flexible IDs contribute to DNA recognition. Dynamic interactions of IDs and DNA can orient proteins ideally for binding to take place (Laity *et al*., 2000).

It has been estimated that 70% of DNA binding proteins have intrinsically disordered tails (Vuzman and Levy, 2012) and it is known that these regions contribute to both specific and nonspecific DNA interactions (Spolar and Record, 1994). These tails are often highly charged and appear to facilitate DNA recognition through this charge. In fact, it has been shown that their impact on DNA binding is wholly dependent on their charge distribution rather than their sequence (Vuzman and Levy, 2012).

1.5.5 Cooperativity of protein-DNA recognition

In eukaryotes gene regulation is often dependent on the binding of many individual transcription factors to the same regulatory region making combinatorial cooperativity an important aspect of transcription factor binding (Sarai and Kono, 2005). The favourable energy of cooperatively bound complexes allows multiple proteins or subunits, which on their own have a low affinity for DNA, to bind in a tight multi-protein complex with DNA (Keleher *et al*., 1988). The ability of transcription factors to form multimers increases the range of

sequence recognition allowing for tighter control of gene regulation (Mangelsdorf and Evans, 1995).

The formation of transcription factor dimers in the presence of DNA is an important aspect of gene regulation (Jankowski *et al*., 2013). Cooperative binding of hetero- and homo- dimers leads to an increase in transcription factor specificity because spacing and oritentation of binding sites become additional factors controlling binding (Kazemian *et al*., 2013) Cooperative binding may occur when binding to DNA results in a conformational change in the monomer allowing dimerisation with a second monomer or when both a nearby binding site on the DNA and the bound monomer attract a second monomer (Georges *et al*., 2010).

1.5.6 Facilitated diffusion

It was noted in the 1970s that if transcription factors searched for their target sequences by free diffusion through solution they would not be able to achieve the association rates required for transcriptional control. This led to the proposal by Berg, Winter and von Hippel that DNA binding proteins find their target sequences through "facilitated diffusion" whereby the protein initially binds the DNA non-specifically and then scans the surrounding DNA by either sliding or hopping along the DNA (Hippel and Berg, 1986). This allows many sites to be checked for the cognate sequence in a single binding event. The nonspecific movement along the DNA is facilitated entirely by electrostatic interactions between the positively charged protein and the negatively charged phosphodiester backbone (Florescu and Joyeux, 2009; Givaty and Levy, 2009). The ideal length of each nonspecific search has been determined to be short (Halford, 2009;

Halford and Marko, 2004). This rate allows multiple sites to be searched while allowing the protein to move away from long non-specific stretches of DNA.

1.5.7 Low affinity DNA binding by transcription factors

While it has long been recognised that transcription factors can bind to DNA sites with either low or high affinity, it has until recently been assumed that low affinity binding events are of no functional consequence. This view has been challenged by several studies which indicate that low affinity binding plays a role in gene regulation. A ChIP study in yeast showed extensive low affinity binding of transcription factors (Tanay, 2006). In *Drosophila* it has been suggested that low affinity binding has as much influence on embryonic development as high affinity binding (Segal *et al*., 2006). Badis *et al*., in their extensive study of mouse transcription factor binding motifs described what they called "secondary motifs" for a large number of transcription factors (Badis *et al*., 2009). These secondary motifs were estimated to be bound to their respective transcription factors with a lower affinity than primary sites.

Low affinity binding can lead to both spatial (Jiang and Levine, 1993; Scardigli, 2003; Struhl, 1989; White *et al*., 2012) and temporal (Gaudet and Mango, 2002; Rowan *et al*., 2010) control of gene regulation during embryogenesis when concentration gradients of factors are crucial. The consequences of the disruption of low affinity binding has been shown *Drosophila* where engineering of high affinity binding sites to replace low affinity sites for the transcription factor Hedgehog (Hh) created patterning defects during embryogenesis (Ramos and Barolo, 2013). The reason for these defects is an imbalance between repressor and

activator binding which leads to incorrect gene expression (Ramos and Barolo, 2013)

1.6 Forkhead proteins

The first protein to be named a forkhead protein was forkhead (Fkh) in *Drosophila*. The *fkh* gene was discovered by Jurgens and Weigel in 1988. Mutations in the *fkh* gene cause abnormal head involution in fly embryos giving rise to the spiked head phenotype which gave the gene its name (Weigel *et al*., 1989). In 2011 more than 2000 forkhead proteins had been catalogued (Benayoun *et al*., 2011; SCOP classification 846832). These proteins span 108 opisthokont species.

FOX (from forkhead box) proteins are a superfamily of transcription factors which either repress or activate transcription through the binding of their forkhead domain (FHD) to DNA. The 19 subclasses (A-S) are grouped according to the homology within the FHD only and as such the classes have highly variable structures within other domains (Hannenhalli and Kaestner, 2009). In 2000 a unified nomenclature was decided on for the FOX proteins (Kaestner *et al*., 2000). According to this nomenclature human FOX proteins are denoted in capitals, for example FOXP2, those from mouse with only the first letter capitalised, for example Foxp2 and all other species with the first letter and the subclass capitalised, for example, FoxP2. The Fox proteins are involved in a number of important cellular processes and diseases. These include signalling, neural development and aging (Benayoun *et al*., 2011). More than half of the FOX families have been implicated in cancer these include FOXA in liver cancer (Li *et*

al., 2012), FOXM in lung (Kim *et al*., 2006) and ovarian cancer (Zhao *et al*., 2014) and FOXO in breast cancer (Zou *et al*., 2008) amongst others.

1.6.1 Interaction of FOX proteins with DNA

FOX proteins perform a number of diverse functions upon DNA binding leading to many binding modes within different families. Some FOX families, the FOXA family, for instance, serve as pioneering factors which open up tightly compacted chromatin in order to facilitate the binding of other proteins involved in transcription. The opening of chromatin is carried out by regions outside the FHD which bind to histones 4 and 5 (Carlsson and Mahlapuu, 2002). Others, such as the FOXP family, function as classical transcriptional repressors (Lopes *et al*., 2006; Shu *et al*., 2001).

The FHD adopts a winged helix motif and is the major DNA recognition motif in FOX proteins responsible for specific binding. It typically consists of three alpha helices, three beta strands and two loops, these loops form the two "wings" of the domain (Benayoun *et al*., 2011). FOX proteins make base specific contacts with the major groove of DNA through Helix 3 while the wings which are rich in basic residues make nonspecific contacts with the phosphodiester backbone (Lai and Clark, 1993; Littler *et al*., 2010; Stroud *et al*., 2006; Tsai *et al*., 2006). The Forkhead box consensus sequence has been defined as: 5′ - (G/A)(T/C)(A/C)AA(C/T)A - 3′ (Kaufmann *et al*., 1995; Overdier *et al*., 1994; Pierrou *et al*., 1994). There is, however, another sequence, 5′ - GACGC - 3′, to which a number of FOX proteins bind (Luo *et al*., 2012; Zhu *et al*., 2009). Though there is a general consensus sequence many families of FOX proteins have been

known to have their own preferences for variations of this consensus (Georges *et al.*, 2010). The mechanism for this family-specific sequence specificity is not clear because the amino acids in contact with different bases are conserved across families (Lai and Clark, 1993; Overdier *et al*., 1994). There is, however, a 20 amino acid region located N-terminally to the recognition helix which has been shown to adopt different secondary structures in different families (Marsden et al., 1998) and confers sequence specificity. Overdier et al. (1994) postulate that the FOX proteins have arisen from three separate clades which has led to a diversity of binding-specificities within the superfamily (Nakagawa *et al*., 2013).

Although most FOX proteins bind to DNA as monomers there are notable exceptions such as the homo- and hetero- dimers formed between members the FOXP family (Li *et al*., 2004) and FOXK1 (Tsai *et al*., 2006). Heterodimers formed by members of different FOX families have also been noted, for example the FoxO3a-G1 heterodimer which is involved in the regulation of cell proliferation (Madureira *et al*., 2006).

1.6.2 The forkhead box P family

The P family has four proteins FOXP1-4. *FOXP1* is an important developmental gene in mice, active in the lung and gut (Shu *et al*., 2001). In humans, FOXP1 is implicated in some cases of mental retardation (Le Fevre *et al*., 2013; Hamdan *et al*., 2010; Horn et al., 2010) and gross motor skill delay (Bacon *et al*., 2014; Carr *et al*., 2010) as well as in B-cell and macrophage development (Hu *et al*., 2006; Sagardoy *et al*., 2013). FOXP2 will be discussed in the next section.

FOXP3 is part of the adaptive immune response and T-cell specialisation (Yagi *et al*., 2004). Specifically, FOXP3 controls the differentiation of T_{reg} cells which are specialised CD4 cells involved in autoimmunity and inflammatory response (for a review of T cell function see Josefowicz *et al*., 2012). Dysregulation of FOXP3 has been implicated in the fatal immune disorder, polyendocrinopathy, enteropathy, X-linked syndrome (IPEX; Bennett *et al*., 2001).

FOXP4 is expressed in the embryonic lung and gut (Lu *et al*., 2002) as well as the brain (Rousso *et al*., 2012; Takahashi *et al*., 2008; Tam *et al*., 2011). Foxp4 is expressed in the adult forebrain of rats (Takahashi *et al*., 2008) and in conjunction with Foxp2 is also expressed during neuron diffentiation in the developing central nervous system of the chick (Rousso *et al*., 2012). In mice FoxP4 is involved in the branching of dendrites of Purkinje cells within the cerebellum (Tam *et al*., 2011).

FOXP1, 2 and 4 have similar domain architecture consisting of a polyglutamine tract, zinc finger, leucine zipper and FHD while FOXP3 has an N-terminal proline rich region which replaces the polyglutamine tract of the other FOXPs as well as a truncated C-terminus compared to the other FOXPs (Bettelli *et al*., 2005; Lopes *et al*., 2006). The FHD of the FOXP family is located near the C-terminus rather than the N-terminus as is the case for other FOX families. The wings of the FHD of the FOXP family are different from other FOX families and winged helix proteins in general in that wing 1 is truncated and wing 2 forms a helix rather than a loop (Stroud *et al*., 2006). A schematic comparison of an HTH motif, winged helix motif and a FOXP FHD is shown in Figure 1.6.

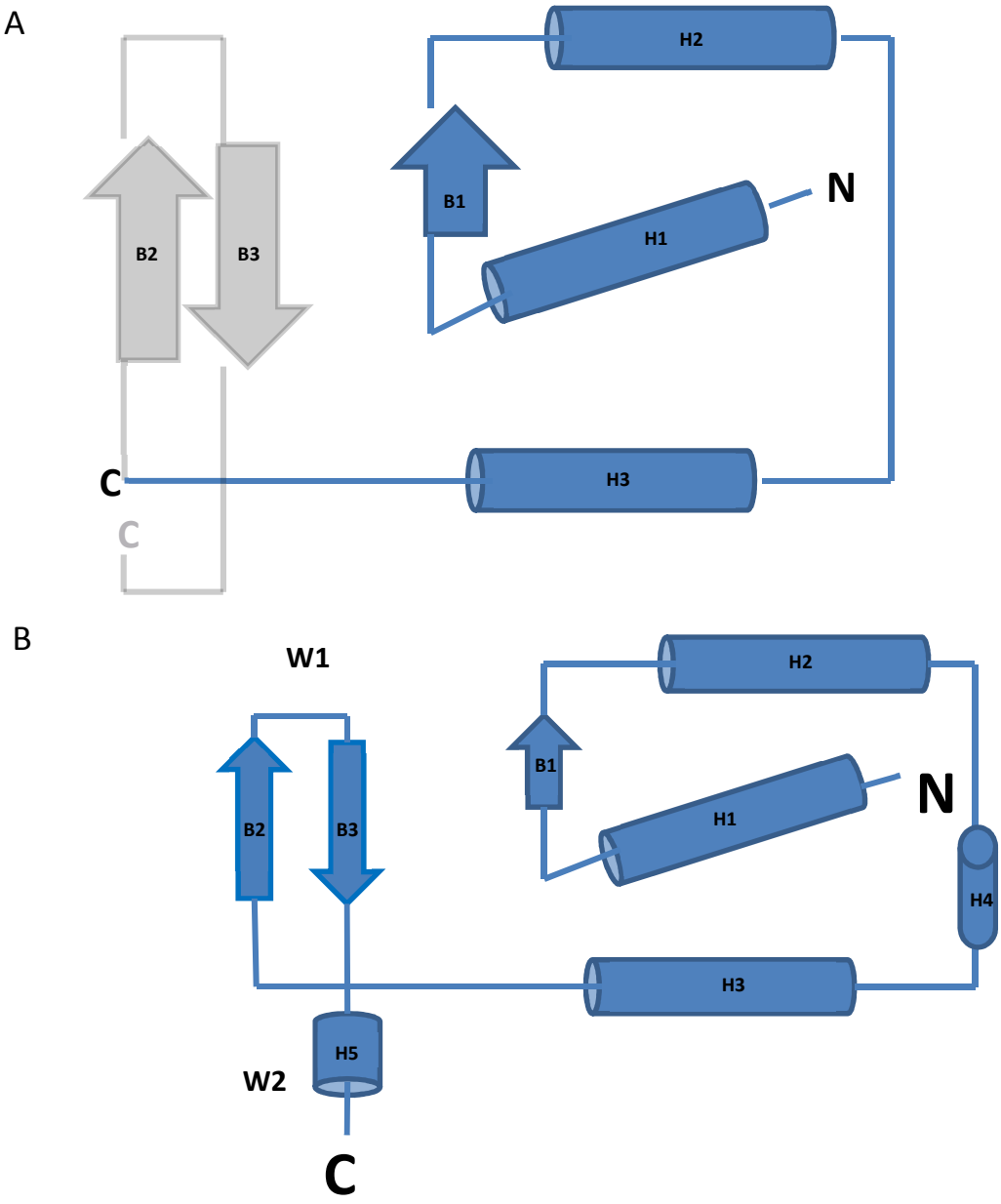


Figure 1.6: Canonical topology of a helix-turn-helix motif, a winged helix motif and a FOXP forkhead domain. A: In a typical winged helix protein the HTH motif (blue) is flanked by two flexible wings, W1 and W2 (grey). B: topology of the FOXP2 FHD in which W1 is truncated and W2 is replaced by a short helix (adapted from Gajiwala and Burley, 2000.

1.6.3 The function of FOXP2

As a transcription factor, the sole function of FOXP2 is to locate and bind specific sequences within the genome. If this binding is functional, networks of other target genes are regulated. Vernes *et al*. (2007) identified biological themes in putative targets of FOXP2 as predicted by ChIP-chip analysis. These themes are membrane bound proteins, proteins involved in homeostasis and locomotive behaviour, as well as axon guidance (Vernes *et al*., 2007). Dysregulation of these networks can lead to numerous pathologies ranging from Specific Language Impairment to breast cancer. When dysregulation of FOXP2 expression occurs it is often difficult to establish the effects on the organism as a whole (French and Fisher, 2014). The role of FOXP2 in language acquisition and other neurological functions, as well as cancers, will be discussed in this section.

The role of FOXP2 in language first came to light when a mutation in the forkhead domain of the protein was found to be the cause of a multigenerational language disorder in the KE family of the UK (Lai *et al*., 2001). Three generations of the KE family suffered from a severe speech and language disorder which was inherited in an autosomal-dominant monogenetic pattern (Hurst and Baraitser, 1990). This disorder has been termed Speech-Language Disorder 1 (SPCH1; MIM 602081) and is characterised by language processing abnormalities as well as restricted orafacial movements affecting speech. The KE family members affected were found to be deficient in the reproduction of phonetic strings (both words and non-words) as well as performing sequences of orofacial movements (Vargha-Khadem *et al*., 1998). Affected family members showed structural abnormalities in a number of regions in the brain, most notably Broca's region which is key to

the production of speech (Watkins *et al*., 2002). Functional magnetic resonance imaging showed that these structural variations corresponded with decreased inactivation of Broca's region during language processing (Liégeois *et al*., 2003)

The disorder causing mutation is that of an arginine replaced by a histidine at residue 553 (R553H) in the DNA binding helix (Lai *et al*., 2001) of the FHD and leads to a disruption of DNA-binding and incorrect cellular localisation of the FOXP2 protein (Vernes *et al*., 2006). More reports of the involvement of FOXP2 in language disorders followed the report of the KE family (Feuk e*t al*., 2006; MacDermot *et al*., 2005; Rice *et al*., 2012; Zeesman and Nowaczyk, 2006). These reports showed patients suffering from verbal dysphasia who had FOXP2 translocations and deletions. To date R553H has been the only point mutation in the FOXP2 FHD to have been implicated in language disorders.

Devanna *et al*. (2014) have shown that FOXP2 expression and the retinoic acid signalling pathway are linked (Devanna *et al*., 2014). Using the Devanna study Benitez-Burraco and Boeckx (2014) suggest that the retinoic acid signalling pathway regulates many putative language related genes. This gene network appears to be under the control of the *RUNX2* gene, the targets of which are involved in brain and skull development.

In addition to evidence in humans of the involvement of FOXP2 in language, studies suggest that the protein also plays a role in the vocalisations of other species such as mice and zebra finch. There is contradictory evidence of the involvement of the mouse KE mutation equivalent (R552H) in subsonic vocalisations in young mice. Groszer *et al.*, (2008) showed it to cause

impairments in motor skill learning and altered synaptic plasticity only, while Fujita and colleagues reported abnormal vocal signalling (Fujita *et al*., 2008). The Foxp2 R552H mutant mice pups of Gaub *et al*., (2010) showed very little difference in subsonic vocalisations from normal pups. This group suggests that pup vocalisations are not affected by Foxp2 mutations because they are innate and not affected by learning circuits in the brain (Gaub *et al*., 2010).A comparison of *FoxP2* mutation studies in mice further complicates the determination of a *FOXP2* phenotype. In the studies reviewed almost all homozygote mutations were lethal while heterozygous mutations produced no phenotype at all or only mild developmental delay (French and Fisher, 2014).

Zebra finch are vocal learners in that songs are taught by male adults to male offspring this makes them a suitable model for instinctive vocal learning (Brainard and Doupe, 2000 and references therein). Area X is the area of the zebra finch brain involved in vocal learning (Sohrabji *et al*., 1990). Zebra finch FoxP2 has been found to have higher levels of expression in Area X than in the rest of the brain (Haesler *et al*., 2004). That FoxP2 is involved in zebra finch song learning is further illustrated by results showing that the gene is up regulated in Area X temporally during vocal learning times (Haesler *et al*., 2004). When FoxP2 is knocked down in juvenile zebra finches, they were unable to learn song accurately (Haesler *et al*., 2007).

FoxP2 is expressed in the developing brains of many species which are not known to use verbal/ auditory pathways including rats (Campbell *et al*., 2009), zebra fish (Bonkowsky and Chien, 2005), medaka (Itakura *et al*., 2008), crocodiles (Haesler

et al., 2004) and frog (Schön *et al*., 2006). This suggests that FOXP2 plays a role in neurological development beyond language.

Neural targets of FOXP2 have been identified in humans by the use of chromatin immunoprecipitation followed by microarray analysis (ChIP-chip) technology. These targets include *FGF8*, a known cortical patterning effector and the homeobox patterning genes *HOXB5* and *HOXB7*, also known to be involved in central nervous system patterning (Spiteri *et al*. 2007). An example of a neural target of FOXP2 is the *CNTNAP2* gene FOXP2 binds to the promoter region of *CNTNAP2* suggesting a role in the regulation of this gene (Vernes and Newbury, 2008). The product of the *CNTNAP2* gene is the transmembrane neurexin CASPR2. This protein has been implicated in human cortical development and may mediate intercellular interactions and laminar organisation (Strauss, 2006). Point mutations in *CNTNAP2* have been associated with aberrant development of language related brain structures in children with autism (Alarcón *et al*., 2008), as well as specific language impairment in children without autism (Vernes and Newbury, 2008).

FOXP2 may also play a role in autism although results have been variable. There has been a significant association between FOXP2 single nucleotide polymorphisms and autism in the Chinese (Gong *et al*., 2004) and Korean (Park *et al*., 2013) populations. However, (Gauthier and Joober, 2003) found no link in Canadian autistic children and in Spain no link was found between autism and FOXP2 either (Toma *et al*., 2013).

FOXP2 has also been associated with psychiatric disorders such as depression and schizophrenia in the Chinese population (Li *et al*., 2013). In addition, a link between schizophrenia and FOXP2 single nucleotide polymorphisms has been found in numerous other studies (McCarthy-Jones *et al*., 2014; Sanjuán *et al*., 2006; Španiel *et al*., 2011; Tolosa *et al*., 2010). The reason for this link may be the effect of FOXP2 on grey matter concentration within schizophrenia related regions of the brain (Španiel et al., 2011).

In addition to their role in neural development, all four members of the FOXP subfamily have been implicated in cancer. FOXP1 is a candidate tumour suppressor (Banham *et al*., 2001; Fox *et al*., 2004), FOXP2 is highly expressed in lymphomas but not in normal lymphocytes (Campbell *et al*., 2010), FOXP3 is involved in adult T-cell leukaemia/ lymphoma (Mansfield *et al*., 2014) and FOXP4 is down-regulated in kidney tumours (Teufel *et al*., 2003).

Yang and colleagues found that the grass carp FoxP1 and FoxP2 are involved in activation of different lymphocyte subpopulations (Yang *et al*., 2010). In humans FOXP1 is involved in B-cell differentiation (Hu *et al*., 2006) and based on the grass carp model it is likely that FOXP2 is also involved. FOXP2 is involved in T-cell regulation through its interactions with the calcium regulated transcription factor NFAT (Wu *et al*., 2006).

Recently a role for FOXP2 in breast cancer has been put forward by Cuiffo *et al*. They found that a class of cells which promote tumour formation and metastasis through interaction with breast cancer cells do so by a micro RNA cascade which leads to down regulation of *FOXP2* in the breast cancer cells. Further, knockdown

of *FOXP2* caused tumour initiation and metastasis in breast cancer (Cuiffo *et al*.. 2014). FOXP2 has also been shown to play a role in the development of certain prostate cancers (Stumm *et al*., 2013).

1.6.4 The structure of FOXP2

The domain architecture of the full length FOXP2 protein consisting of a poly-glutamine region, a zinc finger, a leucine zipper, the FHD and an acid rich tail is shown in Figure 1.7. Between the domains are regions which are predicted to be disordered. It is possible that these regions may contribute to DNA binding via mechanisms discussed for intrinsically disordered regions in Section 1.5.4

The polyglutamine rich region of FOXP2 is one of the longest poly-glutamine tracts known in a human protein. Although expansions of such regions in the genes of other transcription factors such as the TATA binding protein (TBP) have been implicated in disease (Nakamura and Jeong, 2001), the length of this tract is very stable in FOXP2 and has been shown to not be a factor in disease development (Bruce and Margolis, 2002).

The leucine zipper of FOXP2 is involved in hetero- and homo- dimerisation which is essential for transcriptional regulation (Li *et al*., 2004). Heterodimers are able to form between FOXP1, 2 and 4 (Sin *et al*., 2014). The formation of dimers via the leucine zipper is necessary for DNA binding (Li *et al*., 2004) and crucial for transcriptional regulation (Sin *et al*., 2014).

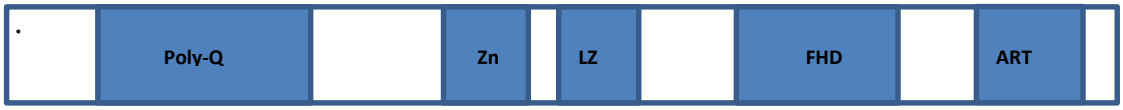


Figure 1.7: Domain architecture of FOXP2. Domains are shown from N to C terminus. Structured domains are shown in blue while white regions are disordered. Poly-Q – glutamine rich region; Zn – zinc finger; LZ – leucine zipper; FHD – forkhead domain; ART- acid rich tail.

1.6.5 Domain swapping in the FOXP2 forkhead domain

The crystal structure of the FOXP2 FHD bound to DNA (Stroud *et al*., 2006) revealed that isolated FOXP2 FHD is able to form domain-swapped dimers (Figure 1.8). This type of dimerisation is not known to occur in the FHD of any other FOX families. Domain swapping occurs when protein chains swap identical units. The swap may involve whole domains or just secondary structural elements (Bennett, 1994). Dimers are formed by domain swapping when there is equal exchange between two monomers. The architecture of individual domains of a domain swapped dimer remains identical to that in the monomer except in the region where the exchange has taken place. The result is essentially two monomers which have folded into one another.

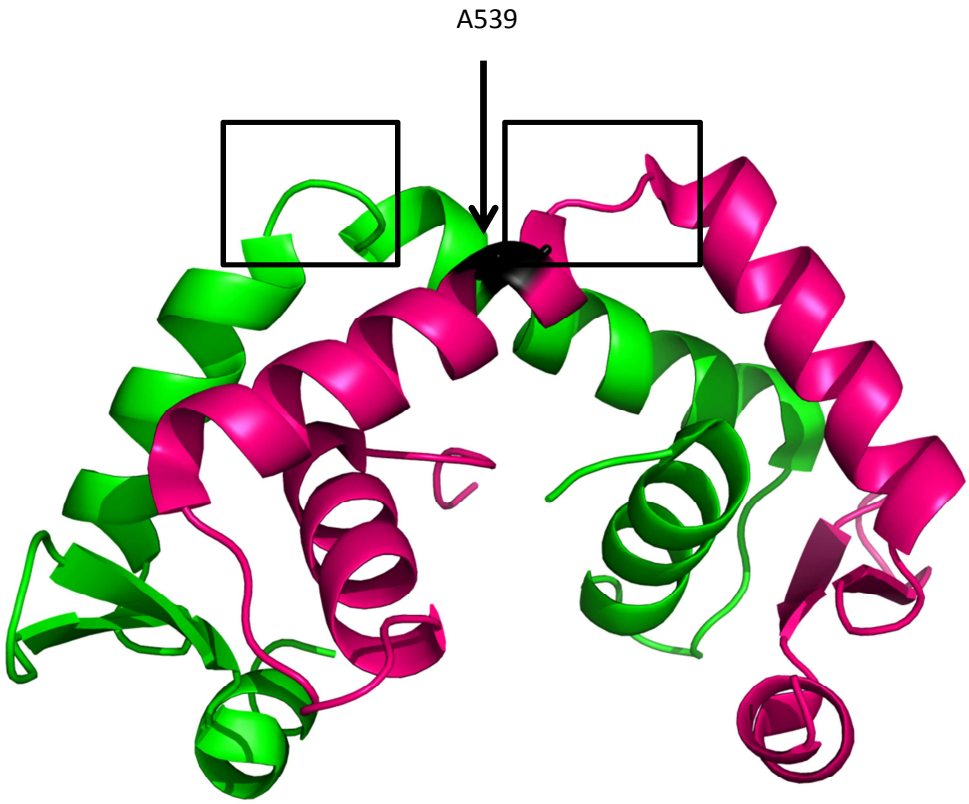


Figure 1.8: Schematic representation of the FOXP2 domain swapped dimer. Two FOXP2 monomers, one shown in green and one shown in pink, swap helix 3 (H3) and strands 2 and 3. Ala539, the mutation of which to a proline prevents domain swapping, is shown in black. The hinge loop region of each monomer is shown by a black box. Visualised using PyMOL (The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.).

Domain swapping in FOXP2 occurs when two monomers exchange helix 2 and strands 2 and 3. Domain swapping occurs when helix 2 extends through the turn separating helix 2 and 3 leading to a 14 residue helix. It is interesting to note that a mutation in this region in FOXP3, which has a high level of sequence identity to FOXP2, has been implicated in IPEX although this area does not directly interact with DNA (Bennett *et al*., 2001). Some IPEX mutations occur in the hinge loop region of the forkhead domain which is necessary for domain swapping (Bandukwala *et al*., 2011). Site-directed mutagenesis of FOXP3 has demonstrated that disruption to domain swapping alone without interference to DNA-binding is enough to lead to T-cell dysfunction (Bandukwala *et al*., 2011) Further, it has been shown that the stable domain swapped dimer of the FOXP3 FHD bridges two pieces of DNA mediating long range chromosomal contacts which are crucial to its function (Chen *et al*., 2015). These findings suggest that domain swapping may play a physiological role in the FOXP family. Given this putative role, it is unlikely, though still possible, that domain swapping is an artefact of the high concentrations of protein used during crystallisation, as is suggested for some other domain swapped structures obtained by crystallisation (Gronenborn, 2009).

Domain swapping in the FOXP FHD appears to be mediated by Ala539 in the hinge loop region, as the mutation of this residue to proline prevents dimerisation (Stroud *et al*., 2006). In members of other FOX families there is a highly conserved proline at this position. Stroud et al. propose that this proline in other classes of FOX proteins acts as a helix breaker preventing the formation of the extended helix which facilitates domain swapping in the FOXP family. Using multi-angle light scattering, Stroud *et al*. demonstrated that the A539P mutant of

the FOXP2 FHD existed solely as a monomer whereas the wild type existed as a mixture of monomers and dimers under the same conditions.

1.6.6 The interaction of the FOXP2 forkhead domain with DNA

Although a Foxp core consensus binding sequence of AAAT has been put forward by Wang *et al*. (2003), it was identified, using the *in vitro* technique of cyclic amplification of ligands, by DNA binding with Foxp1. This core sequence was the one used in the crystal structure of the FOXP2 forkhead domain bound to DNA (Figure 1.9; Stroud *et al*., 2006). Vernes *et al*. (2007) and Spiteri (2007) identified variations of the classical FOX consensus sequence similar to the Wang sequence in ChIP experiments. From data generated by gene expression studies of humanised *FOXP2* expressed in mice Enard and colleagues predicted TATTTAT as a FOXP2 binding sequence (Enard *et al*., 2009). In 2013 Nelson et al. reported a different sequence obtained by microfluidic chip analysis, TGTTTAC, which FOXP2 preferentially bound *in vitro* (Nelson *et al*., 2013).

From the crystal structure of the FOXP2 FHD in complex with DNA, it can be seen that the interaction differs from that other FOX proteins and DNA. Largely this is due to the significant difference in the structure of the wings between FOXP and other FOX families (Stroud *et al*., 2006). Figure 1.10 shows the contacts between the FOXP2 FHD and DNA, as well as those between the FOXA3 FHD and DNA (Stroud *et al*., 2006). Because Wing 1 is truncated and Wing 2 is a short helix in FOXP, the wings make very little contribution to DNA binding (Stroud *et al*., 2006). In other FOX families, such as FOXA, these wings make extensive contacts with the phosphate backbone and the minor groove of the

DNA (Lai and Clark, 1993). This discrepancy led Stroud *et al*. to predict a lower affinity of binding in the FOXP family (Stroud *et al*., 2006).

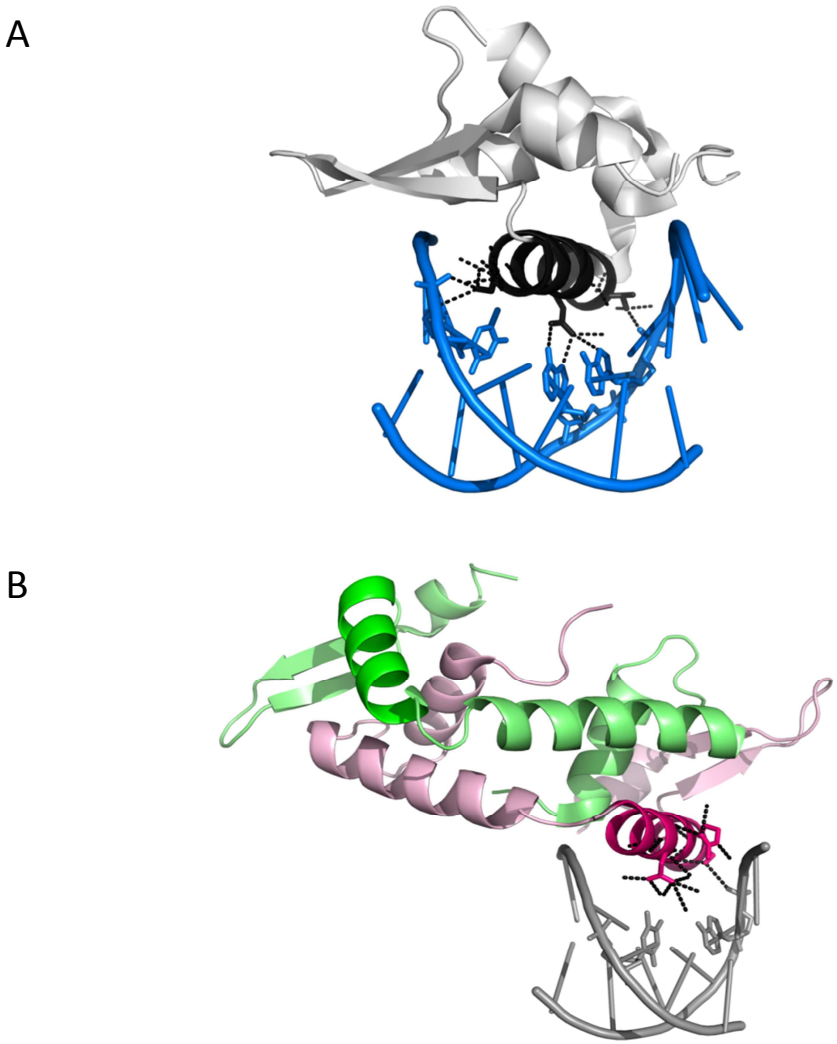


Figure 1.9: Ribbon representation of the FOXP2 forkhead domain bound to DNA (PDB 2A07). FOXP2 monomer (A) and domain swapped dimer (B) in association with the core consensus sequence proposed by Wang *et al*., 2003. The two monomers of the domain swapped dimer are coloured in green and pink respectively. Helix 3 in both monomer and dimer is opaque while the rest of the protein is transparent. Hydrogen bonds formed between the protein and DNA are

indicated by black dashed lines. Figure generated using PyMOL (The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.).

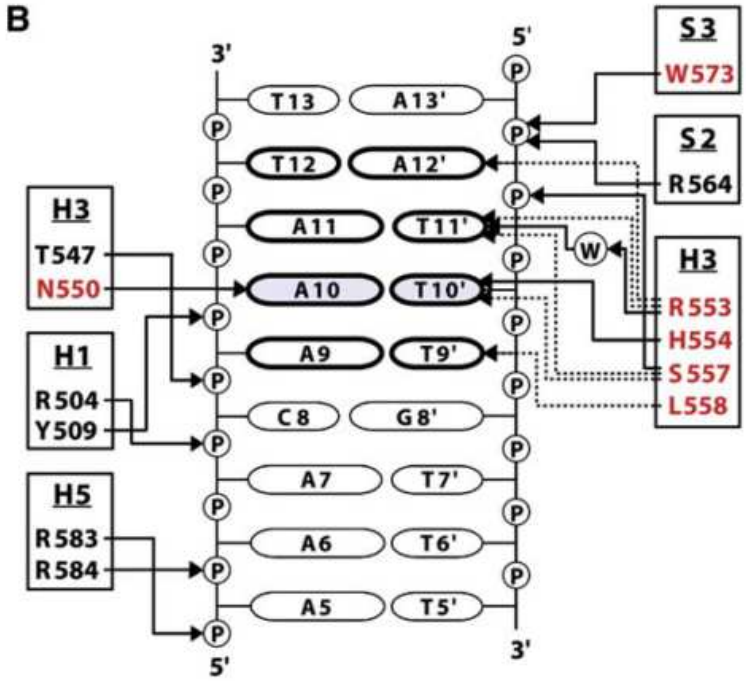


Figure 1.10: DNA contacts made by the FOXP2 forkhead domain. DNA is shown as a ladder with phosphates represented as circles and bases as ovals. Hydrogen bonds are solid lines while van der Waals contacts are dashed lines. The residues which are also involved in DNA recognition in FOXA3 which represents a typical FOX protein are shown in red to highlight the variation in binding patterns between the FOXP2 FHD and other FOX proteins (Taken from Stroud *et al*., 2006).

Despite the differences in wing contacts, the FOXP2 FHD appears to use the same conserved residues as in other FOX families within helix 3 to make major groove contacts (Figure 1.10). These amino acids make contact with different bases to those in other FOX families but this is in agreement with the observation of Georges et al. that FOX families do not achieve family-specific sequence specificity through differences helix 3 protein sequence (Georges *et al*., 2010).

Stroud et al. (2006) speculate that the FOXP2 forkhead domain may recognise a variety of DNA sequences because of the nature of the contacts made with DNA. This would be consistent with the findings of Badis *et al*. (2009) which showed that many transcription factors bind more than one sequence and that each sequence is bound with varying affinity. This allows differential levels of regulation by a single transcription factor.

The dimer in the crystal structure makes fewer contacts with the DNA than the monomer does (Figure 1.9; Stroud *et al*., 2006). This is inconsistent with the crystal structure of FOXP3 in complex with DNA (Bandukwala *et al*., 2011; Chen *et al*., 2015). In the FOXP3 structure all protein is present as domain swapped dimers with each monomer tightly bound to distinct sequences (GTTTCA and AATTTGT) on different DNA molecules forming a bridge between them (Bandukwala *et al*., 2011; Chen *et al*., 2015). In the FOXP2 crystal structure there are six protein molecules in the asymmetric unit of the crystal, two are present as monomers while the other four make up two domain swapped dimers. The monomers and dimers interact with the same sequence and only one monomer of

the domain swapped dimer is in contact with DNA while the recognition helix of the other monomer in the dimer is not in contact (Stroud *et al*., 2006). This is further indication that the DNA sequence used in the crystal structure of the FOXP2 FHD is probably not the functional binding sequence of the protein *in vivo*.

1.7 Problem identification

As a transcription factor, the sole function of FOXP2 is to regulate transcription of target genes. In order to understand this regulation it is necessary to establish the DNA sequences with which the FOXP2 FHD can interact. This will shed light on the specificity of the interaction between the FOXP2 FHD and DNA. The strength and kinetics of binding to these different sequences will give an indication of whether interactions with these sequences are likely to lead to a functional regulatory event.

The specificity and affinity of the interaction of the FOXP2 FHD and DNA will provide information on the mechanism of binding utilised by FOXP2 and the FOXPs in general. To date no kinetic or thermodynamic studies of the binding of any of the FOXP FHDs and DNA have been undertaken and as such no mechanism of binding has been proposed. Identification of the mechanism of DNA binding utilised by FOXP2 may shed light on low affinity binding and variable sequence specificity in the FOX proteins.

2. Aims

2.1 Overall aim:

To establish specific binding mechanisms (kinetic parameters, strength of binding, as well as the type and position of bonds formed) of the FOXP2 forkhead domain and all possible cognate DNA sequences.

2.2 Specific objectives:

1. Overexpress and purify the FOXP2 forkhead domain.

2. Determine the oligomeric state of the FOXP2 forkhead domain

3. Create and purify the A539P monomeric mutant of the FOXP2 forkhead domain

4. Identify novel cognate DNA sequences that bind the FOXP2 forkhead domain

5. Establish kinetic and thermodynamic parameters of binding between the FOXP2 forkhead domain and various cognate DNA sequences in order to determine which sequences the FOXP2 forkhead domain binds with the greatest affinity

6. Use molecular modelling to predict which residues of the FOXP2 forkhead domain interact directly with DNA and which types of bonds are formed.

7. Use structural alignment to predict conformational changes to the FOXP2 forkhead domain upon binding to various sequences.

3 Experimental Procedures

3.1 Creation of monomeric mutant of the FOXP2 forkhead domain

3.1.1 The pET-30 expression system

The pET-30 expression system (Novagen) is isopropyl β-D-1-thiogalactopyranoside (IPTG) inducible and results in a fusion protein with a 6 X His tag and 15 amino acid S-Tag™ at the N-terminus. These tags are cleavable through an enterokinase cleavage site. The pET-30 plasmid also contains the *Kan*^{R} gene which confers kanamycin resistance and allows for selection of transformed cells by kanamycin containing media.

3.1.2 Plasmid extraction

In order to perform site directed mutagenesis on the pET-30 vector containing the FOXP2 FHD (received as a kind gift from Prof. L. Chen, of the University of Southern California Los Angeles, CA), the plasmid was purified using the Genejet Miniprep Kit (Fermentas). Purification was carried out according to the manufacturer's instructions using 1,5 ml aliquots of BL21 (DE3) pLys cells grown in LB for 16 hours. DNA concentrations were determined spectrophotometrically at 260 nm.

3.1.4 Site-directed mutagenesis

In order to create the A539P mutant of the FOXP2 forkhead domain, site directed mutagenesis was performed using the QuikChange® kit (Stratagene). The following primers were designed by changing the CGA codon for alanine into the GGA codon for proline and were synthesised by Inqaba Biotec (Pretoria, South Africa) and used to introduce the A539P mutation into pET-30 containing an insert coding for the wild type FOXP2 forkhead domain:

5’ GGTTTACACGGACATTTCCTTACTTCAGGCGTAATG 3’

5’ CATTACGCC TGAAGTAAGGAAATGTCCGTGTAAACC 3’

The mutated codon for proline is indicated on both primers in red.

The thermo-cycling conditions for mutagenesis were as follows: 2 min at 95 °C; 18 x (20 sec at 95 °C; 10 sec at 60 °C; 2,5 min at 68 °C) 2 min 72 °C. The 50 µl reaction mixture contained 1X Quikchange® Lightening Buffer, 100 ng pET-30 containing the FOXP2 FHD, 125 ng each of forward and reverse primers, 0,2 mM dNTP mix, 1X Quiksolution reagent, and 1 unit of *Pfu Turbo* DNA Polymerase (Quikchange® Lightning Enzyme).

3.1.5 Transformation of competent cells with plasmid DNA

Transformations were carried out using the heat-shock method. Competent cells - *E. coli* T7 Express (New England Biolabs) were thawed on ice. Approximately 50 ng of chilled pET-30 containing the wild type FOXP2 FHD or the A539P FOXP2 FHD insert was added to 50 µl of cells. The cells were then incubated on ice for 30 min and then heat shocked at 42 °C for 45 sec and placed back on ice for a further 5 min. After this, 500 µl of SOC (Super Optimal broth with Catabolite)

media (Novagen; 2 % typtone, 0.5% yeast, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 40 mM glucose), was added and the cells were incubated with agitation at 37 °C for one hour to allow the *Kan*^{R} gene to be expressed. Cells were then plated on LB agar (1% tryptone; 1% NaCl; 0.5% yeast; 1.5% agar) containing 50 µg/ml kanamycin and grown overnight at 37 °C. Plasmid from a colony was purified using the GeneJet™ Mini Prep kit (Fermentas) and sent for Sanger sequencing at Inqaba Biotec (Pretoria, South Africa) in order to confirm the correct insert.

3.2 Overexpression and purification of the FOXP2 forkhead domain and the A539P mutant

3.2.1 Expression trials

To obtain ideal expression conditions of the FOXP2 FHD and the A539P mutant an expression trial was undertaken. Cells - T7 Express (New England Biolabs) - were grown in 2YT (1.6% tryptone; 0.5% NaCl; 0.5% yeast extract; 30 ng/ml kanamycin) at 37 ºC overnight shaking at 250 rpm. A 100 X dilution of the overnight culture was prepared in LB supplemented with kanamycin and grown until the optical density at 600 nm reached 0.6 (mid log phase). Cells were then induced with either 0.5 or 1 mM IPTG and samples taken every 2 hours for 16 hours. These cells were harvested by centrifugation at 10 000 g in a desktop microfuge and then prepared for SDS-PAGE and electrophoresed as described below. Stroud's conditions were confirmed and 1 mM IPTG was used to induce cells for 4 hours in all subsequent expression of both the wild type FOXP2 FHD and the A539P mutant.

3.2.2 Immobilised metal ion chromatography

In order to obtain pure FOXP2 FHD and the A539P mutant immobilised metal ion chromatography (IMAC) was used to capture the 6X His-tag. Four litres of T7 Express cells expressing the wild type FOXP2 FHD or A539P mutant were grown and induced as described in the previous section. Cells were harvested by centrifugation at 4 200 x g for 30 min and resuspended in binding buffer (0.5 M NaCl; 20 mM Tris HCl pH 7.4; 50 mM imidazole) and sonicated in order to lyse them. The soluble protein fraction was then collected after centrifugation at 26 900 x g for 30 min. A 5 ml column of His-Trap resin was charged with nickel according to manufacturer's instructions and equilibrated with ten column volumes of binding buffer. After loading the soluble cell fraction, 5 column volumes of binding buffer were run through the column. The protein was eluted using five column volumes elution buffer (0.5 M NaCl; 20 mM Tris-HCl pH 7.4; 1 M imidazole).

3.2.3 Determination of the FOXP2 forkhead domain and A539P mutant purity and concentration

3.2.3.1 Sodium dodecyl sulphate polyacrylamide gel electrophoresis

Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) was used to determine protein purity and was performed by the discontinuous method. A 12 % acrylamide separating gel was prepared with 0.375 M Tris/HCl buffer (pH 8.8) and 0.1% SDS. A 4% acrylamide stacking gel was prepared with 0.0625 M Tris/HCl buffer (pH 6.8) and 0.1% SDS. Samples were prepared for electrophoresis by diluting them 1:1 with SDS sample buffer (2% SDS, 5% β-

mercaptoethanol, 10% glycerol and 0.02% bromophenol blue) and heating at 95 ºC for 5 minutes in order to fully denature proteins. Molecular weight markers were also prepared in the same manner.

After application of samples to wells, gels were electrophoresed at 120 V until the bromophenol blue had migrated to approximately 1 cm of the bottom of the gel. The electrophoresis buffer was made up of 0.025 M Tris and 0.192 M glycine buffer (pH 8.6) containing 1% SDS. After electrophoresis, gels were stained (2% Coomassie blue R250, 13.3% glacial acetic acid and 18.75% ethanol) for 1 hour and then destained (40% ethanol and 10 % glacial acetic acid) overnight.

3.3 Confirmation of the identity of the wild type FOXP2 forkhead domain and the A539P mutant

In order to confirm the identity of the wild type FOXP2 FHD and the A539P FOXP2 FHD mutant, sequencing by liquid chromatography mass spectrometry/ mass spectrometry (LC MS/MS) was performed as a commercial service at the Council for Scientific and Industrial Research (CSIR), Pretoria. Protein samples (10 µM) were digested overnight using 10 µg/µl trypsin and then made up to a concentration of 2% acetonitrile/ 0.2% formic acid. Fragments obtained from the digest were separated using a Dionex Ultimate 3000 RSLC (rapid separation liquid chromatography) system and then ionised by electron spray ionisation and analysed using a QSTAR ELITE mass spectrometer. The obtained mass spectra were compared to the UniSwiss data base using the Paragon search engine (Shilov *et al*., 2007) in Protein Pilot (version 4.0.8085).

3.4 Confirmation of the structural integrity of the FOXP2 forkhead domain A539P mutant

Fluorescence measurements of the wild type and A539P mutant FOXP2 forkhead domain at a concentration of 5 µM were carried out at 20 °C on a Jasco FP_6300 fluorimeter with slit-widths set to 5 nm. A 295 nm excitation wavelength was used in order to measure fluorescence solely from tryptophan residues.

3.5 Determination of the oligomeric state of the FOXP2 forkhead domain and A539P mutant

Size exclusion chromatography (SEC) was performed on a 150 ml Hiload™ 16/600 75 pg size-exclusion column (GE Healthcare) using the ÄKTA fast protein liquid chromatography system (GE Healthcare). The column was equilibrated using equilibration buffer (20 mM Tris-Cl pH 7.6, 150 mM NaCl) and subsequently loaded with 2 ml of 40 µM protein sample which was eluted at 1 ml/min over one column volume. SEC was performed on the wild type FOXP2 FHD and the A539P mutant.

3.5 Identification of novel FOXP2 cognate DNA sequences

3.5.1 Systematic evolution of ligands by exponential enrichment

A variation of the method of (Rotherham *et al*., 2012) was used to perform SELEX and MonoLEX. Experimental procedures were carried out at the Council for Scientific and Industrial Research (Pretoria, South Africa) The starting library in both cases was a single stranded 90-mer randomised at 49 internal positions flanked by constant regions for polymerase chain reaction (PCR) amplification.

Thus the library sequence was:

5′-GCCTGTTGTGAGCCTCCTAAC(N49)CATGCTTATTCTTGTCTCCC-3′

and the primer sequences were: 5′-GCCTGTTGTGAGCCTCCTAAC-3′ (forward primer) and 5′- GGGAGACAAGAATAAGCATG-3′ (reverse primer). Initial amplification of the library to double stranded DNA as well as amplification of bound DNA in between rounds for SELEX was carried out by PCR in 50 µl reactions (50 ng template DNA; 1X Taq buffer; 0.2 mM dNTPs; 1.5 mM $MgCl_2$; 1 µM forward primer; 1 µM reverse primer; 2.5 U Taq polymerase) with the following thermo-cycling conditions: 95 °C for 3 min, 10 cycles at 95 °C for 1 min, 54 °C for 1 min and 72 °C for 90 sec, followed by a final extension at 72 °C for 8 min.

In the SELEX procedure the DNA library was negatively selected against the nitrocellulose membrane used for partitioning to prevent the detection of DNA which bound non-specifically to the membrane. Binding reactions were carried out with 50 ng of DNA and 50 ng of the wild type FOXP2 FHD in HMCKN binding buffer (20 mM HEPES, 2 mM $MgCl_2$, 2 mM $CaCl_2$, 2 mM KCl and 150 mM NaCl, pH 7.4). The binding mixture was then passed through a clean nitrocellulose membrane equilibrated with HMCKN buffer. The membrane was then washed several times with HMCKN buffer to remove unbound DNA and the remaining bound DNA was recovered from the membrane using 7 M urea and purified using standard phenol:chloroform extraction (using the protocol of Sambrook and Russel, 2006) and isopropanol precipitation. The DNA recovered

from each round was then amplified as described above and used as the starting material for the next round of selection.

In the MonoLEX procedure the DNA library was negatively selected against magnetic epoxylated beads (ReSyn Biotechnologies, Pretoria). The wild type FOXP2 forkhead domain was then amine coupled to the beads according to the manufacturer's instructions. Beads were washed thoroughly with HMCKN buffer and 50 ng of DNA was added. Unbound DNA was removed with several washes of HMCKN buffer and the bound DNA was removed and purified as described for SELEX.

3.5.2 Motif identification

DNA obtained from the final round of SELEX and MonoLEX was sent to the University of Stellenbosch for next generation sequencing (NGS) on the Ion Torrent platform. The two data sets from MonoLEX (DS001) and SELEX (DS002) were cleaned of poor data (Phred score < 20) and trimmed of primer sequence in Galaxy Suite (Blankenberg *et al*., 2010). Any sequences which were greater than 51 or smaller than 48 were discarded as being invalid because the starting library contained 49 randomised bases. The clean datasets were then analysed using DREME (Discriminative Regular Expression Motif Elicitation; Bailey, 2011) which uses the DREME algorithm which is optimised for large datasets and AME (Analysis of Motif Enrichment; McLeay and Bailey, 2010) which uses the PASTAA algorithm to identify enriched motifs within the sequences. Using MAST (Motif Alignment and Search Tool; (Bailey and Gribskov, 1998) the DS001 was then checked for motifs generated from DS002

and DS002 was checked for motifs generated from DS001. Fischers Exact Score was used as an indication of the validity of all motif results.

3.5.3 DNA preparation

Figure 3.1 illustrates the DNA sequences that were used in subsequent work. The sequence obtained in this work (referred to as Webb) has been compared with published FOXP2 FHD binding motifs in the rest of this work.

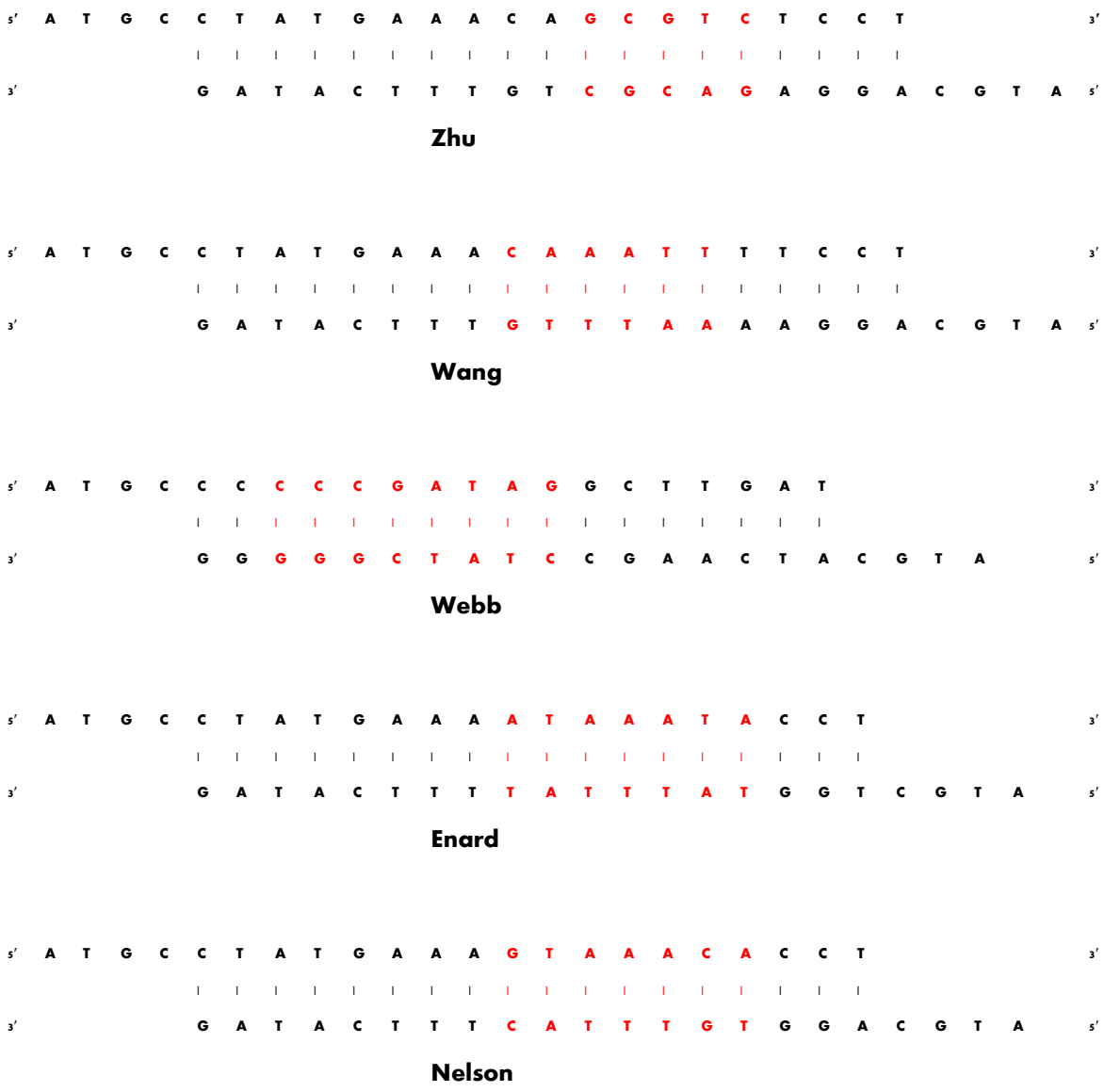


Figure 3.1: DNA sequences used to determine the kinetic parameters of the binding of the FOXP2 FHD and DNA. Overhangs and biotinylation were for coupling purposes required for surface plasmon resonance. Predicted core binding sites are displayed in red.

The oligonucleotides with the sequences listed in Figure 3.1 were synthesised by Integrated DNA Technologies (Coraville, IA). For the Webb sequence, flanking sequence was taken from the sequence dataset while for the published motifs, flanking sequence was taken from the DNA sequence used for crystallisation (Stroud *et al*., 2006). Over-hangs were included to allow mobility of the duplexes on the SPR chip.

3.6 Determination of the kinetic and thermodynamic parameters and binding affinity of the FOXP2 forkhead domain binding various DNA sequences

3.6.1 Surface plasmon resonance

All SPR measurements were carried out on a BIACore 3000 biosensor at the CSIR, Pretoria. For initial screening of putative binding sequences of the FOXP2 FHD, FOX storage buffer was used as the running buffer for SPR (20 mM HEPES; 150 mM NaCl; pH 7.4). Carboxymethyl dextran chips (XanTec CMD500M; XanTec, Dusseldorf, Germany) were used.

In order to immobilise DNA to the chip surface, the DNA sequences listed in Figure 3.1 were synthesised as duplex DNA by IDT (Coralville, IA) with 5′ biotin labels on one of the strands. The biotin label was used to capture the DNA on the chips which had streptavidin amine-coupled to the surface. In order to amine-couple the streptavidin, the biosensor chips were activated by flowing 50 µl of 1:1 0.5 M N-hydroxysuccinimide (NHS) and 0.2 M N-ethyl-N'-(3-dimethylaminopropyl) carbodiimide hydrochloride (EDC) was injected at a flow rate of 10 µl/minute. This was followed by the injection of 50 µl of 1 mg/ml

streptavidin (made up in acetate buffer: 10 mM, pH 4.5), into three of the four flow cells (1 to 3). In order to block uncoupled, active amine sites, ethanolamine-HCl (1 M, pH 8.0) was injected over all four flow cells.

The biotinylated DNA sequences made up in nuclease-free water were then injected onto the chip at a concentration of 10 µM. Each chip contained 3 individual DNA sequences, one per flow cell, with the fourth cell being used as a blank. DNA was injected repeatedly until approximately 500 RU of DNA was coupled.

Fifty microliters of 50 µM FOXP2 FHD A539P mutant was injected over all four flow cells at 10 µl/min. The monomeric mutant was used instead of the wild type in order to ensure that all signals were from the DNA binding event without any possibility of dimerisation events convoluting data. The quantity of binding for each sequence was determined by subtracting the response units (RU) from the blank flow cell and then determining the difference in RU after protein injection compared to prior to injection.

Measurements of the kinetics of binding of the various DNA sequences to the FOXP2 FHD A539P mutant were also measured using SPR on the BIACore 3000 instrument as outlined above. Extensive optimisation of buffer conditions and chip surfaces were required because nonspecific binding of the positively charged FOXP2 FHD A539P mutant to the negatively charged carboxymethyl dextran surface of commonly used sensor chips was likely to occur. This type of nonspecific binding is a common problem when looking at DNA-transcription factor binding using SPR (Majka and Speck, 2007). Many methods of decreasing

nonspecific binding of analyte to the chip surface cannot be used when looking at DNA-transcription factor interactions. Salt concentration of buffers cannot be increased as this prevents electrostatic interactions between positive residues and the negatively charged DNA backbone. It is also not advisable to swap the ligand and the analyte, coupling the protein to the chip and flowing DNA over as arginines and lysines involved in amine coupling are key residues for DNA-protein interaction. A phosphate buffer system was used rather than the standard HEPES buffers supplied by BIACore because it was found during buffer optimisation that HEPES buffers disrupted binding, as evidenced by poor signal. Tween® added to buffers prevented hydrophobic interactions between the protein and chip surfaces. Rather than a standard carboxymethyl dextran chip, a carboxylated chip (XanTec HC30M; XanTec Dusseldorf, Germany) was chosen as it did not have negatively charged dextran and also had lower density of carboxymethyl groups on the chip surface giving it a lower charge density. To further reduce the charge density, a strategy of repeated activation followed directly by blocking before re-activating was used to ensure that many of the carboxymethyl groups on the chip were bound to ethanolamine.

Biosensor chips were prepared as described above with the exception of the buffer system used and that the activation and blocking steps described were performed twice before streptavidin was immobilised. All kinetics experiments were performed using phosphate buffered saline (PBS) containing Tween® 20 (0.138 M NaCl; 0.0027 M KCl; O.05% Tween ® 20; pH 7.4). Increasing concentrations of protein in the range of 9 to 244 nM were injected at 75 µl/min over the chip and allowed to return to equilibrium over 600 seconds. The bound protein was

removed from the chip by a 1.5 M NaCl regeneration step after each concentration. Data was analysed using the BIAEvaluation software.

3.6.2 Isothermal titration calorimetry

ITC was carried out on a MicroCal VP-ITC instrument. The DNA and protein solutions were dialysed extensively against the same batch buffer (20 mM Tris; 150 mM NaCl). The experiment was carried out at 20 °C. The sample cell contained 1.33 ml of 150 µM A539P FOXP2 FHD and the syringe contained 5 µM of the Nelson DNA sequence (see Figure 3.1). These concentrations were determined through several trial runs of the experiment at varying concentrations. The monomeric mutant of the FOXP2 FHD was used to ensure that dimerisation would not convolute the binding signal. Titrations consisted of 40 injections of 6 µl with 300 seconds between injections. Data was fit to a sigmoidal function using the MatLab software.

3.7 *In silico* prediction of the residues and bases involved in the interaction between the FOXP2 forkhead domain and various DNA sequences

3.7.1 Modelling of DNA sequences

The 5 DNA sequences from Figure 3.1 were modelled as double stranded B-DNA with Watson Crick pairing using the 3D-DART (3DNA-Driven DNA Analysis and Rebuilding Tool) webserver (http://haddock.science.uu.nl/dna/dna.php) which makes use of the 3DNA programme (Dijk and Bonvin, 2009). The webserver gives output as PDB files which are optimised for use with the HADDOCK (High Ambiguity Driven protein-protein DOCKing) webserver (See Section 3.6.2).

3.7.2 Molecular docking

Macro-molecular docking uses computer algorithms to predict contact between two molecules by starting with the properties of the free unbound molecules. The sequences constructed using 3D-DART (see Section 3.6.1) were docked with chain J of the FOXP2 FHD PDB file (2A07) using the HADDOCK (High Ambiguity Driven protein-protein DOCKing) webserver (www.haddock.science.uu.nl/services/HADDOCK/haddock.php). Docking was done using the Easy Interface which allows users to input the two molecules of choice and decides on the best parameters. The residue involved in the KE mutation A553 was chosen as an active or involved residue within the protein.

3.8 Structural alignment

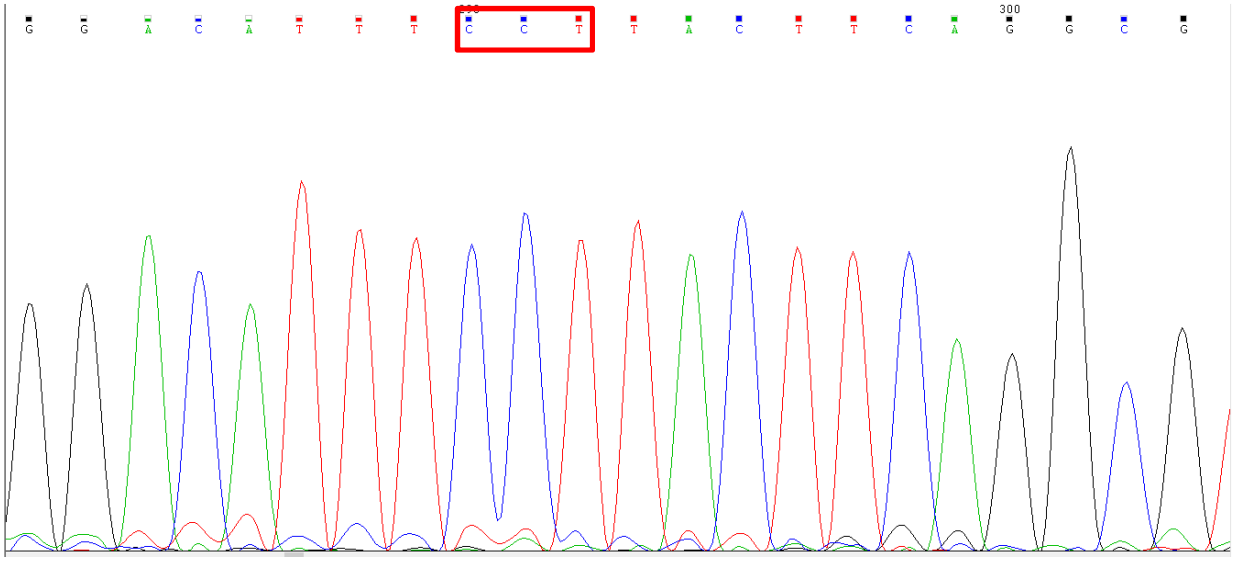
In order to predict any changes to the structure of the FOXP2 FHD upon binding to the various DNA sequences studied, the MatchMaker extension of Chimera (Meng *et al*., 2006) was used to align the structures generated in Section 3.6.2 and the resulting alignment was viewed using the Multalign extension of the programme. The global structural alignment was performed using the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) with the BLOSUM-62 weight matrix (Henikoff *et al*., 1994) with 30% weighting on the secondary structure and 70% on the residue similarity.

4. Results

4.1 Construction of the FOXP2 forkhead domain A539P mutant

The FOXP2 FHD monomeric mutant (A539P) was successfully generated by site-directed mutagenesis and was confirmed by a commercial Sanger sequencing service at Inqaba Biotec, Pretoria (Figure 4.1). Translation of the insert sequence confirmed that Ala539 was replaced by a proline.

A



B

```
A539P        IVRPPFTYATLIRQAIMESSDRQLTLNEIYSWFTRTFPYFRRNAATWKNAVRHNLSLHKC  358
             IVRPPFTYATLIRQAIMESSDRQLTLNEIYSWFTRTF YFRRNAATWKNAVRHNLSLHKC
Wild type    IVRPPFTYATLIRQAIMESSDRQLTLNEIYSWFTRTFAYFRRNAATWKNAVRHNLSLHKC  60

A539P        FVRVENVKGAVWTVDEVEYQKRRSQ  433
             FVRVENVKGAVWTVDEVEYQKRRSQ
Wild type    FVRVENVKGAVWTVDEVEYQKRRSQ  85
```

Figure 4.1 Sequence of the FOXP2 forkhead domain A539P mutant insert in pET-30. Confirmation of the construction of pET-30 containing the FOXP FHD A539P mutant insert by Sanger sequencing. The mutant was constructed by site directed mutagenesis to replace a guanine with a cytosine in the insert. This mutated the GCT codon on the reverse strand which codes for alanine to the CCT codon (shown in red) which codes for proline on the reverse. A: chromatogram of sequencing results. B: Alignment indicating the translated sequence obtained following site-directed mutagenesis. The Alignment was performed using the Basic Local Alignment Search Tool (Altschul *et al.*, 1997).

4.2 Overexpression and purification of the FOXP2 forkhead domain and the A539P mutant

Induction trials determined that optimum expression of the wild type FOXP2 FHD was achieved 4 hours post-induction with 1 mM IPTG (Figure 4.2). This was expected as Stroud *et al*. reported overexpression conditions of 3-5 hours with 1 mM IPTG induction (Stroud *et al*., 2006). There was a marked decrease in the amount of protein present in whole cell lysate from inductions longer than 6 hours indicating that after this time the cells actively began breaking down the protein. This could be due to improper folding of the protein at high concentrations leading to break down of the improperly folded proteins (Baneyx, 1999). Overexpression conditions of 4 hours growth post-induction with 1 mM IPTG were subsequently used for expression of both the FOXP2 FHD and the FOXP2 FHD A539P mutant. Purification of both proteins was successfully achieved with IMAC (Figure 4.3). The proteins were determined to have >90% purity using densitometry. The proteins were both estimated to be just over 15 kDa which is in agreement with the predicted size of 14.9 kDa, based on the amino acid sequence of the tagged FOXP2 FHD. The identity of the proteins was confirmed using mass spectroscopy (Figure 4.4). To ensure that no DNA contamination was present in the protein samples, A_{280} to A_{260} ratios were calculated and samples with ratios of greater than 1.5 were deemed to be free from DNA. This was important as the FOXP2 FHD is capable of nonspecific DNA binding in *E. coli*. This was prevented by using high salt concentration (0.5 M NaCl) in all purification buffers. No DNase was used during purification in case it caused interference in subsequent DNA binding experiments.

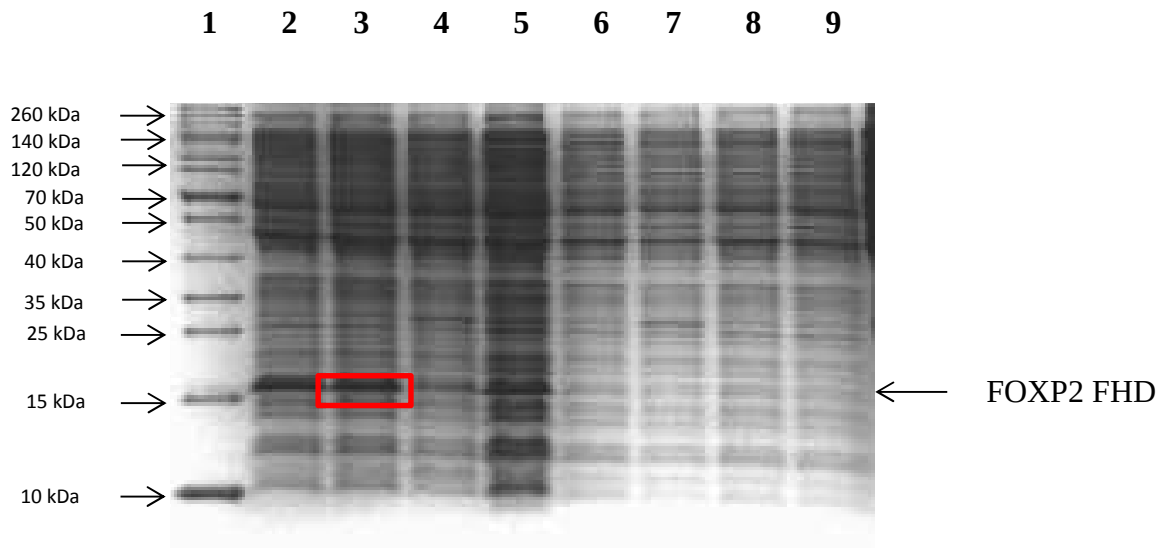


Figure 4.2 Expression trials of the wild type FOXP2 forkhead domain. Electrophoresis of whole cell extract from T7 Express cells transformed with pET-30 containing the FOXP2 FHD insert and induced with 1 M IPTG. The expression condition used in subsequent work is shown by a red box. Lane 1: Thermo Spectra Broad Range Protein Ladder; Lane 2: 2 hour induction; Lane 3: 4 hours induction; Lane 4: 6 hours induction; Lane 5: 8 hours induction; Lane 6: 10 hours induction; Lane 7: 12 hours induction; Lane 8: 14 hours induction; Lane 9: 16 hours induction. Proteins are stained with Comassie Brilliant Blue.

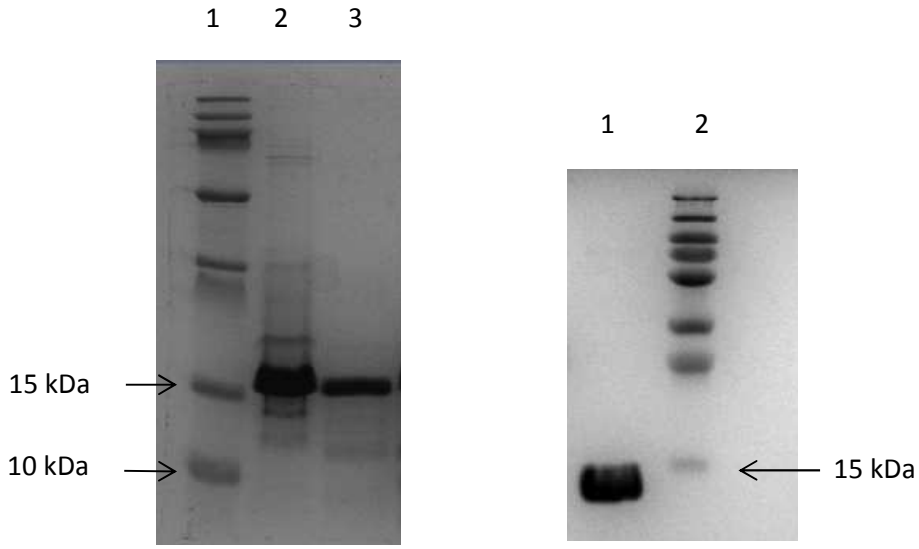
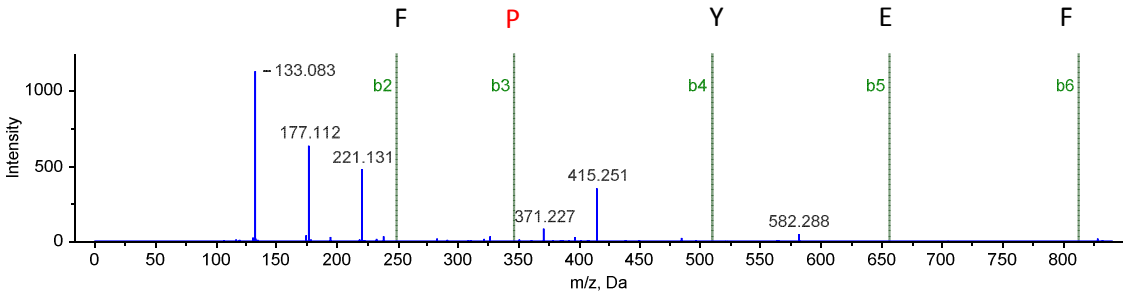


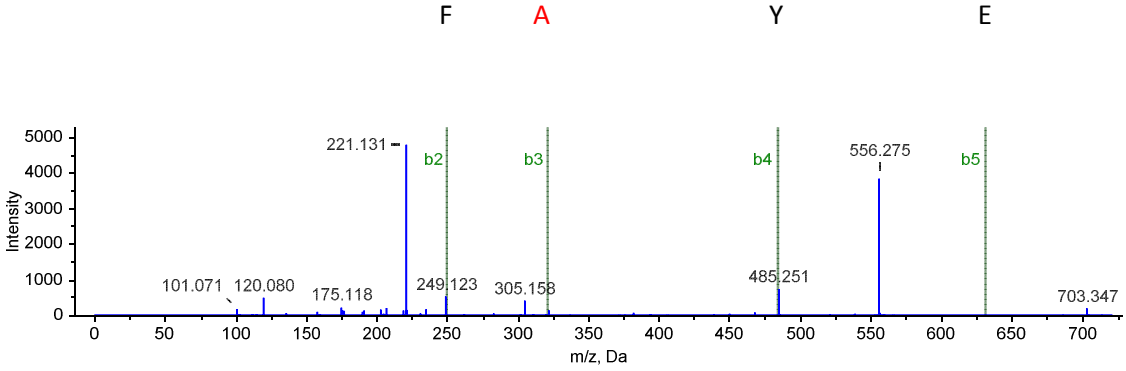
Figure 4.3 Purification of the FOXP2 forkhead domain and the FOXP2 forkhead domain A539P mutant. Electrophoresis of protein purified using immobilised metal affinity chromatography. Left: Lane 1: Thermo Spectra Low Range Protein Marker; Lane 2: soluble fraction of T7 Express cells expressing the FOXP2 FHD; Lane 3: Purified FOXP2 FHD. Right: Lane 1: Purified FOXP2 FHD A539 mutant; Lane 2: Thermo Spectra Low range Protein Marker. Proteins are stained with Comassie Brilliant Blue.

4.3 Confirmation of the identity of the wild type FOXP2 FHD and the A539P mutant by mass spectroscopy

The protein sequence of the wild type FOXP2 and the A539P mutant was used to confirm the identity of the two proteins. This was done by standard protein sequencing using mass spectroscopy. After trypsin digest, LC MS/MS was performed on proteins obtained from purification as a commercial service at the CSIR, Pretoria. Full sequence coverage was obtained for both the wild type FOXP2 FHD and the A539P. The A539P mutation was confirmed (Figure 4.4). Confidence levels of >99.9% were obtained for these sequences.



The FOXP2 forkhead domain A539P mutant



The wild type FOXP2 forkhead domain

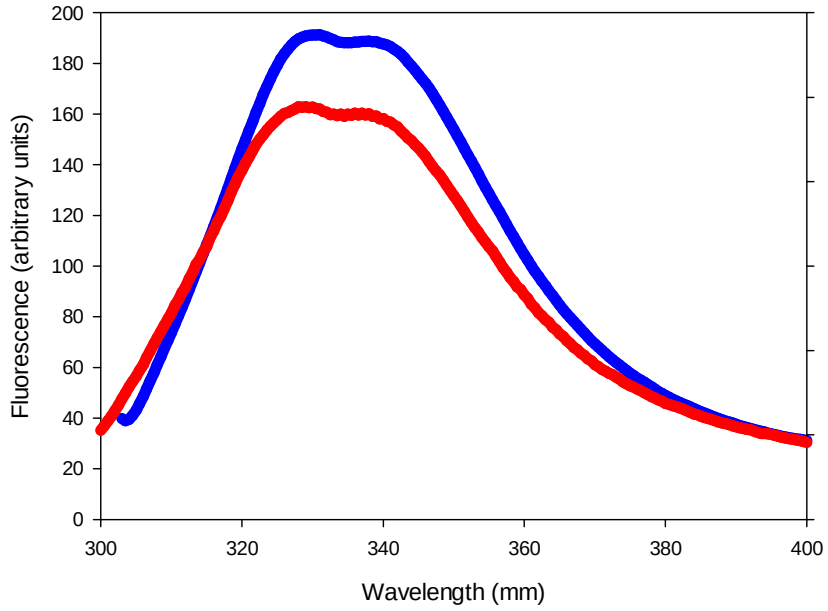
Figure 4.4 Mass spectra of the FOXP2 forkhead domain and the FOXP2 Forkhead domain A539P mutant. Sequence of the FOXP2 FHD and the FOXP2 FHD A539P as determined by liquid chromatography-mass spectrometry/mass spectrometry and visualised using the Protein Pilot software package. The alanine in wild type and proline in the A539P mutant are shown in red.

4.4 Confirmation of the structural integrity of the FOXP2 forkhead domain

It was important to assess whether the A539P mutation caused structural changes to the FOXP2 FHD other than preventing domain swapping as these changes would confound later data. In order to determine whether the A539P mutation caused folding abnormalities when compared with the wild type FOXP2 FHD, intrinsic fluorescence was used. This method was favoured as both proteins contain 3 tryptophans which are located in different helices of the protein. Thus intrinsic fluorescence gives information about the local environment in three separate areas of the proteins making it a good indicator of global structural integrity. Both the wild type and A539P mutant of the FOXP2 FHD display similar fluorescence spectra with an emission maximum of approximately 340 nm (Figure 4.5 A). This emission wavelength is in agreement with the location of the tryptophans within the protein which are neither fully buried nor exposed (Figure 4.5 B). The three tryptophans in the FOXP2 FHD are located on two separate helices and a strand. Ala539 is located between the two tryptophan containing helices. The decrease in the intensity of fluorescence of the mutant by approximately 15% when compared to the wild type protein is a reflection of a change to the immediate environment of one or more of the tryptophan side chains, for instance the increase in proximity to residues which quench tryptophan fluorescence (Callis and Liu, 2004; Chen and Barkley, 1998). Many residues are known to quench tryptophan fluorescence, namely: lysine, tyrosine, glutamine, asparagine, aspartic and glutamic acids, histidine and cysteine (Chen and Barkley, 1998). This change in sidechain packing bears no relation to the global structure

of the protein; this can be seen by the lack of wavelength shift between the two proteins. The maintenance of global structural integrity in the FOXP2 FHD A539P mutant allowed it to be used in subsequent work as a suitable model for the monomeric wild type FOXP2 FHD.

A



B

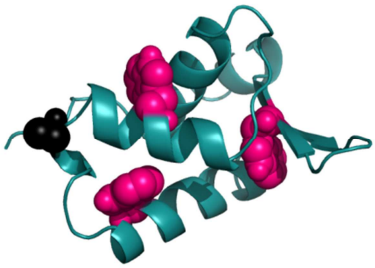


Figure 4.5 Intrinsic Fluorescence of the FOXP2 forkhead domain and the FOXP2 forkhead domain A539P mutant. A**:** Fluorescence spectra of the FOXP2 FHD (2 µM; blue) and the FOXP2 FHD A539P mutant (2 µM; red) excited at 295 nm. Both the FOXP2 FHD and the A539P mutant show emission maxima at 330 nm. Measurements were made in 20 mM Tris; 150 mM NaCl at 20 °C. B: Structure of the FOXP2 FHD. The three tryptophans used to assess global structure via fluorescence are shown in magenta while Ala539 is shown in black. PDB 2A07 visulalised using PyMOL (The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.).

4.5 Determination of the oligomeric state of the FOXP2 forkhead domain

Size exclusion chromatography confirmed the oligomeric state of both the wild type and A539P mutant of the FOXP2 FHD to be predominantly monomeric (Figure 4.6). The estimated size of the monomer according to a standard curve constructed from size standards was approximately 14 kDa. There was also small peak in the wild type protein corresponding to the dimer which measured at approximately 25 kDa. This indicates that predominantly monomer and a small proportion of dimer were present in the wild type FOXP2 FHD while the A539P mutant was exclusively monomeric. The concentration of protein used in SEC was 40 µM, indicating that at concentrations of 40 µM or lower, the A539P FOXP2 FHD is exclusively monomeric. Since most subsequent binding studies were done within this concentration range, A539P was confirmed as a suitable model for the monomeric wild type FOXP2 FHD.

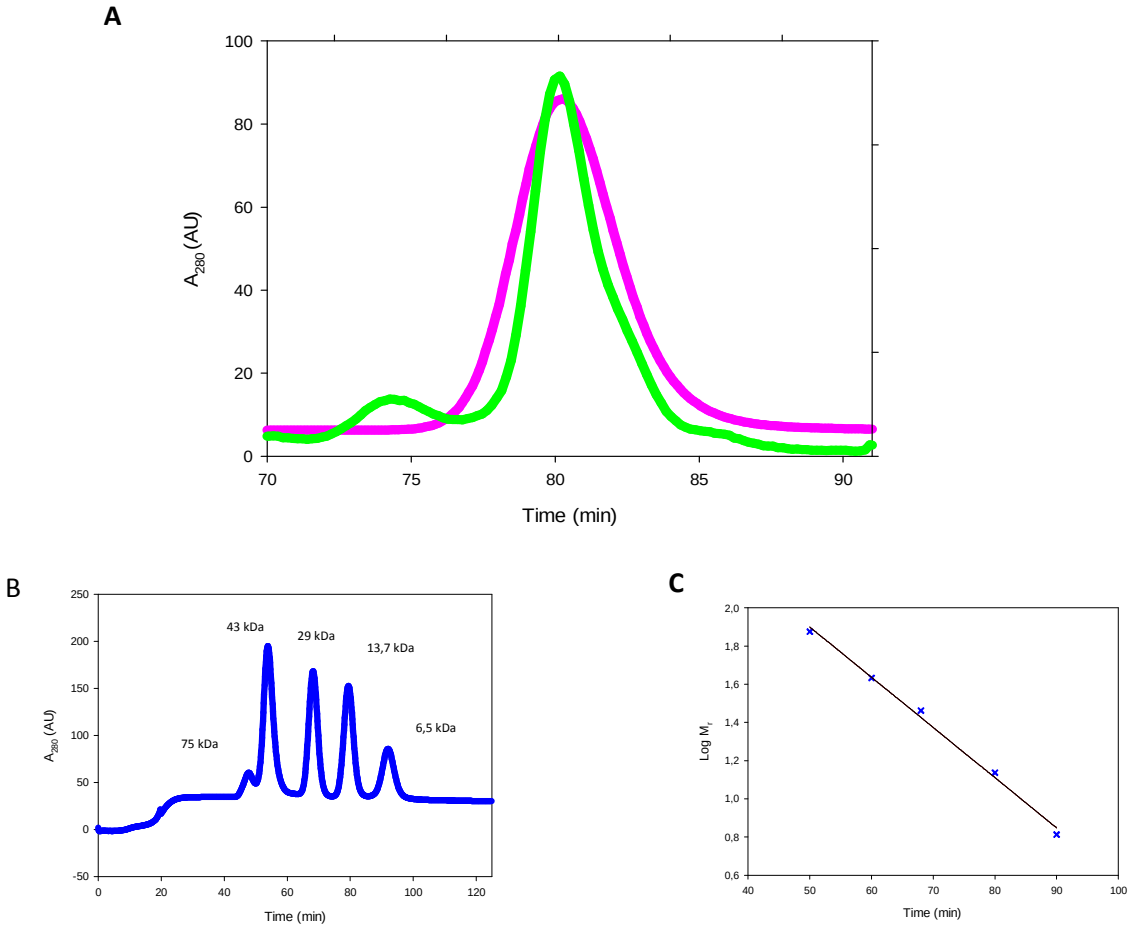


Figure 4.6 Confirmation of the oligomeric state of the wild type FOXP2 forkhead domain and the A539P mutant. A: The FOXP2 FHD (40 µM; green) and the FOXP2 FHD A539P mutant (40 µM; purple) eluted from a GE Healthcare Hiload™ 16/600 75 pg size-exclusion column B: Elution profile of the size exclusion chromatography standards. GE Healthcare Low Molecular Weight Gel Filtration Calibration standards from first eluted: conalbumin (75 kDa); ovalbumin (43 kDa); carbonic anhydrase (29 kDa); ribonuclease A (13,7 kDa); Aprotinin (6,5 kDa). C: Standard curve obtained from calibration standards used for sizing of the FOXP2 FHD and the A539P mutant. Linear fit is shown in blue.. From this curve it was estimated that the monomer was approximately 14 kDa and the dimer was approximately 25 kDa.

4.6 Identification of novel FOXP2 forkhead domain binding motifs

Because of the relatively large number of reported cognate DNA sequences, an investigation of potential binding sites using systematic evolution of ligands by exponential enrichment was undertaken. This particular method of identifying target motifs had not previously been undertaken for the FOXP2 FHD. Systematic evolution of ligands by exponential enrichment (SELEX) is a method of *in vitro* identification of nucleic acid binding partners of protein targets. The binding material can be single or double stranded DNA or RNA. In the case of transcription factor binding, double stranded DNA is the molecule of interest. The general procedure of SELEX consists of the binding of a randomised DNA library with the target protein and the recovery and amplification of bound DNA which is then used as the starting DNA for the next round of selection (Tuerk and Gold, 1990). A single round form of SELEX, known as MonoLEX, was put forward by Nitsche *et al*. and has been successfully used as a more cost efficient method of SELEX (Nitsche *et al*., 2007). MonoLEX requires extremely efficient recovery of bound and unbound DNA as there is only a single round to remove nonspecific binders, thus the method uses affinity chromatography rather than nitrocellulose membranes as its recovery method (Nitsche *et al*., 2007). SELEX and MonoLEX were chosen over other methods of binding motif identification such as chromosomal imunopercipitation (ChIP) and microarray based technologies because of the simplicity and cost effectiveness of setting up the technique. In addition, a number of FOXP2 ChIP studies have been performed (Vernes *et al*., 2007; Spiteri *et al*., 2007) but no studies using SELEX and this particular protein have been published.

Multiple round SELEX led to enrichment of the DNA pool until round 4 (Figure 4.7). Round 5 led to a substantial loss of DNA. Three attempts at further enrichment at round 5 were unsuccessful. Thus, round 4 DNA was sequenced. NGS provides distinct advantages over traditional Sanger sequencing. The most important for its use in motif identification is the ability to identify thousands of unique sequences from a mixed sequence sample in a single sequencing reaction. Sanger sequencing produces indiscernible mixed sequence if the sample contains more than one sequence. This limitation can only be overcome in Sanger sequencing by cloning individual sequences through PCR amplification and AT cloning to obtain samples of plasmid containing an insert of each of the sequences. This is incredibly time consuming and costly. NGS is able to obtain thousands of unique sequences because sequencing takes place on a solid surface rather than in solution. This is accomplished on a chip with either beads or glass micro-channels. Each bead or channel has only one molecule of DNA bound which is then amplified to produce spatially isolated pools of unique sequences across the chip. Each of these pools produces a discernible signal upon sequencing (Mardis, 2013).

The three NGS platforms readily accessible in South Africa are the Roche 454, the Illumina and the Ion Torrent. These systems can be compared in terms of cost, error rate and number of reads per run (sequence coverage). The Illumina is the most expensive platform, closely followed by the Roche 454, with the Ion Torrent being significantly more cost effective; however both the Illumina and Roche 454 platforms have much lower error rates (<0.5%) and higher sequence coverage (millions of reads) than the Ion Torrent which has a 1% error rate and only

produces less than 100 000 reads (Mardis, 2013; Metzker, 2010). The Ion Torrent platform was chosen as it is cost effective and only moderate sequence coverage was required because of the high redundancy of the sequence sets caused by amplification of recovered DNA.

MonoLEX DNA from the single round was also sent for sequencing. Ion Torrent sequencing returned 270 000 and 230 000 sequences for the two experiments respectively. Both sequence sets showed roughly the same sequence variation with approximately 20 000 unique sequences per set. The high redundancy of the datasets was due to initial amplification of the library by PCR.

After cleaning of the datasets using Galaxy Suite (Blankenberg *et al.*, 2010) the top nine motifs which were present in both datasets were identified using DREME and AME (Table 4.1). These sequences were varied in length and base composition. The length of motifs was constrained within DREME and AME to six to ten base pairs. None of the motifs identified matched published FOXP2 binding motifs or those of any of the FOX families. This was expected as Nakagawa in 2013 suggested that the binding specificities of the FOX proteins extended beyond the canonical FOX binding sequence. The nine motifs generated could not be considered FOXP2 binding motifs until binding of the FOXP2 FHD and these motifs were empirically confirmed.

These nine sequences were screened for binding using surface plasmon resonance. Surface plasmon resonance (SPR) is a spectrometric technique which measures biomolecular interactions in real time allowing affinity constants and the kinetic parameters of binding to be established.

Table 4.1 DNA motifs identified by systematic evolution of ligands by exponential enrichment

Sequence number	Sequence	E-value*	Consensus Logo**
1	**CGTATAWG**	**1.9e-1153**	
2	**GACTCATC**	**9.3e-982**	
3	**CCCGATAG**	**4.1e-921**	
4	**ACCAAGC**	**7.9e-646**	
5	**CCATCTTA**	**3.1e-461**	
6	**GGTCGA**	**2.3e-307**	
7	**ATGGGG**	**4.3e-073**	
8	**CATATA**	**8.5e-072**	
9	**ANTGAGTC**	**1.7e-059**	

*The E-value of the match of a sequence in a database to a group of motifs is defined as the expected number of sequences in a random database of the same size that would match as well as the sequence.

**The logos use standard consensus logo representation in which the height of each base represents the information contained at that position of the sequence.

The two sequences that were found to bind to the FOXP2 FHD are boxed in red

IUPAC code is used where N represents any base and W represents A or T.

Plasmons are electromagnetic waves which occur at the interface between a metal and gas or liquid. When light is passed through a prism below the metal surface some light energy is absorbed by the plasmons causing them to resonate. This causes a dip in the reflected light intensity at that particular angle known as the SPR angle. Any changes at the interface between the metal and liquid or gas cause a change to the SPR angle; thus allowing any adsorption to or desorption from the surface to be detected (Rich and Myszka, 2000). The measurement of the shift in the SPR angle when analyte is flowed through narrow channels or flow cells over a chip with a thin metal film which has ligand bound to it is the basic principle of automated SPR systems.

SPR is an attractive tool for the measurement of biomolecular interactions as it is a direct form of measurement without labelling and uses very small quantities of ligand and analyte. Commercially available systems such as the BIACore used in this study make use of chips which are functionalised by the manufacturer allowing easy coupling of the ligand.

Screening was conducted by analysis of the response units generated by a single injection of 50 µM FOXP2 FHD over flow cells containing each DNA sequence when compared to injection over a flow cell without DNA. Two sequences were found to bind the FOXP2 forkhead domain, Sequence 3 and Sequence 9 (Figure 4.8), as seen by an increase in response units after injection of protein. However, when kinetic analysis was attempted, only Sequence 3 was found to give satisfactory binding (data not shown). This is likely to be due to nonspecific binding of sequence 9 to the carboxymethyl groups on the chip during initial screening which was not reproducible under the more stringent conditions used

for kinetic analysis. This sequence (Sequence 3) was named the Webb sequence and was used for further binding analysis in comparison with other published FOXP2 binding sequences shown in Figure 3.1.

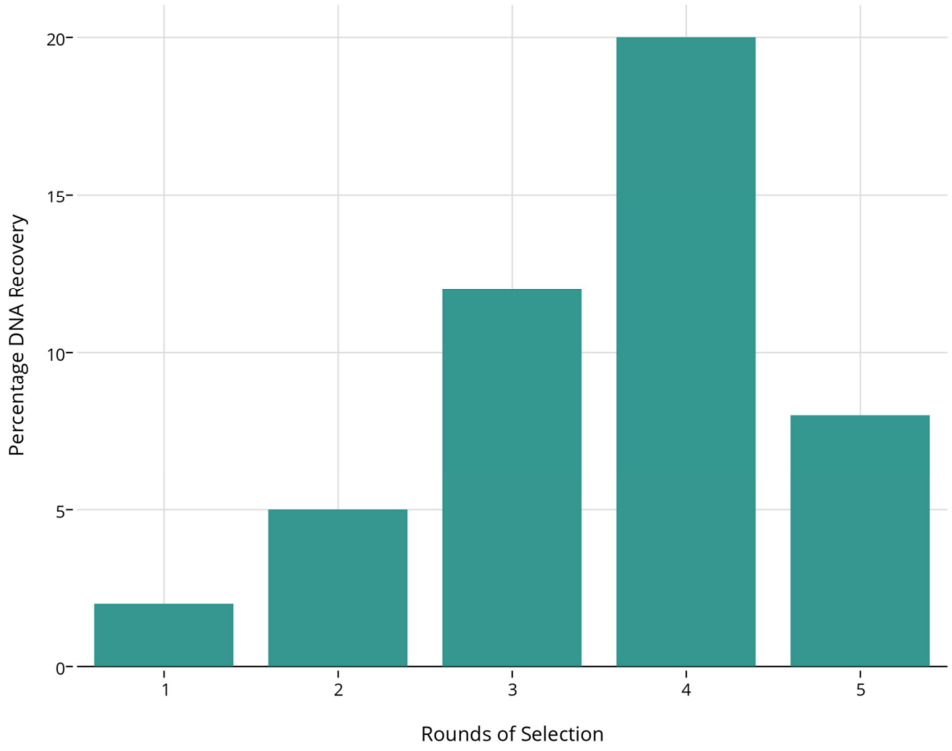


Figure 4.7 Percentage enrichment of systematic evolution of ligands by exponential enrichment rounds. Amount of DNA recovered from each round of SELEX expressed as a percentage of the total amount of library initially mixed with the FOXP2 FHD. Amount of DNA recovered was calculated using absorbance at 260 nm of samples after phenol:chloroform extraction and isopropanol precipitation. Round 4 was chosen for subsequent sequencing and motif analysis.

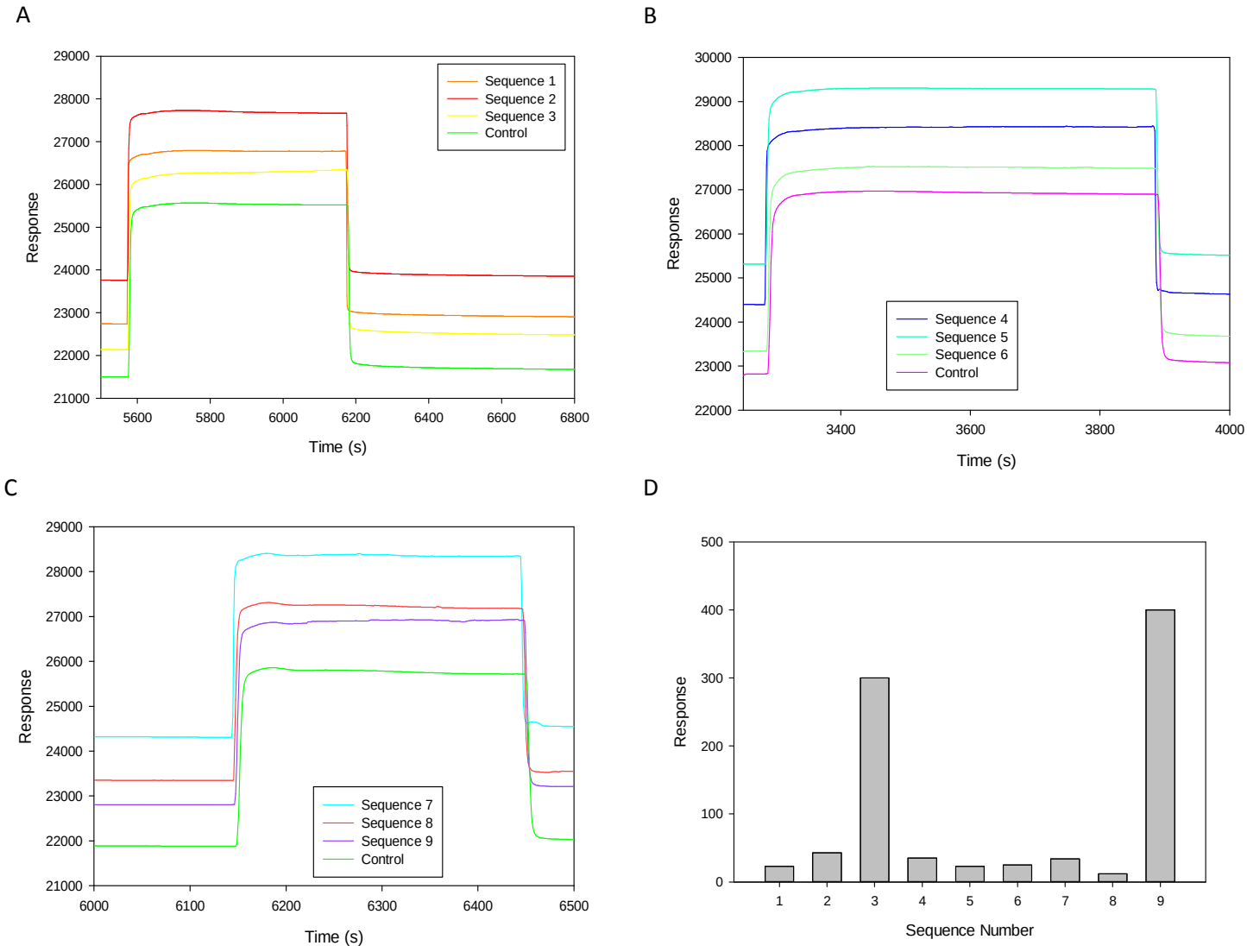


Figure 4.8 Screening of identified DNA motifs for FOXP2 forkhead domain binding. Level of binding as measured by surface plasmon resonance of the FOXP2 forkhead domain and DNA sequences identified as potential FOXP2 FHD binding sequences. A – C: Sensograms of binding for sequences 1-9, where an increase in response units after injection of 50 µM FOXP2 FHD indicates binding. Control flow cells did not contain DNA D: The binding values as determined by the difference in response before and after binding. Sequences 3 and 9 showed binding to the FOXP2 FHD.

4.7 Rates and affinities of the FOXP2 forkhead domain and various DNA sequences

In order to compare the types of binding among the different DNA sequences listed in Figure 3.1 which FOXP2 is able to bind, kinetic analysis was done using SPR. This type of analysis establishes dissociation constants (K_d) as well as association (k_1 or k_a) and dissociation rates (k_{-1} or k_d) from multiple injections of protein at a range of concentrations over immobilised DNA. In a reversible DNA-protein binding interaction

$$\mathrm{DNA} + \mathrm{Protein} \underset{k_d}{\overset{k_a}{\rightleftharpoons}} \mathrm{DNA} \cdot \mathrm{Protein},$$

The association rate is the rate at which the protein binds to the DNA and the dissociation rate is the rate at which the protein becomes unbound from the DNA. The dissociation constant is the ratio of the dissociation rate to the association rate and is used as a measure of affinity where the smaller the dissociation constant, the greater the binding. In SPR, the rates from a 1:1 binding reaction are calculated using the Langmuir equation which was derived based on adsorption of a gas across a solid surface which accurately describes the adsorption of an analyte in liquid phase to a solid surface in SPR (Myszka, 1997).

Table 4.2: DNA sequences used to determine the kinetic parameters of the binding of the FOXP2 forkhead domain and DNA

Name	Method	Source
Wang	Cyclic amplification and selection of targets (CAST)	Wang *et al*., 2003
Nelson	Microfluidic micro array	Nelson *et al*., 2013
Enard	Gene expression studies	Enard *et al*., 2009
Zhu	Known FOX binding site	Zhu *et al*., 2009
Webb	SELEX	This work

The sequences used in this study were obtained by a number of techniques (Table 3.1) SELEX, CAST, and microfluidic chip analysis are all *in vitro* techniques which check binding against a randomised synthetic library. The motifs acquired from these techniques are relatively high affinity binding motifs, due to multiple washing steps throughout the procedures. However, these motifs cannot be mapped back to the genome and are not necessarily representative of binding under cellular conditions (Wang *et al*., 2011). ChIP on the other hand is an *in vitro* technique which captures all DNA sequences to which the protein is bound at the time of crosslinking. For this reason a relatively large number of DNA sequences are obtained, many of which are bound transiently and non-specifically. For this and other reasons ChIP data is not always reproducible (Wang *et al*., 2011). Other factors include uniformity of cross-linking, specificity of the antibody which

detects the protein, efficiency of chromatin immunoprecipitation and biological variability (Peng *et al.*, 2010).

Data obtained from this kinetic analysis by SPR fitted well to the Langmuir 1:1 model of binding (Figures 4.9). All fits performed resulted in chi squared values of <10% of the maximum signal which indicates reasonable fits. The affinities and rates obtained varied across the five sequences (Enard, Nelson, Zhu, Webb and Wang; Table 4.2). These sequences were investigated as they are published binding motifs of the FOXP2 FHD. A randomised control sequence showed very little binding and it was not possible to fit data to any binding model. This indicates that the consistent binding of the sequences tested was specific as no consistent binding was seen with an arbitrary sequence.

The association rates of the sequences were similar with the exception of Zhu which has a much faster association rate than the Enard, Nelson and Webb sequences, and the Wang sequence which has a much slower rate than the other sequences (Figure 4.10). The dissociation rates of the Enard, Zhu, Wang and Webb sequences showed no significant variation while the rate of the Nelson sequence was slower. The Enard sequence displayed the lowest binding affinity measured by the dissociation constant followed by the Wang sequence while the Nelson, Zhu and Webb sequences had dissociation constants within error of one another.

This means that the FOXP2 FHD binds to the Enard sequence at a moderate rate and dissociates quickly spending little time bound and thus binding with low affinity. The Wang sequence is bound by the FOXP2 FHD at a relatively slow

rate but dissociates at a moderate rate leading to intermediate affinity binding. The Zhu, Webb and Nelson sequences show relatively high affinity binding to the FOXP2 FHD. In the case of the Zhu sequence this is due to a fast association rate coupled to a moderate dissociation rate and in Nelson this is due to a moderate association rate with a slow dissociation rate.

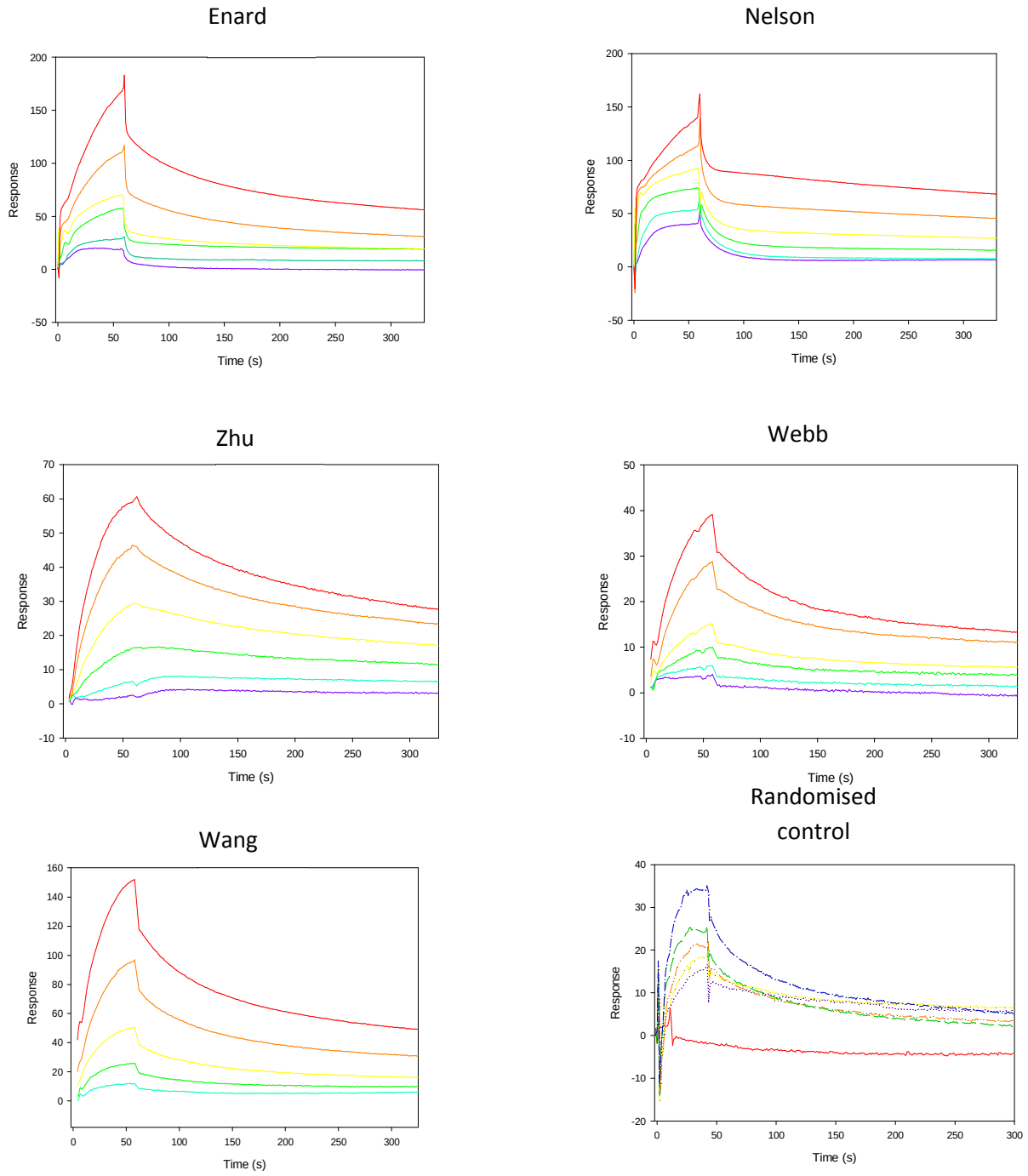


Figure 4.9 The kinetics of binding of the FOXP2 forkhead domain A539P mutant and various DNA sequences. Surface plasmon resonance analysis of the binding of the FOXP2 forkhead domain and the various DNA sequences. Injections of 9 µM (blue); 18 µM (cyan); 36 µM (green); 72 µM (yellow); 144 µM (orange) and 288 µM (red) of the FOXP2 FHD were flowed over immobilised DNA. These sensorgrams represent the average of three replicates and have been zeroed against a flow cell containing no DNA. The A539P mutant was used so as to prevent contaminating signal from dimerisation.

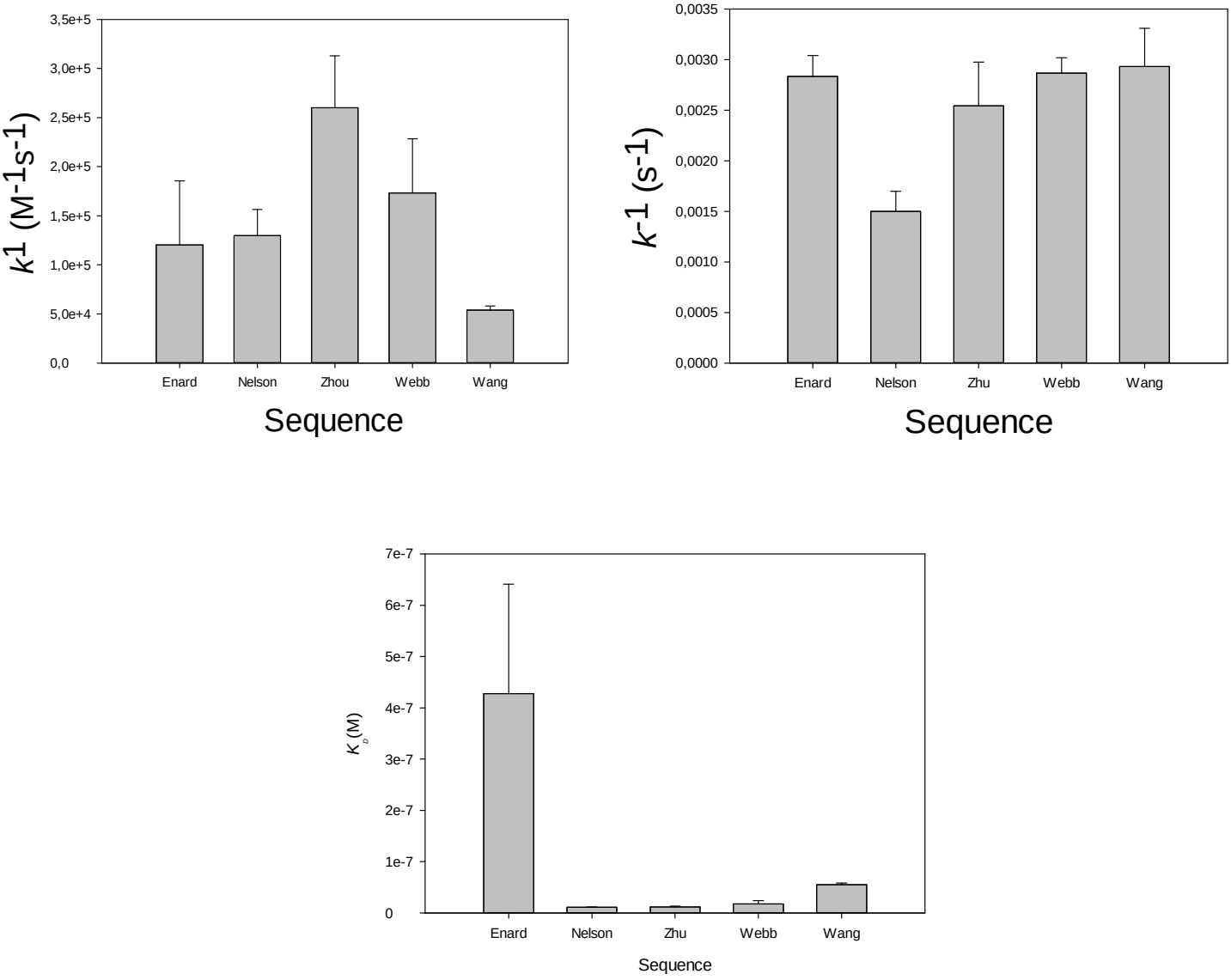


Figure 4.10 Rates and affinities of the FOXP2 forkhead domain and various DNA sequences. Comparison of the association rates (top left), dissociation rates (top right) and dissociation constants (bottom) of the FOXP2 FHD and the Enard, Nelson, Zhu, Wang and Webb DNA sequences as obtained by surface plasmon resonance utilising a Langmuir 1:1 model of binding.

Table 4.3 Affinities and rates of the FOXP2 forkhead domain and various DNA sequences

Sequence	k_1 ($M^{-1}s^{-1}$ x 10^5)	k_{-1} (ms^{-1})	K_d (nM)
Enard TATTTAT	1.5 ± 0.2	2.5 ± 0.2	438 ± 107
Nelson CATTTGT	1.3 ± 0.1	1.5 ± 0.2	11.5 ± 0.5
Zhu CGCAG	2.6 ± 0.4	2.5 ± 0.5	10.3 ± 2
Webb GGGCTATC	1.7 ± 0.5	2.9 ± 0.2	18 ± 2
Wang GTTTAA	0.54 ± 0.04	2.9 ± 0.3	55 ± 3

4.8 Isothermal titration calorimetry

Isothermal titration calorimetry (ITC) measures the amount of energy required to maintain a constant temperature difference of 0 °C between a cell containing the reaction under study and a reference cell. Initially the reaction cell contains only one of the binding partners, which then has the second binding partner added to it in a stepwise fashion by a syringe. ITC directly measures the energy absorbed or released by the reaction being studied and provides thermodynamic parameters, binding affinity and stoichiometry (Leavitt and Freire, 2001). ITC has become the gold standard in characterising biomolecular interactions because it is highly accurate and requires no chemical (such as labelling) or physical (such as immobilisation on a surface) modification of either binding partner (Doyle, 1997).

From the thermogram generated from an ITC experiment, the binding affinity (K_d) and the enthalpy of binding (ΔH) are calculated while ΔS and ΔG are calculated

from Equation 3, where R is the gas constant; T is the temperature of the reaction; ΔG is the Gibbs free energy of binding and ΔS is the entropy of binding.

$$-RT ln K_a = \Delta G = \Delta H - T\Delta S \quad (3)$$

ITC was performed in order to confirm the 1:1 binding that was seen using SPR and to obtain thermodynamic parameters of binding. The binding isotherm obtained from ITC was lacking data points in the presaturation region of the curve and thus produced an incomplete sigmoid (Figure 4.11). The data was found to have a good fit with the sigmoidal equation:

$$1{,}2356\ \mathrm{erfc}[1{,}69619\ (0{,}381124 - x)] - 2.21186$$

Where ΔH is given by the difference between asympototes or

$$2(1{,}2356) \approx 2{,}5 \text{ kcal/ mol}$$

K_d is given by the slope or

$$\frac{2(1{,}2356 \times 1{,}69619)}{\sqrt{\pi}} = 423 \text{ nM}$$

And

ΔG = -RT lnK = -0,5 kcal/ mol

Since

ΔG = ΔH –TΔS

Therefore

ΔS= 0.01 kcal/ mol/ K

Because of the incomplete nature of the sigmoid obtained from the thermogram it is unlikely that these values are reliable, this can be seen by the large discrepancy between the K_d obtained for the binding of the FOXP2 FHD A539P mutant and the Nelson DNA sequence from this experiment and the one obtained from SPR (423 nM and 11.5 nM respectively). However, general trends from the data can be seen. The data fit a 1:1 binding model. This validates the use of the Langmuir model used to fit SPR data. The unfavourable enthalpy is compensated for by favourable entropy. Entropically driven reactions in DNA-protein interactions are indicative of distortions to the DNA backbone upon binding (Jen-Jacobson *et al*., 2000). The distortion of the DNA backbone in this experiment may be caused by the lack of Mg_2Cl in buffers used. Mg_2Cl typically stabilises the DNA backbone within the cell (Hartwig, 2001).

Although exact thermodynamic parameters were not generated from the ITC data, it is important to note that this is the first thermodynamic study to be conducted on the FOXP2 FHD and provides a very important starting point from which to further optimise this type of experiment. Optimisation of buffers as well as an increase in the concentration of DNA in the sample cell may help to obtain the required data in the presaturation region of the curve.

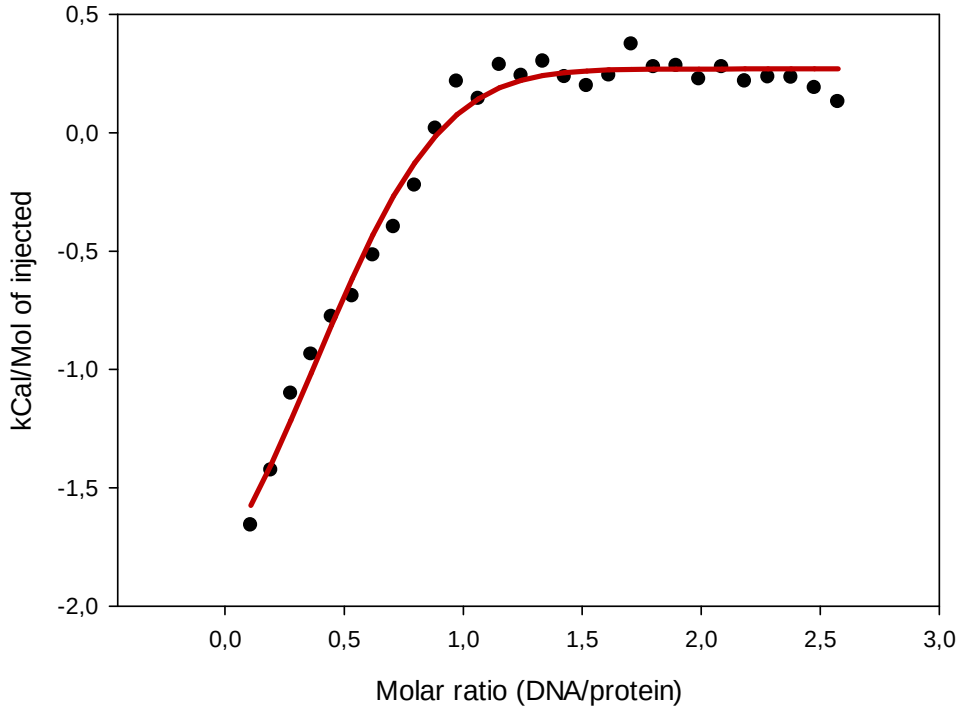


Figure 4.11.: Binding isotherm of the FOXP2 FHD A539P with the Nelson DNA sequence. An incomplete sigmoid was obtained when 150 µM FOXP2 FHD A539P mutant was injected into 5 µM Nelson DNA sequence. The data (black dots) was fitted with the equation $1{,}2356\,\mathrm{erfc}[1{,}69619\,(0{,}381124 - x)] - 2.21186$ (red line). Because of the incomplete nature of the sigmoid, it was not possible to calculate reliable thermodynamic parameters from the data.

4.9 *In silico* prediction of the bonds formed between the FOXP2 forkhead domain and various DNA sequences

In order to provide a possible explanation of the varied rates and affinities of binding between the FOXP2 FHD and the various DNA sequences studied, molecular docking was performed. Macro-molecular docking uses computer algorithms to predict contact between two molecules by starting with the properties of the free unbound molecules. The sequences constructed using 3D-DART (see Section 3.6.1) were docked with chain J of the FOXP2 FHD PDB file (2A07) using the HADDOCK (High Ambiguity Driven protein-protein DOCKing) webserver:

www.haddock.science.uu.nl/services/HADDOCK/haddock.php

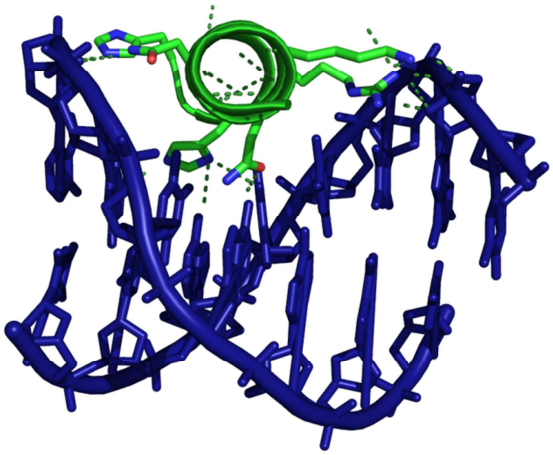
which allows conformational changes to both molecules during complex formation. While many docking programs only take into account the theoretical properties of the starting molecules, HADDOCK is data driven and therefore takes into account experimental data. HADDOCK makes use of Ambiguous Interaction Restraints (AIR) to drive docking (de Vries *et al.*, 2010). The user defines active residues known to be involved in the interaction which the programme then uses to define AIR.

Arg553 of the FOXP2 FHD was defined as an active residue to be used as a starting point for docking. The reason why this residue was chosen is that experimentally it is known that mutation of this residue causes disruption to binding as it is the cause of Speech-language Disorder 1 (Lai *et al.*, 2001) . The disruption of the interaction between the FOXP2 FHD and DNA by the R553H

mutation is caused by a drastic perturbation in the electrostatic potential disrupting charge complementarity between the protein and the DNA backbone (Banerjee-Basu and Baxevanis, 2004).

The models used were chosen according to the best HADDOCK scores, which measure the quality of the model. In order to confirm that HADDOCK could accurately predict structures of DNA-protein complexes, the model generated for the Wang DNA sequence was aligned to the crystal structure of the FOXP2 FHD in complex with this sequence and it was found that the bonds were the same.

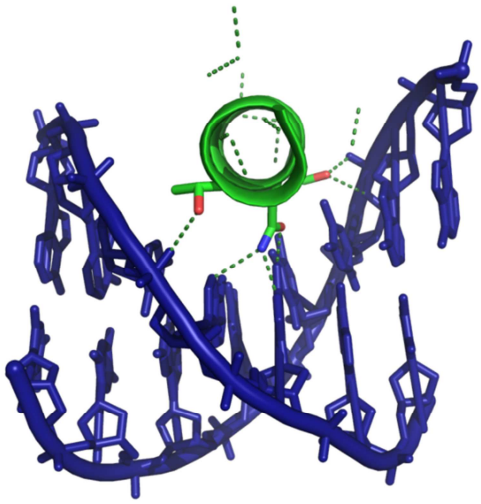
The docking of the Nelson, Zhu, Webb and Wang sequences were successful (Figures 4.12), however, the model obtained for the Enard sequence was deemed to be unsatisfactory as it contained hydrogen bonds of < 1.5 Å, as calculated using PyMOL, which is not possible due to the size of the atoms involved in the bonds and no further analysis was conducted on the interaction between the FOXP2 FHD and the Enard sequence. The reason for this failure is not clear, however, HADDOCK generates models by first docking the two molecules as rigid bodies and then, once the best model has been found, continues refinement of the model with flexibility in both molecules. The structure of the FOXP2 FHD used in modelling was that of the crystal structure in which the protein is bound to the Wang sequence. It is possible that the FOXP2 FHD binds to Enard in a relaxed state or in a different conformation which HADDOCK was unable to predict.



Nelson Sequence

Base Interaction		**Backbone Interaction**	
Residue	*Base*	*Residue*	*Base*
Asn550	A33	Lys549	G9
His554	T12 A13 G31	Arg553	T10 G9
		Asn555	A30
		His559	A30

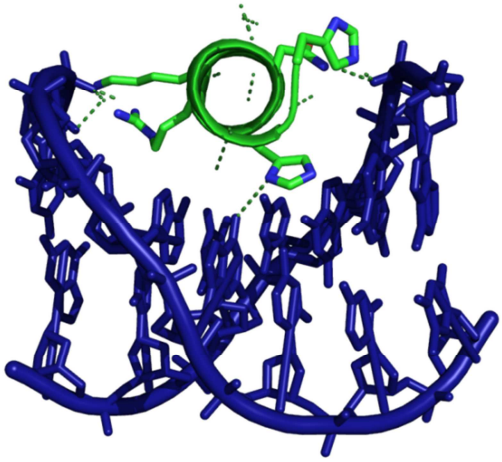
Figure 4.12 In silico prediction of the interaction between the FOXP2 forkhead domain and the Nelson DNA sequence. Molecular docking of the FOXP2 FHD and the Nelson DNA sequence modelled using the HADDOCK webserver (de Vries *et al*., 2010) and visualised using PyMOL (The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.).



Wang Sequence

Base Interaction		**Backbone Interaction**	
Residue	*Base*	*Residue*	*Base*
Asn550	A13 A13 A12	Thr547	A12
		Ser557	T10

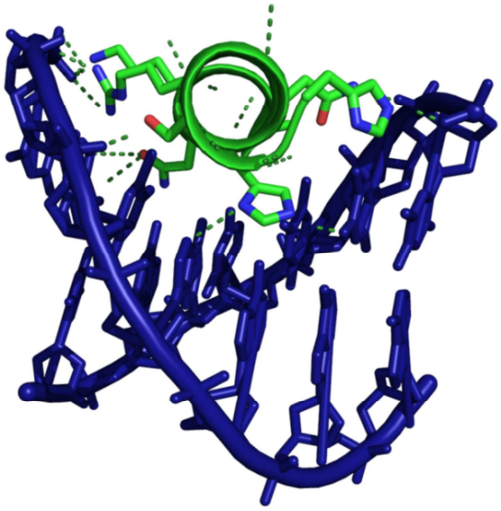
Figure 4.13 In silico prediction of the interaction between the FOXP2 forkhead domain and the Wang DNA sequence. Molecular docking of the FOXP2 FHD and the Wang DNA sequence modelled using the HADDOCK webserver (de Vries *et al*., 2010) and visualised using PyMOL (The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.).



Webb Sequence

Base Interactions		**Backbone Interactions**	
Residue	*Base*	*Residue*	*Base*
His554	G8	Lys549	C4 C4
		Arg553	C5 C5
		Asn555	G31
		His559	G31

Figure 4.14 In silico prediction of the interaction between the FOXP2 forkhead domain and the Webb sequence. Molecular docking of the FOXP2 FHD and the Webb DNA sequence modelled using the HADDOCK webserver (de Vries *et al.*, 2010) and visualised using PyMOL (The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.).



Zhu Sequence

Base Interactions		**Backbone Interactions**	
Residue	*Base*	*Residue*	*Base*
Asn550	A7 A7	Lys549	T6 T6
His554	G9 T36	Arg553	T6 T6
		Asn555	T36
		Ser557	T8
		His559	T35

Figure 4.15 In silico prediction of the interaction between the FOXP2 forkhead domain and the Zhu sequence. Molecular docking of the FOXP2 FHD and the Zhu DNA sequence modelled using the HADDOCK webserver (de Vries *et al*., 2010) and visualised using PyMOL (The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.).

All modelled sequences showed hydrogen bonding with both the phosphate backbone of the DNA as well as with nitrogenous bases of the DNA (Table 4.4). In all four models only residues Asn550 and His554 were involved in base specific interactions. The model of the Wang sequence showed the fewest number of predicted base-specific hydrogen bonds with only one residue of the protein interacting with specific bases of the DNA. In the Wang model only Asn550 is predicted to be involved in sequence specific hydrogen bonding while in the Webb model only His554 is involved. In the Nelson and Zhu sequences interactions are predicted with both Asn550 and His554.

In terms of backbone interactions most sequences make more backbone contacts are made than base specific interactions and the backbone interactions involve Lys549, Arg553, Asn555 and His559. Only the Wang sequence deviates from both these trends. This sequence shows more base specific than backbone interactions and these backbone interactions are with Thr547 and Ser557. The Zhu sequence also has a backbone interaction with Ser557. Serine and threonine are typical amino acids involved in DNA-binding (Luscombe, 2001). The vast majority of backbone binding in all sequences is predicted to occur at A and T bases with the exception of the Webb sequence which has backbone interactions at only G and C bases.

Table 4.4 Hydrogen bonds formed between the FOXP2 forkhead domain and various DNA sequences as predicted by molecular modelling

Sequence	Base Interaction		Backbone Interaction	
Nelson	*Residue*	*Base*	*Residue*	*Base*
	Asn550	A33	Lys549	G9
	His554	T12	Arg553	T10
		A13		G9
		G31	Asn555	A30
			His559	A30
	Total 4		**Total 5**	
Wang	*Residue*	*Base*	*Residue*	*Base*
	Asn550	A13	Thr547	A12
		A13	Ser557	T10
		A12		
	Total 3		**Total 2**	
Webb	*Residue*	*Base*	*Residue*	*Base*
	His554	G8	Lys549	C4
				C4
			Arg553	C5
				C5
			Asn555	G31
			His559	G31
	Total 1		**Total 6**	
Zhu	*Residue*	*Base*	*Residue*	*Base*
	Asn550	A7	Lys549	T6
		A7		T6
	His554	G9	Arg553	T6
		T36		T6
			Asn555	T36
			Ser557	T8
			His559	T35
	Total 4		**Total 7**	

An attempt to find a structural basis of the varied rates and affinities was made by looking for trends between the bonds predicted from molecular docking and the dissociation rates, association rates and the dissociation constants calculated for the binding of the FOXP2 FHD and the various DNA sequences.

A relationship was observed between the number of predicted backbone interactions and the association rate (Figure 4.16A) with the sequences with the greatest number of predicted backbone interactions having the fastest association rates. It was considered whether the slowest dissociation rates corresponded with the greatest number of predicted base specific interactions but this was not the case. However, it does appear that the dissociation rate is slower the greater the number of predicted base specific interactions between His554 and DNA (Figure 4.16B). It was also observed that the greater the total number of predicted interactions the greater the binding affinity (Figure 4.16C)

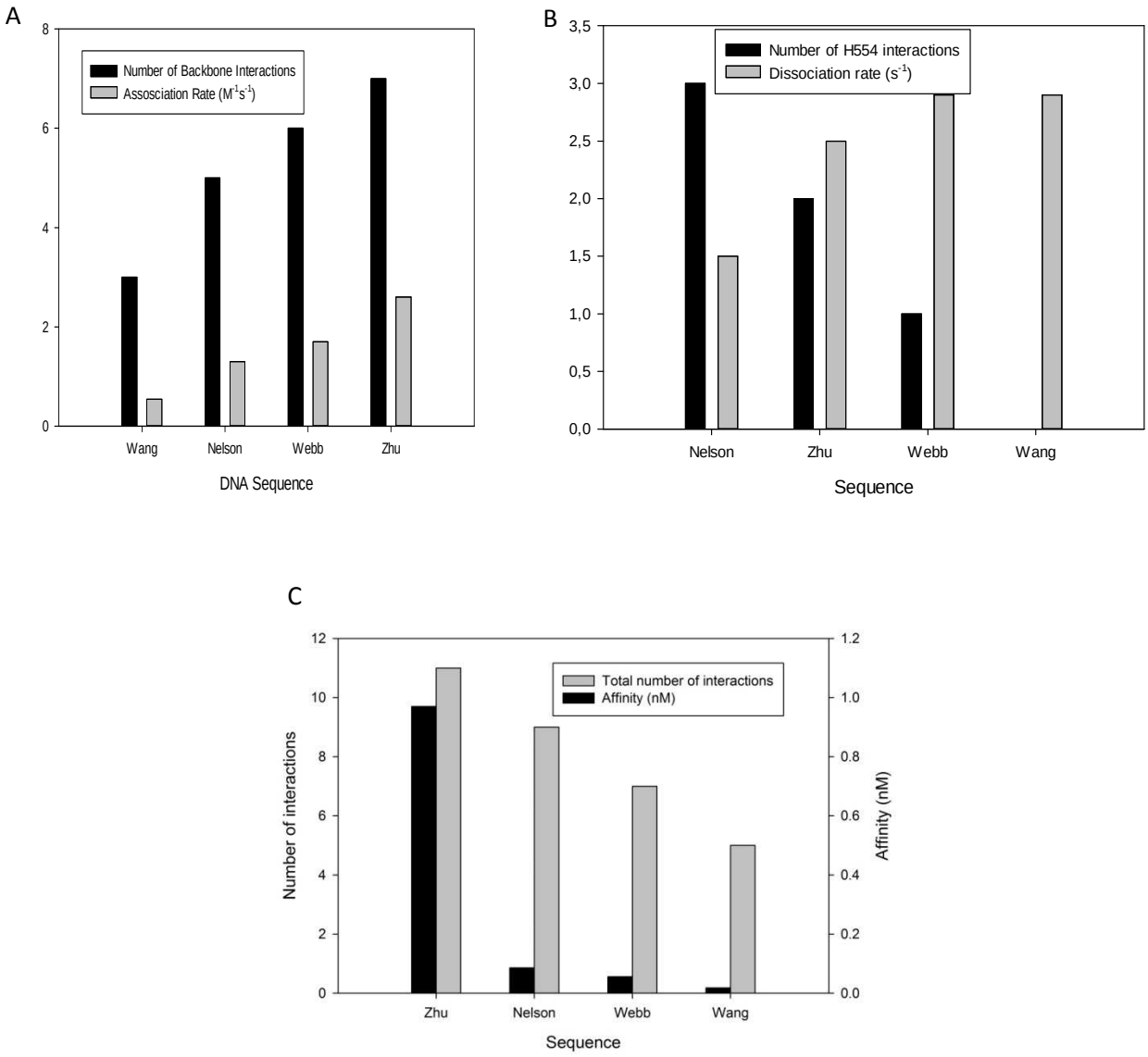


Figure 4.16 The relationship between rates and affinities of various DNA sequences and the FOXP2 forkhead domain and the number of backbone and base interactions predicted. A) The association rates of binding are related to the number of backbone interactions predicted. B) There is a weak relationship between the dissociation rates of binding and the number of base interactions predicted with H554. C) The affinity of binding shows a relationship to the total (backbone and base) interactions predicted.

4.10 Structural alignment of the predicted structures of the FOXP2 forkhead domain and various DNA sequences

In order to predict structural changes to the protein upon binding to different DNA sequences, the backbone of the models generated in HADDOCK were aligned using Chimera. No major structural changes to the protein were observed in the models of the protein bound to any of the DNA sequences (Figure 4.17). The root-mean square deviation (RMSD) of this alignment was calculated to be 0.73 Å. This indicates that structural rearrangement of the protein is not required in order for binding to different DNA sequences. This is applicable to the sequences which were successfully modelled (Nelson, Wang, Webb and Zhu) while no conclusions about the structural basis of the Enard sequence can be drawn owing to the lack of an adequate model.

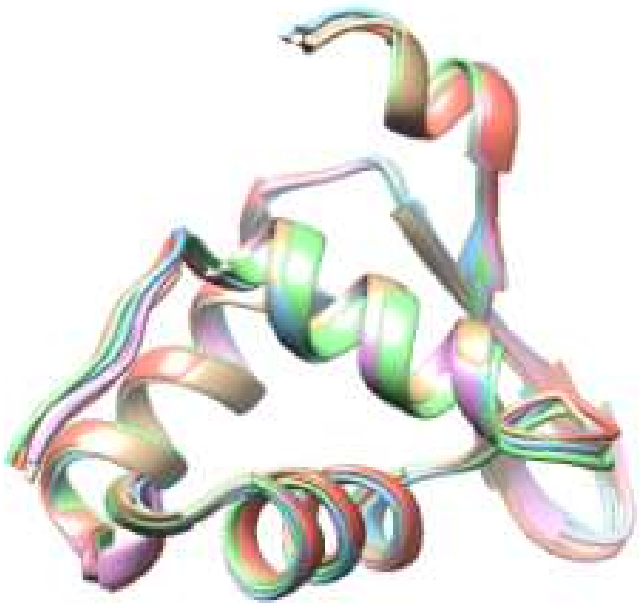


Figure 4.17 Structural alignment of the backbone of the FOXP2 forkhead domain bound to various DNA sequences. Models generated using HADDOCK of the FOXP2 FHD bound to the Nelson (pink), Wang (green), Webb (blue) and Zhu (orange) DNA sequences. It can be seen that there are no major structural changes to the protein upon binding to different sequences. Alignment generated and viewed using Chimera (Meng *et al*., 2006). The RMSD for the alignment was calculated to be 0.73 Å.

5 Discussion

5.1 The FOXP2 FHD can bind a variety of distinct sequences

Using SPR it was shown that the FOXP2 FHD is capable of binding a number of unrelated sequences (Figure 4.9). These sequences range from a canonical FOX consensus sequence (Wang) to a novel sequence identified in this work which has never been reported as a FOX binding motif (Webb). This finding indicates that FOXP2 does not recognise a single consensus sequence but rather is bound to numerous sequences within the genome.

It is unsurprising that the FOXP2 FHD is able to recognise and bind a variety of distinct sequences as numerous studies indicate that many transcription factors recognise more than one sequence (Badis *et al.*, 2009; Mirny et al., 2009; Todeschini *et al*., 2014). Chromatin immuno precipitation (ChIP) studies of numerous transcription factors yield thousands of binding sequences rather than a single consensus sequence (Todeschini *et al*., 2014). A study of 100 mouse transcription factors found that nearly half of transcription factors investigated were able to recognise multiple distinct DNA sequences (Badis *et al*., 2009). Furthermore, Wunderlich and Mirny in their 2009 study of nearly 1 000 transcription factors found that eukaryotic transcription factors occupy sites with a variety of sequence motifs (Mirny *et al*., 2009).

It is necessary for transcription factors to recognise a number of sequences because DNA binding proteins are very rarely free in solution within the cell but are almost always in contact with DNA (Lin and Riggs, 1975; Kao-Huang and Revzin, 1977; Phair *et al*., 2004; Elf *et al*., 2007). This continuous contact gives

rise to three different kinds of protein-DNA interactions: specific functional, specific non-functional and nonspecific non-functional (Todeschini *et al*., 2014). In this context, functional binding is defined as binding which gives rise to gene regulation. It is likely that the sequences investigated in this study give rise to both specific functional and specific non-functional binding.

Although non-functional binding does not directly give rise to gene regulation it may be extremely important for cellular function by titrating the concentration of transcription factors available to bind specific functional targets (Todeschini *et al*., 2014). This is supported by the empirical evidence that when concentrations of transcription factors are either halved or doubled, transcription of target genes is changed by an order much smaller than two (Veitia *et al*., 2013) and in bacteria, a recent study using the well-studied *Lac1* repressor competitive binding sites, led to a titration of occupied functional binding sites (Brewster *et al*., 2014). This study showed that concentrations of the transcription factor were affected by competition between varied sites and the relative affinity of the protein for these sites. Controlling transcription factor concentration in this manner minimises the dosage effect of genes allowing whole genome duplications to be tolerated without affecting the transcription levels within the cells (Schnable *et al*., 2011).

5.2 The FOXP2 FHD binds distinct sequences with varied affinity and rates

From kinetic analysis by SPR it was shown that the FOXP2 FHD binds each of the sequences studied with unique rates and affinities (Table 4.3).This means that the FOXP2 FHD is bound to each of the sequences for different time periods ranging from relatively short for the Enard sequence to longer periods for the

Webb, Zhu and Nelson sequences, the Wang sequence shows intermediate binding rates.

In order to explain the structural basis of this variation, models of the FOXP2 FHD bound to each of the sequences were created (Figure 4.12-15). From these models the number of hydrogen bonds formed between the FOXP2 FHD and each sequence were predicted (Table 4.4). There appears to be a relationship between the association rates of the FOXP2 FHD and various DNA sequences and the number of backbone interactions predicted in each interaction (Figure 4.16). With the exception of the Wang sequence, most of the backbone interactions are with the positively charged arginine and lysine residues (Table 4.4). The Wang sequence also has the slowest association rate with the FOXP2 FHD, highlighting the importance of electrostatic interactions in the association rate.

Electrostatic interactions are often linked to shape readout of the DNA minor groove by the protein. This is probably not the case in the FOXP2 FHD as no interactions with the minor groove were observed in the models constructed (Figure 4.12-15). Stroud *et al*., 2006 noted that the interaction of the FOXP2 FHD differed from that of other FOX proteins. The wings of other FOXs make extensive contact with the minor groove. However, in FOXP2 these contacts are not possible because Wing 1 is truncated to a simple turn and Wing 2 is a short helix (Stroud *et al*., 2006).

The relationship between the association of rate of the FOXP2 FHD binding to DNA and backbone interactions explains why the disease causing mutation, R553H, completely disrupts DNA binding even though that particular residue is

not directly involved in hydrogen bonding with DNA as seen in the crystal structure of the FOXP2 FHD in complex with the Wang DNA. This is further illustrated by the predicted model with the Zhu sequence, in addition to which this residue makes only backbone interactions in the predicted models with the Nelson and Webb sequences.. Banerjee-Basu and Baxervanis found in 2004 that the R553H mutation changes the electrostatic potential of the surface of the protein. A disruption to the electrostatics of the protein would prevent interactions between the protein and the backbone of the DNA. This would decrease the association rate of binding effectively preventing interaction despite base specific interactions remaining unaffected. The majority of interactions between transcription factors and DNA are between positively charged residues of the protein and the negatively charged phosphodiester backbone of the DNA (Luscombe and Austin, 2000). In addition this electrostatic interaction is necessary for facilitated diffusion (Hippel and Berg, 1986). The importance of the electrostatics of the FOXP2 FHD is evidenced by the predicted models of its interaction with DNA. In all of the models more protein residues are involved in interactions with the backbone than with the nitrogenous bases of the DNA (Table 4.4).

The dissociation rates of the FOXP2 FHD appear to have a relationship to the number of base specific bonds formed between His554 and the DNA (Figure 4.16). The Nelson sequence, which shows the slowest dissociation, has three bonds predicted with His554 while the Wang sequence, which has the fastest dissociation rate, is not predicted to form any bonds with His554. Further experimentation involving the mutational analysis outside the scope of this project is required to elucidate the exact role of His554 in FOXP2 FHD DNA-binding but

from the available evidence, this residue plays a key role in hydrogen bonding to DNA. This residue also makes conserved interactions in DNA-binding in the structures of numerous other FOX FHDs including FOXA1 (Lai and Clark, 1993), FOXK1 (Tsai *et al*., 2006) and FOXO1 (Tsai *et al*., 2007).

The total number of interactions (backbone and base) formed between the protein and the DNA is relative to the association constant (Figure 4.16). This means that the more interactions predicted between the protein and the DNA the higher the affinity of binding. This is logical as each interaction contributes to the affinity of binding.

It is important to note that although the sequences investigated bound the FOXP2 FHD with unique affinities which indicates varied strength of binding for different sequences, all of the interactions between the FOXP2 FHD and the sequences investigated showed specificity. Specificity in a DNA-protein interaction refers to the ability of the protein to distinguish specific sequences from the background sequence of the genome (Rohs, 2010). This specificity is evidenced by the lack of binding between the FOXP2 FHD and a randomised control (Figure 4.9).

5.3 The FOXP2 FHD is monomeric in solution

It was shown by SEC of the FOXP2 FHD at a concentration of 40 µM that the FOXP2 FHD domain is monomeric when in solution at concentrations higher than those required for DNA-binding according to the K_ds obtained from SPR which are in the nanomolar range. While it has been established that the leucine zipper is required for dimerisation it is unclear whether these dimers have domain swapped FHDs. Given the findings presented here as well as the length of the flexible region between the leucine zipper and the FHD of FOXP2 it at first seems unlikely that FOXP2 acts as a domain swapped dimer. The region between the FHD and the leucine zipper of FOXP2 is approximately 100 amino acids and is predicted to be unstructured. Thus dimerisation at the leucine zipper would not bring the FHDs of the individual monomers into the close contact required in order for domain swapping to take place.

However, there is sound reasoning to support the physiological role of FOXP2 domain swapped dimers. Mutations in the hinge loop region of the FOXP3 FHD which are thought to prevent domain swapping have been linked to IPEX (Bandukwala *et al*., 2011). This means that domain swapping in the FOXP3 FHD, which shares 74% protein sequence identity with FOXP2, plays a crucial physiological role (Chen *et al*., 2015).

From an evolutionary stand point it is advantageous for the DNA binding domains of transcription factors to act as homo- and hetero- dimers. Dimeric DNA binding domains have much longer consensus sequences and effectively increase specificity. This is of particular importance in large families of highly

homologous transcription factors, such as the FOX family, where many members of different classes recognise highly similar consensus sequences (Georges *et al*., 2010). In the FOXP family both hetero- and homo- dimers are known to form and perform a physiological function within the family (Li *et al*., 2004; Sin *et al*., 2014).

Cooperative binding provides a possible explanation of how domain swapped dimers may be the physiologically active form of FOXP2 A possible cause of cooperative binding which leads to domain swapped dimers in the case of the FOXP2 FHD may be partial monomer unfolding in "fly casting" during DNA binding. Fly casting is a hypothesized mechanism of molecular recognition which has been experimentally shown to be relevant in protein-DNA interactions. In this mechanism the protein initially interacts with the DNA in a partially unfolded state by nonspecific interactions while searching for the correct recognition sequence. Upon specific binding with the recognition sequence the protein reaches its folded form. Partially unfolded monomers of other proteins have been shown to form dimers upon DNA binding (Kohler and Metallo, 1999; Rentzeperis *et al*., 1993).

Because a partially unfolded monomer is required for domain swapping, dimerisation upon cooperative DNA binding of partially unfolded monomers searching for their recognition sequences by fly casting is a logical process which could explain why the FOXP2 FHD exists as a monomer in solution but is likely active as a dimer. Kohler and Metallor explain that it is favourable for dimers to form after DNA binding rather than in solution because preformed dimers searching for a specific sequence have slower dissociation rates from nonspecific

sequences than monomers do causing dimers to be kinetically hindered in locating specific sequence (Kohler and Metallor, 1993).

5.4 Proposed types of binding between the FOXP2 forkhead domain and DNA

Given the findings presented here it is possible to propose three separate types of DNA binding for the FOXP2 FHD. These are low affinity specific binding; moderate affinity specific binding and high affinity specific binding (Figure 5.2).

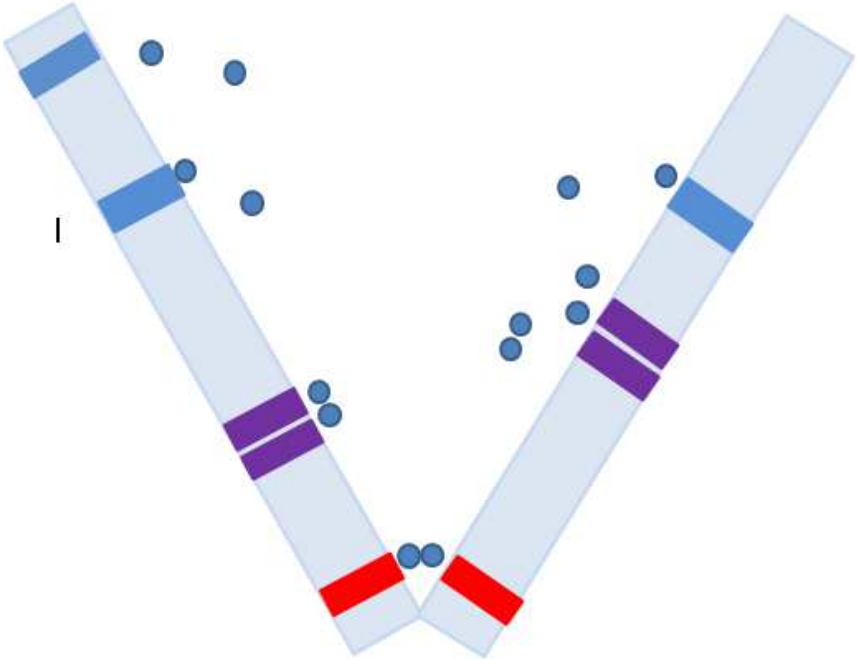


Figure 5.2 Proposed mechanism of the FOXP2 forkhead domain DNA-binding. In this putative model the FOXP2 FHD binds low affinity sites (blue) as monomer rapidly cycling on and off the DNA before binding to moderate affinity sites (purple) located close to the low affinity sites. At closely spaced moderate affinity sites dimerisation by domain swapping can occur and the dimer can dissociate to bind two high affinity sites (red) on separate DNA molecules bridging the two pieces of DNA.

During low affinity binding, the FHD probably binds to extremely low affinity sites on DNA at a rapid rate and scans nearby sequences in a partially unfolded state. This type of binding is probably non-functional in terms of gene regulation but assists in facilitated diffusion and allows the protein to find other moderate affinity sites. This type of binding is exemplified by the binding with the Enard sequence given by the rates and affinities obtained as well as the lack of a suitable model for this sequence with the folded FOXP2 FHD. Although this type of binding is low affinity, it is not entirely nonspecific still allowing control of spatial location of the protein within the nucleus according to the 3D location of the low affinity binding sites within the genome. Nonspecific binding, on the other hand would lead to a random distribution of the protein throughout the nucleus. These low affinity sites, as they are not predicted to have a functional consequence, could in theory bind anywhere in a gene and do not necessarily occur within promoters and would thus have been ignored in ChIP studies which identify possible target genes by looking specifically for proteins bound in promoter regions of genes.

In moderate affinity binding the FOXP2 FHD binds specific sites with moderate affinity. If moderate affinity binding sites of FOXP2 are highly enriched within a certain genomic region, the concentration of the protein on the DNA could be extremely high making it possible for monomers scanning the DNA to come into frequent contact and form domain swapped dimers. The concentration of FOXP2 FHD could further be increased within a specific genomic region by the presence of other more tightly bound DNA binding proteins restricting movement along the DNA. This type of binding is probably not functional in terms of transcription but

concentrates the protein on the DNA allowing a fully formed dimer to bind nearby high affinity sites. This type of binding is exemplified by the binding seen in this work with the Wang sequence. This is shown by the moderate affinity measured. The crystal structure of the FOXP2 FHD in complex with the Wang sequence (Figure 1.9) supports the idea that the FOXP2 FHD binds the Wang sequence as a monomer but can dimerise and then dissociate. In this structure the monomer forms more bonds than the dimer indicating that the dimer will dissociate more rapidly.

In high affinity binding, each monomer of the domain swapped dimer could bind specific sites with high affinity and regulates transcription at these sites. The two FHDs of the dimer probably bind sites on two different DNA molecules as seen in the crystal structure of the FOXP3 domain swapped dimer in complex with DNA (Bandukwala *et al*., 2011). This type of binding is exemplified by the binding seen in this work with the Webb, Nelson and Zhu sequences. This is shown by the relatively high affinities of the FOXP2 FHD for these sequences (Table 4.2) when compared to the Wang and Enard sequences.

These types of binding agree with facilitated diffusion but in addition overcome the challenge of proteins becoming trapped at non-specific sites distant from the target sequence. Esadze *et al*. found that at physiological ionic concentrations, site search by facilitated diffusion was hampered by proteins becoming trapped at sites spatially removed from the target sequence significantly slowing the search process (Esadze *et al*., 2014). Trapping would not occur if high affinity target sequences were located within a short scanning distance from lower affinity less specific sites which draw the protein onto the DNA. This model would keep the

scanning length to the short stretches shown to be most efficient for locating specific sites (Halford, 2009).

These types of binding are also in agreement with the colocalisation mechanism of Mirny and colleagues which proposes that the efficiency of target search is dramatically reduced when the number of searches on non-specific DNA is limited (Mirny *et al*., 2009). In their scheme, which was based on bacterial transcription, where transcription and translation are coupled, the number of searches is limited by spatial organisation of the chromosome allowing transcription factors to be produced proximally to their targets. However, in eukaryotes this is not possible because of translation and transcription being uncoupled. The model presented here limits the number of searches by concentrating the FOXP2 FHD in limited space within the nucleus through clustering of low, moderate and high affinity binding sites. As the protein moves from low to moderate to high affinity binding sites within the genome, its location becomes more fixed because affinity is a direct measure of how mobile the protein is within the genome. Low affinity binding allows the protein to move relatively freely as the time for which it is bound to sequence is short. In contrast high affinity binding fixes the location of the protein because the protein spends a longer time bound to a specific sequence.

This model pays particular attention to the spatial location and concentration of the protein. In recent years the spatial organisation of the nucleus has become an important topic of research within biophysics. The spatial organisation of the nucleus is relevant to the study of genetic regulation because the nucleus is so crowded with macromolecules (proteins, DNA, transcription machinery, RNA

and ribosomes) that diffusion is not possible and because of extensive looping genetic distances between elements cannot be calculated linearly but need to take the 3D fold of the genome into account. For a review of these concepts see (Gorkin *et al*., 2014). Rao *et al*. (2014) have created a 3D map of the human genome within several cell lines and have shown that a very strict spatial compartmentalisation of the nucleus is present (Rao *et al*., 2014). This makes the regulation of the location of transcription factors, which are imported into the nucleus after translation in the cytosol, very important for genetic regulation.

Given that a large number of transcription factors exhibit cooperative dimerisation and bind more than one sequence it is possible that this mechanism of binding may be broadly applicable to other DNA binding proteins. Further experiments into DNA sliding and cooperative binding using single molecule fluorescence of the FOXP2 FHD as well as functional gene regulation (for instance promoter studies using reporter genes and analysis of published ChiP data for sequences thought to have functional consequences) are required in order to validate these types of DNA binding of the FOXP2 FHD.

5.5 Possible relevance of proposed mechanism to full length FOXP2

Although three modes of binding for the FOXP2 FHD have been suggested here based on the oligomeric state of the domain, its ability to bind a number of distinct DNA sequences and the affinities and rates of binding with these sequences, it cannot be forgotten that FOXP2 is a multidomain protein with other domains which are theoretically capable of DNA binding such as the leucine zipper and zinc finger. If these domains do bind DNA, the mechanism of DNA binding for

FOXP2 is likely to be far more complicated. Proteins with more than one type of DNA binding domain are capable of moving rapidly along DNA sequences (intersegmental transfer) by a mechanism known as bridging (Doucleff and Clore, 2008). During bridging, separate DNA binding domains are simultaneously bound to different cognate sequences on the same long stretch of DNA or on entirely separate strands which have been brought close together by looping. The two sequences are then held in close proximity to one another until the DNA binding domain with the lower affinity dissociates and goes in search of a new cognate sequence. This process can be repeated many times with each DNA binding domain dissociating at its own rate (Hippel and Berg, 1986). Interestingly the process of bringing DNA on separate chromosomes into contact has recently been highlighted by the finding that "gene kissing" of multiple DNA segments is required for the transcription of groups of genes which are co-regulated (Fanucchi *et al*., 2013). In support of this the domain swapped dimer of the FOXP3 has been shown to facilitate long range chromosomal movements required for the contact of co-regulated genes (Chen *et al*., 2015). However, only through the elucidation of binding mechanisms of each individual DNA binding domain, in a manner similar to the work completed here, will it be possible to understand the binding behaviour of the full multidomain protein.

6. Conclusion

From the work presented here it has been shown that the FOXP2 FHD is likely to bind DNA by more than one mechanism. There is evidence that the FOXP2 FHD can recognise a variety of DNA sequences, including a novel sequence (Webb). This motif has not previously been reported as a binding motif of the FOXP2

FHD or any other FOX protein and was identified in this work. Furthermore, evidence was presented that each of these DNA sequences binds the FOXP2 FHD with different rates and affinities. Electrostatic interactions between positively charged amino acids and the DNA backbone, as well as base-specific interactions between His554 and the DNA appear to be key in determining rates and affinities of binding interactions of the FOXP2 FHD and DNA. Based on these findings, three types of DNA-binding are proposed for the FOXP2 FHD. These types are: low affinity, non-functional binding; moderate affinity, non-functional binding and high affinity, functional binding. It is probable that each type of binding serves to control the spatial location of the protein within the nucleus, as well as the local concentration of protein. This work has shown that low affinity binding and varied sequence specificity may play a role in gene regulation by FOXP2. The proposed mechanism of binding for the FHD might have a future impact on the understanding of the binding and function of full length FOXP2.

7. References:

Alarcón, M., Abrahams, B.S., Stone, J.L., Duvall, J.A., Perederiy, J. V, Bomar, J.M., Sebat, J., Wigler, M., Martin, C.L., Ledbetter, D.H., et al. (2008). Linkage, association, and gene-expression analyses identify CNTNAP2 as an autism-susceptibility gene. Am. J. Hum. Genet. *82*, 150–159.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST : a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Aravind, L., and Anantharaman, V. (2005). The many faces of the helix-turn-helix domain: Transcription regulation and beyond⋆. FEMS Microbiol. *29*, 231–262.

Arnott, S., and Hukins, D. (1972). Optimised parameters for A-DNA and B-DNA. Biochem. Biophys. Res. Commun. *47*, 1505–1509.

Arnott, S., Bond, P., Selsing, E., and Smith, P. (1976). Models of triple-stranded polynucleotides with optimised stereochemistry. Nucleic Acids Res. *3*, 2459–2470.

Bacon, C., Schneider, M., Magueresse, C. Le, and Froehlich, H. (2014). Brain-specific Foxp1 deletion impairs neuronal development and causes autistic-like behaviour. Mol. Psychiatry *0049*, 1–8.

Badis, G., Berger, M., and Philippakis, A. (2009). Diversity and complexity in DNA recognition by transcription factors. Science (80-.). *324*, 1720–1723.

Bailey, T.L. (2011). DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics *27*, 1653–1659.

Bailey, T., and Gribskov, M. (1998). Combining evidence using p-values: application to sequence homology searches. Bioinformatics *14*, 1–5.

Ban, C., Ramakrishnan, B., and Sandaralingam, M. (1994). A single 2'=hydroxyl group converts B-DNA to A-DNA. J. Mol. Biol. *236*, 275–285.

Bandukwala, H.S., Wu, Y., Feuerer, M., Chen, Y., Barboza, B., Ghosh, S., Stroud, J.C., Benoist, C., Mathis, D., Rao, A., et al. (2011). Structure of a Domain-Swapped FOXP3 Dimer on DNA and Its Function in Regulatory T Cells. Immunity *34*, 479–491.

Banerjee-Basu, S., and Baxevanis, A.D. (2004). Structural analysis of disease-causing mutations in the P-subfamily of forkhead transcription factors. Proteins *54*, 639–647.

Baneyx, F. (1999). Recombinant protein expression in Escherichia coli. Curr. Opin. Biotechnol. *60*, 411–421.

Banham, A.H., Beasley, N., Campo, E., Gene, S., Fernandez, P.L., Fidler, C., Gatter, K., Jones, M., Mason, D.Y., Prime, J.E., et al. (2001). The FOXP1 Winged Helix Transcription

Factor Is a Novel Candidate Tumor Suppressor Gene on Chromosome 3p. Cancer Res. *61*, 8820–8829.

Benayoun, B. a, Caburet, S., and Veitia, R. a (2011). Forkhead transcription factors: key players in health and disease. Trends Genet. *27*, 224–232.

Bennett, M. (1994). Domain swapping: entangling alliances between proteins. Proc. … *91*, 3127–3131.

Bennett, C.L., Brunkow, M.E., Ramsdell, F., O'Briant, K.C., Zhu, Q., Fuleihan, R.L., Shigeoka, a O., Ochs, H.D., and Chance, P.F. (2001). A rare polyadenylation signal mutation of the FOXP3 gene (AAUAAA-->AAUGAA) leads to the IPEX syndrome. Immunogenetics *53*, 435–439.

Benos, P. V, Lapedes, A.S., and Stormo, G.D. (2002). Is there a code for protein-DNA recognition? Probab(ilistical)ly. Bioessays *24*, 466–475.

Berg, O., and Hippel, P. von (1988). Selection of DNA binding sites by regulatory proteins: II. The binding specificity of cyclic AMP receptor protein to recognition sites. J. Mol. Biol. *200*, 709–723.

Bettelli, E., Dastrange, M., and Oukka, M. (2005). Foxp3 interacts with nuclear factor of activated T cells and NF-kappa B to repress cytokine gene expression and effector functions of T helper cells. Proc. Natl. Acad. Sci. U. S. A. *102*, 5138–5143.

Biffi, G., and Tannahill, D. (2013). Quantitative visualization of DNA G-quadruplex structures in human cells. Nat. Chem. *5*, 182–186.

Billeter, M. (1997). Homeodomain-type DNA recognition. Prog. Biophys. Mol. Biol. *66*, 211–225.

Blankenberg, D., Gordon, A., Von Kuster, G., Coraor, N., Taylor, J., and Nekrutenko, A. (2010). Manipulation of FASTQ data with Galaxy. Bioinformatics *26*, 1783–1785.

Bonkowsky, J.L., and Chien, C.-B. (2005). Molecular cloning and developmental expression of foxP2 in zebrafish. Dev. Dyn. *234*, 740–746.

Brainard, M.S., and Doupe, a J. (2000). Auditory feedback in learning and maintenance of vocal behaviour. Nat. Rev. Neurosci. *1*, 31–40.

Brennan, R., and Roderick, S. (1990). Protein-DNA conformational changes in the crystal structure of a lambda Cro-operator complex. Proc. Natl. Acad. Sci. *87*, 8165–8169.

Brennan, R.G., and Matthews, W. (1989). The Helix-Turn-Helix DNA Binding Motif. *264*, 22–25.

Brewster, R.C., Weinert, F.M., Garcia, H.G., Song, D., Rydenfelt, M., and Phillips, R. (2014). The transcription factor titration effect dictates level of gene expression. Cell *156*, 1312–1323.

Bruce, H. a, and Margolis, R.L. (2002). FOXP2: novel exons, splice variants, and CAG repeat length stability. Hum. Genet. *111*, 136–144.

Bustamante, C.D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M.T., Glanowski, S., Tanenbaum, D.M., White, T.J., Sninsky, J.J., Hernandez, R.D., et al. (2005). Natural selection on protein-coding genes in the human genome. Nature *437*, 1153–1157.

Callis, P.R., and Liu, T. (2004). Quantitative Prediction of Fluorescence Quantum Yields for Tryptophan in Proteins. J. Phys. Chem. B *108*, 4248–4259.

Campbell, A.J., Lyne, L., Brown, P.J., Launchbury, R.J., Bignone, P., Chi, J., Roncador, G., Lawrie, C.H., Gatter, K.C., Kusec, R., et al. (2010). Aberrant expression of the neuronal transcription factor FOXP2 in neoplastic plasma cells. Br. J. Haematol. *149*, 221–230.

Campbell, P., Reep, R.L., Stoll, M.L., Ophir, A.G., and Phelps, S.M. (2009). Conservation and diversity of Foxp2 expression in muroid rodents: functional implications. J. Comp. Neurol. *512*, 84–100.

Carlsson, P., and Mahlapuu, M. (2002). Forkhead Transcription Factors: Key Players in Development and Metabolism. Dev. Biol. *250*, 1–23.

Carr, C.W., Moreno-De-Luca, D., Parker, C., Zimmerman, H.H., Ledbetter, N., Martin, C.L., Dobyns, W.B., and Abdul-Rahman, O. a (2010). Chiari I malformation, delayed gross motor skills, severe speech delay, and epileptiform discharges in a child with FOXP1 haploinsufficiency. Eur. J. Hum. Genet. *18*, 1216–1220.

Castillo Bosch, P., Segura-Bayona, S., Koole, W., van Heteren, J.T., Dewar, J.M., Tijsterman, M., and Knipscheer, P. (2014). FANCJ promotes DNA synthesis through G-quadruplex structures. EMBO J. *33*, 2521–2533.

Chen, Y., and Barkley, M. (1998). Toward understanding tryptophan fluorescence in proteins. Biochemistry *2960*, 9976–9982.

Chen, Y., Chen, C., Zhang, Z., Liu, C.-C., Johnson, M.E., Espinoza, C. a., Edsall, L.E., Ren, B., Zhou, X.J., Grant, S.F. a., et al. (2015). DNA binding by FOXP3 domain-swapped dimer suggests mechanisms of long-range chromosomal interactions. Nucleic Acids Res. *43*, 1268–1282.

Coulocheri, S., and Pigis, D. (2007). Hydrogen bonds in protein–DNA complexes: where geometry meets plasticity. Biochimie *89*, 1291–1303.

Cuiffo, B.G., Cam-, A., Taverna, D., Antoine, E., Campagne, A., Bell, G.W., Lembo, A., Orso, F., Lien, E.C., Bhasin, M.K., et al. MSC-Regulated MicroRNAs Converge on the Transcription Factor FOXP2 and Promote Breast Article MSC-Regulated MicroRNAs Converge on the Transcription Factor FOXP2 and Promote Breast Cancer Metastasis. Stem Cell 1–13.

Devanna, P., Middelbeek, J., and Vernes, S.C. (2014). FOXP2 drives neuronal differentiation by interacting with retinoic acid signaling pathways. Front. Cell. Neurosci. *8*, 1–13.

Dickerson, R. (1998). DNA bending: the prevalence of kinkiness and the virtues of normality. Nucleic Acids Res. *26*, 1906–1926.

Dickerson, R.E. (1983). Base sequence and helix structure variation in B and A DNA. J. Mol. Biol. *166*, 419–441.

Dijk, M. Van, and Bonvin, A. (2009). 3D-DART: a DNA structure modelling server. Nucleic Acids Res. *37*, 235–239.

Doucleff, M., and Clore, G.M. (2008). Global jumping and domain-specific intersegment transfer between DNA cognate sites of the multidomain transcription factor Oct-1. Proc. Natl. Acad. Sci. *105*, 13871–13876.

Doyle, M. (1997). Characterization of binding interactions by isothermal titration calorimetry. Curr. Opin. Biotechnol.

Duckett, D.R., Murchie, A.I.H., Diekmann, S., von Kitzing, E., Kemper, B., and Lilley, D.M.J. (1988). The structure of the holliday junction, and its resolution. Cell *55*, 79–89.

Elf, J., Li, G.-W., and Xie, X.S. (2007). Probing transcription factor dynamics at the single-molecule level in a living cell. Science *316*, 1191–1194.

Enard, W., Gehre, S., Hammerschmidt, K., Halter, S.M., Blass, T., Somel, M., Br??ckner, M.K., Schreiweis, C., Winter, C., Sohr, R., et al. (2009). A Humanized Version of Foxp2 Affects Cortico-Basal Ganglia Circuits in Mice. Cell *137*, 961–971.

Esadze, A., Kemme, C. a, Kolomeisky, A.B., and Iwahara, J. (2014). Positive and negative impacts of nonspecific sites during target location by a sequence-specific DNA-binding protein: origin of the optimal search at physiological ionic strength. Nucleic Acids Res. *42*, 7039–7046.

Fanucchi, S., Shibayama, Y., Burd, S., Weinberg, M.S., and Mhlanga, M.M. (2013). Chromosomal contact permits transcription between coregulated genes. Cell *155*, 606–620.

Felsenfeld, G., Davies, D., and Rich, A. (1957). Formation of a three-stranded polynucleotide molecule. J. Am. *79*, 2022–2023.

Feuk, L., Kalervo, A., and Lipsanen-Nyman, M. (2006). Absence of a Paternally Inherited FOXP2 Gene in Developmental Verbal Dyspraxia. Am. J. Hum. Genet. *79*, 965–972.

Le Fevre, A.K., Taylor, S., Malek, N.H., Horn, D., Carr, C.W., Abdul-Rahman, O. a., O'Donnell, S., Burgess, T., Shaw, M., Gecz, J., et al. (2013). FOXP1 mutations cause intellectual disability and a recognizable phenotype. Am. J. Med. Genet. Part A *161*, 3166–3175.

Florescu, A.-M., and Joyeux, M. (2009). Description of nonspecific DNA-protein interaction and facilitated diffusion with a dynamical model. J. Chem. Phys. *130*, 015103.

Fox, S.B., Brown, P., Han, C., Ashe, S., Leek, R.D., Harris, A.L., and Banham, A.H. (2004). Expression of the Forkhead Transcription Factor FOXP1 Is Associated with Estrogen Receptor α and Improved Survival in Primary Human Breast Carcinomas Expression of the Forkhead Transcription Factor FOXP1 Is Associated with Estrogen Receptor and Improved. Clin. Cancer Res. *10*, 3521–3527.

Frank-Kamenetskii, M.D., and Mirkin, S.M. (1995). Triplex DNA structures. Annu. Rev. Biochem. *64*, 65–95.

French, C. a., and Fisher, S.E. (2014). What can mice tell us about Foxp2 function? Curr. Opin. Neurobiol. *28*, 72–79.

Fujita, E., Tanabe, Y., and Shiota, A. (2008). Ultrasonic vocalization impairment of Foxp2 (R552H) knockin mice related to speech-language disorder and abnormality of Purkinje cells. Proc. … *2*, 2–7.

Fuxreiter, M., Simon, I., and Bondos, S. (2011). Dynamic protein-DNA recognition: beyond what can be seen. Trends Biochem. Sci. *36*, 415–423.

Gajiwala, K.S., and Burley, S.K. (2000). Winged helix proteins. Curr. Opin. Struct. Biol. *10*, 110–116.

Gao, F., Foat, B., and Bussemaker, H. (2004). Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. BMC Bioinformatics *10*, 1–10.

Gaub, S., Groszer, M., Fisher, S.E., and Ehret, G. (2010). The structure of innate vocalizations in Foxp2-deficient mouse pups. Genes. Brain. Behav. *9*, 390–401.

Gaudet, J., and Mango, S.E. (2002). Regulation of organogenesis by the Caenorhabditis elegans FoxA protein PHA-4. Science *295*, 821–825.

Gauthier, J., and Joober, R. (2003). Mutation screening of FOXP2 in individuals diagnosed with autistic disorder. Am. J. Med. Genet. *175*, 172–175.

Georges, A.B., Benayoun, B. a, Caburet, S., and Veitia, R. a (2010). Generic binding sites, generic DNA-binding domains: where does specific promoter recognition come from? FASEB J. *24*, 346–356.

Givaty, O., and Levy, Y. (2009). Protein Sliding along DNA : Dynamics and Structural Characterization. J. Mol. Biol. *385*, 1087–1097.

Gong, X., Jia, M., Ruan, Y., and Shuang, M. (2004). Association between the FOXP2 gene and autistic disorder in Chinese population. Am. J. Hum. Genet. *116*, 113–116.

Gorkin, D.U., Leung, D., and Ren, B. (2014). The 3D genome in transcriptional regulation and pluripotency. Cell Stem Cell *14*, 771–775.

Gronenborn, A. (2009). Protein acrobatics in pairs–dimerization via domain swapping. Curr. Opin. Struct. Biol. *19*, 39–49.

Groszer, M., Keays, D. a., Deacon, R.M.J., de Bono, J.P., Prasad-Mulcare, S., Gaub, S., Baum, M.G., French, C. a., Nicod, J., Coventry, J. a., et al. (2008). Impaired Synaptic Plasticity and Motor Learning in Mice with a Point Mutation Implicated in Human Speech Deficits. Curr. Biol. *18*, 354–362.

Haesler, S., Wada, K., Nshdejan, a, Morrisey, E.E., Lints, T., Jarvis, E.D., and Scharff, C. (2004). FoxP2 expression in avian vocal learners and non-learners. J. Neurosci. *24*, 3164–3175.

Haesler, S., Rochefort, C., Georgi, B., Licznerski, P., Osten, P., and Scharff, C. (2007). Incomplete and inaccurate vocal imitation after knockdown of FoxP2 in songbird basal ganglia nucleus Area X. PLoS Biol. *5*, e321.

Halford, S.E. (2009). An end to 40 years of mistakes in DNA-protein association kinetics? Biochem. Soc. Trans. *37*, 343–348.

Halford, S.E., and Marko, J.F. (2004). How do site-specific DNA-binding proteins find their targets? Nucleic Acids Res. *32*, 3040–3052.

Hamdan, F.F., Daoud, H., Rochefort, D., Piton, A., Gauthier, J., Langlois, M., Foomani, G., Dobrzeniecka, S., Krebs, M.O., Joober, R., et al. (2010). De novo mutations in FOXP1 in cases with intellectual disability, autism, and language impairment. Am. J. Hum. Genet. *87*, 671–678.

Hannenhalli, S., and Kaestner, K.H. (2009). The evolution of Fox genes and their role in development and disease. Nat. Rev. Genet. *10*, 233–240.

Harami, G.M., Gyimesi, M., and Kovács, M. (2013). From keys to bulldozers: expanding roles for winged helix domains in nucleic-acid-binding proteins. Trends Biochem. Sci. *38*, 364–371.

Harrison, S., and Aggarwal, A. (1990). DNA recognition by proteins with the helix-turn-helix motif. Annu. Rev. Biochem. *59*, 933–969.

Harteis, S., and Schneider, S. (2014). Making the bend: DNA tertiary structure and protein-DNA interactions. Int. J. Mol. Sci. *15*, 12335–12363.

Hartwig, A. (2001). Role of magnesium in genomic stability. Mutat. Res. Mol. Mech. Mutagen. *475*, 113–121.

Hayden, M., and Ghosh, S. (2012). NF-κB, the first quarter-century: remarkable progress and outstanding questions. Genes Dev. *26*, 203–234.

Henikoff, S., Henikoff, J.G., and Hutchinson, F. (1994). Position-based Sequence Weights. J. Mol. Biol. *243*, 574–578.

Hippel, P. Von, and Berg, O. (1986). On the specificity of DNA-protein interactions. Proc. Natl. Acad. Sci. *83*, 1608–1612.

Horn, D., Kapeller, J., Rivera-Brugués, N., Moog, U., Lorenz-Depiereux, B., Eck, S., Hempel, M., Wagenstaller, J., Gawthrope, A., Monaco, A.P., et al. (2010). Identification of FOXP1 deletions in three unrelated patients with mental retardation and significant speech and language deficits. Hum. Mutat. *31*, E1851–E1860.

Hu, H., Wang, B., Borde, M., Nardone, J., Maika, S., Allred, L., Tucker, P.W., and Rao, A. (2006). Foxp1 is an essential transcriptional regulator of B cell development. Nat. Immunol. *7*, 819–826.

Hurst, J., and Baraitser, M. (1990). An extended family with a dominantly inherited speech disorder. Dev. Med. Child Neurol. *32*, 347–355.

Itakura, T., Chandra, A., Yang, Z., Xue, X., Wang, B., Kimura, W., Hikosaka, K., Inohaya, K., Kudo, A., Uezato, T., et al. (2008). The medaka FoxP2, a homologue of human language gene FOXP2, has a diverged structure and function. J. Biochem. *143*, 407–416.

Jacobo-Molina, A., and Ding, J. (1993). Crystal structure of human immunodeficiency virus type 1 reverse transcriptase complexed with double-stranded DNA at 3.0 A resolution shows bent DNA. Proc. Natl. Acad. Sci. *90*, 6320–6324.

Jankowski, A., Szczurek, E., Jauch, R., Tiuryn, J., and Prabhakar, S. (2013). Comprehensive prediction in 78 human cell lines reveals rigidity and compactness of transcription factor dimers. Genome Res. *23*, 1307–1318.

Jayaram, B., Sharp, K. a, and Honig, B. (1989). The electrostatic potential of B-DNA. Biopolymers *28*, 975–993.

Jiang, J., and Levine, M. (1993). Binding affinities and cooperative interactions with bHLH activators delimit threshold responses to the dorsal gradient morphogen. Cell *72*, 741–752.

Jones, S., van Heyningen, P., Berman, H.M., and Thornton, J.M. (1999). Protein-DNA interactions: A structural analysis. J. Mol. Biol. *287*, 877–896.

Jordan, S.R., and Pabo, C. (1988). structure of the lambda Complex at 2.5 A Resolution : Details of the Interactions. Science (80-.). *242*, 893–899.

Josefowicz, S.Z., Lu, L.-F., and Rudensky, A.Y. (2012). Regulatory T Cells: Mechanisms of Differentiation and Function. Annu. Rev. Immunol. *30*, 531–564.

Kaestner, K.H., Knöchel, W., Martínez, D.E., Kno, W., and Martı, D.E. (2000). Unified nomenclature for the winged heix/forkhead transcription factors. 142–146.

Kao-Huang, Y., and Revzin, A. (1977). Nonspecific DNA binding of genome-regulating proteins as a biological control mechanism: measurement of DNA-bound Escherichia coli lac repressor in vivo. Proc. Natl. Acad. Sci. *74*, 4228–4232.

Kathiresan, S., and Srivastava, D. (2012). Genetics of human cardiovascular disease. Cell *148*, 1242–1257.

Kaufmann, E., Müller, D., and Knöchel, W. (1995). DNA recognition site analysis of Xenopus wingedhelix proteins. J. Mol. Biol. *248*, 239–254.

Kazemian, M., Pham, H., Wolfe, S. a, Brodsky, M.H., and Sinha, S. (2013). Widespread evidence of cooperative DNA binding by transcription factors in Drosophila development. Nucleic Acids Res. *41*, 8237–8252.

Keleher, C., Goutte, C., and Johnson, A. (1988). The yeast cell-type-specific repressor α2 acts cooperatively with a non-cell-type-specific protein. Cell *53*, 927–936.

Kim, K., and Khayrutdinov, B. (2011). Solution structure of the Zβ domain of human DNA-dependent activator of IFN-regulatory factors and its binding modes to B-and Z-DNAs. Proc. Natl. Acad. Sci. *108*, 6921–6926.

Kim, I.-M., Ackerson, T., Ramakrishna, S., Tretiakova, M., Wang, I.-C., Kalin, T. V, Major, M.L., Gusarova, G. a, Yoder, H.M., Costa, R.H., et al. (2006). The Forkhead Box m1 transcription factor stimulates the proliferation of tumor cells during development of lung cancer. Cancer Res. *66*, 2153–2161.

Kitano, K., Kim, S., and Hakoshima, T. (2010). Structural basis for DNA strand separation by the unconventional winged-helix domain of RecQ helicase WRN. Structure *18*, 177–187.

Kohler, J., and Metallo, S. (1999). DNA specificity enhanced by sequential binding of protein monomers. Proc. Natl. Acad. Sci. *96*, 11735–11739.

Kyewski, B., and Klein, L. (2006). A central role for central tolerance. Annu. Rev. Immunol. *24*, 571–606.

Lai, E., and Clark, K. (1993). Hepatocyte nuclear factor 3/fork head or“ winged helix” proteins: a family of transcription factors of diverse biologic function. Proc. Natl. Acad. Sci. *90*, 10421–10423.

Lai, C.S., Fisher, S.E., Hurst, J.A., Vargha-Khadem, F., and Monaco, A.P. (2001). A forkhead-domain gene is mutated in a severe speech and language disorder. Nature *413*, 519–523.

Laity, J.H., Dyson, H.J., and Wright, P.E. (2000). DNA-induced a -Helix Capping in Conserved Linker Sequences is a Determinant of Binding Affinity in Cys 2 -His 2 Zinc Fingers. J. Mol. Biol. *295*, 719–727.

Lander, E., Linton, L., Birren, B., and Nusbaum, C. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Leavitt, S., and Freire, E. (2001). Direct measurement of protein binding energetics by isothermal titration calorimetry. Curr. Opin. Struct. Biol. *11*, 560–566.

Lee, T.I., and Young, R. a (2013). Transcriptional regulation and its misregulation in disease. Cell *152*, 1237–1251.

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 Transcriptional and DNA Binding Activity of the Foxp1 / 2 / 4 Family Is Modulated by Heterotypic and Homotypic Protein Inte. Mol. Cell. Biol. *24*, 809–822.

Li, T., Zeng, Z., Zhao, Q., Wang, T., Huang, K., Li, J., Li, Y., Liu, J., Wei, Z., Wang, Y., et al. (2013). FoxP2 is significantly associated with schizophrenia and major depression in the Chinese Han population. World J. Biol. Psychiatry *14*, 146–150.

Li, Z., Tuteja, G., Schug, J., and Kaestner, K.H. (2012). Foxa1 and Foxa2 are essential for sexual dimorphism in liver cancer. Cell *148*, 72–83.

Liégeois, F., Baldeweg, T., Connelly, A., Gadian, D.G., Mishkin, M., and Vargha-Khadem, F. (2003). Language fMRI abnormalities associated with FOXP2 gene mutation. Nat. Neurosci. *6*, 1230–1237.

Lin, S., and Riggs, A.D. (1975). The General Affinity of lac Repressor for E . coli DNA : Implications for Gene Regulation in Procaryotes and Eucaryotes. Cell *4*, 107–111.

Littler, D.R., Alvarez-Fernández, M., Stein, a, Hibbert, R.G., Heidebrecht, T., Aloy, P., Medema, R.H., and Perrakis, a (2010). Structure of the FoxM1 DNA-recognition domain bound to a promoter sequence. Nucleic Acids Res. *38*, 4527–4538.

Littlewood, T.D., Kreuzaler, P., and Evan, G.I. (2012). All things to all people. Cell *151*, 11–13.

Lopes, J., Kriegsman, B., Foiani, M., and Nicolas, A. (2011). G-quadruplex-induced instability during leading- strand replication. *30*, 4033–4046.

Lopes, J.E., Torgerson, T.R., Schubert, L. a., Anover, S.D., Ocheltree, E.L., Ochs, H.D., and Ziegler, S.F. (2006). Analysis of FOXP3 Reveals Multiple Domains Required for Its Function as a Transcriptional Repressor. J. Immunol. *177*, 3133–3142.

Lu, M.M., Li, S., Yang, H., and Morrisey, E.E. (2002). Foxp4 : a novel member of the Foxp subfamily of winged-helix genes co-expressed with Foxp1 and Foxp2 in pulmonary and gut tissues. Mech. Dev. *1198*, 197–202.

Luo, H., Jin, K., Xie, Z., Qiu, F., Li, S., Zou, M., Cai, L., Hozumi, K., Shima, D.T., and Xiang, M. (2012). Forkhead box N4 (Foxn4) activates Dll4-Notch signaling to suppress

photoreceptor cell fates of early retinal progenitors. Proc. Natl. Acad. Sci. U. S. A. *109*, E553–E562.

Luscombe, N. (2001). Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. Nucleic Acids Res. *29*, 2860–2874.

Luscombe, N., and Austin, S. (2000). An overview of the structures of protein-DNA complexes. Genome 1–37.

MacDermot, K.D., Bonora, E., Sykes, N., Coupe, A.-M., Lai, C.S.L., Vernes, S.C., Vargha-Khadem, F., McKenzie, F., Smith, R.L., Monaco, A.P., et al. (2005). Identification of FOXP2 truncation as a novel cause of developmental speech and language deficits. Am. J. Hum. Genet. *76*, 1074–1080.

Madureira, P. a, Varshochi, R., Constantinidou, D., Francis, R.E., Coombes, R.C., Yao, K.-M., and Lam, E.W.-F. (2006). The Forkhead box M1 protein regulates the transcription of the estrogen receptor alpha in breast cancer cells. J. Biol. Chem. *281*, 25167–25176.

Majka, J., and Speck, C. (2007). Analysis of protein–DNA interactions using surface plasmon resonance. Anal. Protein–DNA Interact. *104*, 13–36.

Mangelsdorf, D., and Evans, R. (1995). The RXR heterodimers and orphan receptors. Cell *83*, 841–850.

Mansfield, J., Slater, C., and Byers, R. (2014). Validation and clinical correlation of triplex CD3, CD8 and FOXP3 IHC of tumor-infiltrating lymphocytes in follicular lymphoma. J. Immunother. Cancer *2*, P261.

Mardis, E.R. (2013). Next-generation sequencing platforms. Annu. Rev. Anal. Chem. (Palo Alto. Calif). *6*, 287–303.

Mark, W.-Y., Liao, J.C.C., Lu, Y., Ayed, A., Laister, R., Szymczyna, B., Chakrabartty, A., and Arrowsmith, C.H. (2005). Characterization of segments from the central region of BRCA1: an intrinsically disordered scaffold for multiple protein-protein and protein-DNA interactions? J. Mol. Biol. *345*, 275–287.

Marsden, I., Jin, C., and Liao, X. (1998). Structural changes in the region directly adjacent to the DNA-binding helix highlight a possible mechanism to explain the observed changes in the sequence-specific binding of winged helix proteins. J. Mol. Biol. *278*, 293–299.

Matos, J., and West, S.C. (2014). Holliday junction resolution : Regulation in space and time. DNA Repair (Amst). *19*, 176–181.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.

McCarthy-Jones, S., Green, M.J., Scott, R.J., Tooney, P. a, Cairns, M.J., Wu, J.Q., Oldmeadow, C., and Carr, V. (2014). Preliminary evidence of an interaction between the FOXP2 gene and childhood emotional abuse predicting likelihood of auditory verbal hallucinations in schizophrenia. J. Psychiatr. Res. *50*, 66–72.

McLeay, R., and Bailey, T. (2010). Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. BMC Bioinformatics.

Meng, E.C., Pettersen, E.F., Couch, G.S., Huang, C.C., and Ferrin, T.E. (2006). Tools for integrated sequence-structure analysis with UCSF Chimera. BMC Bioinformatics *7*, 339.

Metzker, M.L. (2010). Sequencing technologies - the next generation. Nat. Rev. Genet. *11*, 31–46.

Mirny, L., Slutsky, M., Wunderlich, Z., Tafvizi, A., Leith, J., and Kosmrlj, A. (2009). How a protein searches for its site on DNA: the mechanism of facilitated diffusion. J. Phys. A Math. Theor. *42*, 434013.

Miyata, Y., Fukuhara, A., Otsuki, M., and Shimomura, I. (2013). Expression of activating transcription factor 2 in inflammatory macrophages in obese adipose tissue. Obesity *21*, 731–736.

Myszka, D. (1997). Kinetic analysis of macromolecular interactions using surface plasmon resonance biosensors. Curr. Opin. Biotechnol. 50–57.

Nakagawa, S., Gisselbrecht, S.S., Rogers, J.M., Hartl, D.L., and Bulyk, M.L. (2013). DNA-binding specificity changes in the evolution of forkhead transcription factors. Proc. Natl. Acad. Sci. U. S. A. *110*, 12349–12354.

Nakamura, K., and Jeong, S. (2001). SCA17, a novel autosomal dominant cerebellar ataxia caused by an expanded polyglutamine in TATA-binding protein. Hum. Mol. … *10*, 1441–1448.

Needleman, S., and Wunsch, C. (1970). A general method applicable to search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. *48*, 443–453.

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*, 5991–6004.

Nitsche, A., Kurth, A., Dunkhorst, A., Pänke, O., Sielaff, H., Junge, W., Muth, D., Scheller, F., Stöcklein, W., Dahmen, C., et al. (2007). One-step selection of Vaccinia virus-binding DNA aptamers by MonoLEX. BMC Biotechnol. *7*, 48.

O'Shea, E., Klemm, J., Kim, P., and Alber, T. (1991). X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. Science (80). *254*, 539–544.

Ohlendorf, D., and Matthew, J. (1985). Electrostatics and flexibility in protein-DNA interactions. Adv. Biophys. *20*, 137–151.

Overdier, D.G., Porcella, A., and Costa, R.H. (1994). The DNA-binding specificity of the hepatocyte nuclear factor 3 / forkhead domain is influenced by amino-acid residues adjacent to the recognition helix . The DNA-Binding Specificity of the Hepatocyte Nuclear Factor 3 / forkhead Domain Is Influenced by Ami. Mol. Cell. Biol. *14*, 2755–2766.

Paeschke, K., Simonsson, T., Postberg, J., Rhodes, D., and Lipps, H.J. (2005). Telomere end-binding proteins control the formation of G-quadruplex DNA structures in vivo. Nat. Struct. Mol. Biol. *12*, 847–854.

Park, Y., Won, S., Nam, M., Chung, J.-H., and Kwack, K. (2013). Interaction Between MAOA and FOXP2 in Association With Autism and Verbal Communication in a Korean Population. J. Child Neurol.

Passner, J., and Steitz, T. (1997). The structure of a CAP–DNA complex having two cAMP molecules bound to each monomer. Proc. Natl. … *94*, 2843–2847.

Pavletich, N., and Pabo, C. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. Science (80-.). *252*, 809–817.

Pavletich, N., and Pabo, C. (1993). Crystal structure of a five-finger GLI-DNA complex: new perspectives on zinc fingers. Science (80-.). *261*, 1701–1707.

Peng, S., Kuroda, M.I., and Park, P.J. (2010). Quantized correlation coefficient for measuring reproducibility of ChIP-chip data. BMC Bioinformatics *11*, 399.

Phair, R.D., Scaffidi, P., Elbi, C., Dey, A., Ozato, K., Brown, D.T., Bustin, M., Misteli, T., Vecerova, J., and Hager, G. (2004). Global Nature of Dynamic Protein-Chromatin Interactions In Vivo : Three-Dimensional Genome Scanning and Dynamic Interaction Networks of Chromatin Proteins Global Nature of Dynamic Protein-Chromatin Interactions In Vivo : Three-Dimensional Genome Scanning. Mol. Cell. Biol. *24*, 6393–6402.

Pierrou, S., Hellqvist, M., Samuelsson, L., Enerbäck, S., and Carlsson, P. (1994). Cloning and characterization of seven human forkhead proteins: binding site specificity and DNA bending. EMBO J. *13*, 5002–5012.

Putnam, C., Clancy, S., and Tsuruta, H. (2001). Structure and mechanism of the RuvB Holliday junction branch migration motor. J. Mol. Biol. *311*, 297–310.

Ramos, A., and Barolo, S. (2013). Low-affinity transcription factor binding sites shape morphogen responses and enhancer evolution. Philos. Trans. ….

Reddy, C., Das, A., and Jayaram, B. (2001). Do water molecules mediate protein-DNA recognition? J. Mol. Biol. *314*, 619–632.

Rentzeperis, D., Ho, J., and Marky, L. a. (1993). Contribution of loops and nicks to the formation of DNA dumbbells: Melting behavior and ligand binding. Biochemistry *32*, 2564–2572.

Rice, G.M., Raca, G., Jakielski, K.J., Laffin, J.J., Iyama-Kurtycz, C.M., Hartley, S.L., Sprague, R.E., Heintzelman, A.T., and Shriberg, L.D. (2012). Phenotype of FOXP2 haploinsufficiency in a mother and son. Am. J. Med. Genet. Part A *158 A*, 174–181.

Rich, R.L., and Myszka, D.G. (2000). Advances in surface plasmon resonance biosensor analysis. Curr. Opin. Biotechnol. *11*, 54–61.

Rohs, R., West, S., Sosinsky, A., and Liu, P. (2009). The role of DNA shape in protein–DNA recognition. Nature *461*, 1248–1253.

Rohs, R., Jin, X., West, S.M., Joshi, R., Honig, B., and Mann, R.S. (2010). Origins of specificity in protein-DNA recognition. Annu. Rev. Biochem. *79*, 233–269.

Rothenburg, S., and Schwartz, T. (2002). Complex regulation of the human gene for the Z-DNA binding protein DLM-1. Nucleic Acids Res. *30*, 993–1000.

Rotherham, L.S., Maserumule, C., Dheda, K., Theron, J., and Khati, M. (2012). Selection and application of ssDNA aptamers to detect active TB from sputum samples. PLoS One *7*, e46862.

Rousso, D.L., Pearson, C.A., Gaber, Z.B., Miquelajauregui, A., Li, S., Portera-Cailliau, C., Morrisey, E.E., and Novitch, B.G. (2012). Foxp-mediated suppression of N-cadherin regulates neuroepithelial character and progenitor maintenance in the CNS. Neuron *74*, 314–330.

Rowan, S., Siggers, T., Lachke, S. a, Yue, Y., Bulyk, M.L., and Maas, R.L. (2010). Precise temporal control of the eye regulatory gene Pax6 via enhancer-binding site affinity. Genes Dev. *24*, 980–985.

Rump, P., Niessen, R.C., Verbruggen, K.T., Brouwer, O.F., De Raad, M., and Hordijk, R. (2011). A novel mutation in MED12 causes FG syndrome (Opitz-Kaveggia syndrome). Clin. Genet. *79*, 183–188.

Sagardoy, A., Martinez-Ferrandis, J.I., Roa, S., Bunting, K.L., Aznar, M.A., Elemento, O., Shaknovich, R., Fontán, L., Fresquet, V., Perez-Roger, I., et al. (2013). Downregulation of FOXP1 is required during germinal center B-cell function. Blood *121*, 4311–4320.

Sanjuán, J., Tolosa, A., and González, J. (2006). Association between FOXP2 polymorphisms and schizophrenia with auditory hallucinations. Psychiatr. … *16*, 67–72.

Sarai, A., and Kono, H. (2005). Protein-DNA recognition patterns and predictions. Annu. Rev. Biophys. Biomol. Struct. *34*, 379–398.

Scardigli, R. (2003). Direct and concentration-dependent regulation of the proneural gene Neurogenin2 by Pax6. Development *130*, 3269–3281.

Schnable, J.C., Pedersen, B.S., Subramaniam, S., and Freeling, M. (2011). Dose–Sensitivity, Conserved Non-Coding Sequences, and Duplicate Gene Retention Through Multiple Tetraploidies in the Grasses. Front. Plant Sci. *2*, 1–7.

Schön, C., Wochnik, A., Rössner, A., Donow, C., and Knöchel, W. (2006). The FoxP subclass in Xenopus laevis development. Dev. Genes Evol. *216*, 641–646.

Seeman, N.C., Rosenberg, J.M., and Rich, a. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. Proc. Natl. Acad. Sci. *73*, 804–808.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I.K., Wang, J.-P.Z., and Widom, J. (2006). A genomic code for nucleosome positioning. Nature *442*, 772–778.

Sharrocks, A. (2001). The ETS-domain transcription factor family. Nat. Rev. Mol. Cell Biol. *2*.

Shilov, I. V, Seymour, S.L., Patel, A. a, Loboda, A., Tang, W.H., Keating, S.P., Hunter, C.L., Nuwaysir, L.M., and Schaeffer, D. a (2007). The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. Mol. Cell. Proteomics *6*, 1638–1655.

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*, 27488–27497.

Sin, C., Li, H., and Crawford, D. a (2014). Transcriptional Regulation by FOXP1, FOXP2, and FOXP4 Dimerization. J. Mol. Neurosci. *In Press*.

Sohrabji, F., Nordeen, E.J., and Nordeen, K.W. (1990). Selective impairment of song learning following lesions of a forebrain nucleus in the juvenile zebra finch. Behav. Neural Biol. *53*, 51–63.

Španiel, F., Horáček, J., Tintěra, J., Ibrahim, I., Novák, T., Čermák, J., Klírová, M., and Höschl, C. (2011). Genetic variation in FOXP2 alters grey matter concentrations in schizophrenia patients. Neurosci. Lett. *493*, 131–135.

Spiriti, J., and van der Vaart, A. (2012). DNA Bending through Roll Angles Is Independent of Adjacent Base Pairs. J. Phys. Chem. Lett. *3*, 3029–3033.

Spiteri, E., Konopka, G., and Coppola, G. (2007). Identification of the Transcriptional Targets of FOXP2, a Gene Linked to Speech and Language, in Developing Human Brain. Am. J. Hum. Genet. *81*, 1144–1157.

Spolar, R.S., and Record, M.T. (1994). Coupling of local folding to site-specific binding of proteins to DNA. Science *263*, 777–784.

Strauss, K. (2006). Recessive symptomatic focal epilepsy and mutant contactin-associated protein-like 2. N. Engl. J. Med. *354*, 1370–1377.

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*, 159–166.

Struhl, K. (1989). Helix-turn-helix, zinc-finger, and leucine-zipper motifs for eukaryotic transcriptional regulatory proteins. Trends Biochem. Sci. *14*, 137–140.

Stumm, L., Burkhardt, L., Steurer, S., Simon, R., Adam, M., Becker, A., Sauter, G., Minner, S., Schlomm, T., Sirma, H., et al. (2013). Strong expression of the neuronal transcription factor FOXP2 is linked to an increased risk of early PSA recurrence in ERG fusion-negative cancers. J. Clin. Pathol. *66*, 563–568.

Takahashi, K., Liu, F.-C., Hirokawa, K., and Takahashi, H. (2008). Expression of Foxp4 in the developing and adult rat forebrain. J. Neurosci. Res. *86*, 3106–3116.

Takaoka, A., Wang, Z., Choi, M.K., Yanai, H., Negishi, H., Ban, T., Lu, Y., Miyagishi, M., Kodama, T., Honda, K., et al. (2007). DAI (DLM-1/ZBP1) is a cytosolic DNA sensor and an activator of innate immune response. Nature *448*, 501–505.

Tam, W.Y., Leung, C.K.Y., Tong, K.K., and Kwan, K.M. (2011). Foxp4 is essential in maintenance of purkinje cell dendritic arborization in the mouse cerebellum. Neuroscience *172*, 562–571.

Tanay, A. (2006). Extensive low-affinity transcriptional interactions in the yeast genome. Genome Res. *16*, 962–972.

Teplova, M., Yuan, Y.-R., Phan, A.T., Malinina, L., Ilin, S., Teplov, A., and Patel, D.J. (2006). Structural basis for recognition and sequestration of UUU(OH) 3’ temini of nascent RNA polymerase III transcripts by La, a rheumatic disease autoantigen. Mol. Cell *21*, 75–85.

Teufel, A., Wong, E. a., Mukhopadhyay, M., Malik, N., and Westphal, H. (2003). FoxP4, a novel forkhead transcription factor. Biochim. Biophys. Acta - Gene Struct. Expr. *1627*, 147–152.

Todeschini, A.-L., Georges, A., and Veitia, R. a (2014). Transcription factors: specific DNA binding and specific gene regulation. Trends Genet. *30*, 211–219.

Tolosa, A., Sanjuán, J., Dagnall, A.M., Moltó, M.D., Herrero, N., and de Frutos, R. (2010). FOXP2 gene and language impairment in schizophrenia: association and epigenetic studies. BMC Med. Genet. *11*, 114.

Toma, C., Hervás, A., Torrico, B., Balmaña, N., Salgado, M., Maristany, M., Vilella, E., Martínez-Leal, R., Planelles, M.I., Cuscó, I., et al. (2013). Analysis of two language-related genes in autism: a case-control association study of FOXP2 and CNTNAP2. Psychiatr. Genet. *23*, 82–85.

Travers, A. (1989). DNA conformation and protein binding. Annu. Rev. Biochem. *58*, 427–452.

Triezenberg, S. (1995). Structure and function of transcriptional activation domains. Curr. Opin. Genet. Dev. *5*, 190–196.

Tsai, K.-L., Huang, C.-Y., Chang, C.-H., Sun, Y.-J., Chuang, W.-J., and Hsiao, C.-D. (2006). Crystal structure of the human FOXK1a-DNA complex and its implications on the diverse binding specificity of winged helix/forkhead proteins. J. Biol. Chem. *281*, 17400–17409.

Tuerk, C., and Gold, L. (1990). Systemic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science (80-.). *249*, 505–510.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S. a, and Luscombe, N.M. (2009). A census of human transcription factors: function, expression and evolution. Nat. Rev. Genet. *10*, 252–263.

Vargha-Khadem, F., Watkins, K.E., Price, C.J., Ashburner, J., Alcock, K.J., Connelly, a, Frackowiak, R.S., Friston, K.J., Pembrey, M.E., Mishkin, M., et al. (1998). Neural basis of an inherited speech and language disorder. Proc. Natl. Acad. Sci. U. S. A. *95*, 12695–12700.

Veitia, R. a., Bottani, S., and Birchler, J. a. (2013). Gene dosage effects: Nonlinearities, genetic interactions, and dosage compensation. Trends Genet. *29*, 385–393.

Vernes, S., and Newbury, D. (2008). A functional genetic link between distinct developmental language disorders. N. Engl. J. Med. *359*, 2337–2345.

Vernes, S., Spiteri, E., and Nicod, J. (2007). High-Throughput Analysis of Promoter Occupancy Reveals Direct Neural Targets of FOXP2, a Gene Mutated in Speech and Language Disorders. Am. J. Hum. Genet. *81*, 1232–1250.

Vernes, S.C., Nicod, J., Elahi, F.M., Coventry, J. a., Kenny, N., Coupe, A.M., Bird, L.E., Davies, K.E., and Fisher, S.E. (2006). Functional genetic analysis of mutations implicated in a human speech and language disorder. Hum. Mol. Genet. *15*, 3154–3167.

De Vries, S.J., van Dijk, M., and Bonvin, A.M.J.J. (2010). The HADDOCK web server for data-driven biomolecular docking. Nat. Protoc. *5*, 883–897.

Vuzman, D., and Levy, Y. (2012). Intrinsically disordered regions as affinity tuners in protein-DNA interactions. Mol. Biosyst. *8*, 47–57.

Wang, A., Quigley, G., and Kolpak, F. (1981). Left-handed double helical DNA: variations in the backbone conformation. Science (80-.). *211*.

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*, 24259–24268.

Wang, J., Lu, J., Gu, G., and Liu, Y. (2011). In vitro DNA-binding profile of transcription factors: methods and new insights. J. Endocrinol. *210*, 15–27.

Watkins, K.E., Vargha-Khadem, F., Ashburner, J., Passingham, R.E., Connelly, a, Friston, K.J., Frackowiak, R.S.J., Mishkin, M., and Gadian, D.G. (2002). MRI analysis of an inherited speech and language disorder: structural brain abnormalities. Brain *125*, 465–478.

Weber, I.T., and Steitz, T. a. (1987). Structure of a complex of catabolite gene activator protein and cyclic AMP refined at 2.5 Å resolution. J. Mol. Biol. *198*, 311–326.

Weigel, D., Jürgens, G., Küttner, F., Seifert, E., and Jäckle, H. (1989). The homeotic gene fork head encodes a nuclear protein and is expressed in the terminal regions of the Drosophila embryo. Cell *57*, 645–658.

White, M. a, Parker, D.S., Barolo, S., and Cohen, B. a (2012). A model of spatially restricted transcription in opposing gradients of activators and repressors. Mol. Syst. Biol. *8*, 614.

Wilson, K. (1992). Escherichia coli biotin holoenzyme synthetase/bio repressor crystal structure delineates the biotin-and DNA-binding domains. Proc. … *89*, 9257–9261.

Wilson, D., Charoensawan, V., Kummerfeld, S.K., and Teichmann, S. a. (2008). DBD - Taxonomically broad transcription factor predictions: New content and functionality. Nucleic Acids Res. *36*, 88–92.

Wolfe, S. a., Grant, R. a., Elrod-Erickson, M., and Pabo, C.O. (2001). Beyond the "recognition code": Structures of two Cys2His2 zinc finger/TaTa box complexes. Structure *9*, 717–723.

Wolin, S., and Cedervall, T. (2002). The LA protein. Annu. Rev. Biochem. 375–403.

Wu, Y., Borde, M., Heissmeyer, V., Feuerer, M., Lapan, A.D., Stroud, J.C., Bates, D.L., Guo, L., Han, A., Ziegler, S.F., et al. (2006). FOXP3 controls regulatory T cell function through cooperation with NFAT. Cell *126*, 375–387.

Xiong, Y., and Sundaralingam, M. (2001). Protein–nucleic acid interaction: major groove recognition determinants. eLS 1–8.

Yagi, H., Nomura, T., Nakamura, K., Yamazaki, S., Kitawaki, T., Hori, S., Maeda, M., Onodera, M., Uchiyama, T., Fujii, S., et al. (2004). Crucial role of FOXP3 in the development and function of human CD25+CD4+ regulatory T cells. Int. Immunol. *16*, 1643–1656.

Yamada, K., Ariyoshi, M., and Morikawa, K. (2004). Three-dimensional structural views of branch migration and resolution in DNA homologous recombination. Curr. Opin. Struct. Biol. 130–137.

Yang, M., Wang, Y., Wang, X., Chen, C., and Zhou, H. (2010). Characterization of grass carp (Ctenopharyngodon idellus) Foxp1a/1b/2: evidence for their involvement in the activation of peripheral blood lymphocyte subpopulations. Fish Shellfish Immunol. *28*, 289–295.

Zeesman, S., and Nowaczyk, M. (2006). Speech and language impairment and oromotor dyspraxia due to deletion of 7q31 that involves FOXP2. Am. J. … *514*, 509–514.

Zhao, F., Siu, M.K.Y., Jiang, L., Tam, K.F., Ngan, H.Y.S., Le, X.F., Wong, O.G.W., Wong, E.S.Y., Gomes, A.R., Bella, L., et al. (2014). Overexpression of Forkhead Box Protein M1 (FOXM1) in Ovarian Cancer Correlates with Poor Patient Survival and Contributes to Paclitaxel Resistance. PLoS One *9*, e113478.

Zheng, N., and Fraenkel, E. (1999). Structural basis of DNA recognition by the heterodimeric cell cycle transcription factor E2F–DP. Genes … 666–674.

Zhu, C., Byers, K.J.R.P., McCord, R.P., Shi, Z., Berger, M.F., Newburger, D.E., Saulrieta, K., Smith, Z., Shah, M. V, Radhakrishnan, M., et al. (2009). High-resolution DNA-binding specificity analysis of yeast transcription factors. Genome Res. *19*, 556–566.

Zou, Y., Tsai, W.-B., Cheng, C.-J., Hsu, C., Chung, Y.M., Li, P.-C., Lin, S.-H., and Hu, M.C.T. (2008). Forkhead box transcription factor FOXO3a suppresses estrogen-dependent breast cancer cell proliferation and tumorigenesis. Breast Cancer Res. *10*, R21.